Automatic Particle Picking of Biological Molecules Imaged by Electron Microscopy

Jasmine Banks[1], Rosalba Rothnagel[2] and Ben Hankamer[2]
[1]Advanced Computational Modelling Centre,
University of Queensland, Brisbane 4072, Australia.
jbanks@acmc.uq.edu.au
[2]Institute for Molecular Biosciences,
University of Queensland, Brisbane 4072, Australia.
{r.rothnagel, b.hankamer}@imb.uq.edu.au

Abstract

One of the next great challenges of cell biology is the determination of the enormous number of protein structures encoded in genomes. In recent years, advances in electron cryo-microscopy and high-resolution single particle analysis have developed to the point where they now provide a methodology for high resolution structure determination. Using this approach, images of randomly oriented single particles are aligned computationally to reconstruct 3-D structures of proteins and even whole viruses. One of the limiting factors in obtaining high-resolution reconstructions is obtaining a large enough representative dataset ($> 100,000$ particles). Traditionally particles have been manually picked which is an extremely labour intensive process. The problem is made especially difficult by the low signal-to-noise ratio of the images.
This paper describes the development of automatic particle picking software, which has been tested with both negatively stained and cryo-electron micrographs. This algorithm has been shown to be capable of selecting most of the particles, with few false positives. Further work will involve extending the software to detect differently shaped and oriented particles.

Keywords: electron microscopy, single particle analysis, automatic particle picking, correlation.

1 Introduction

With the completion of a large number of genome sequences, one of the next great challenges of cell biology is the determination of the protein structures that they encode. The human genome project alone identified ∼35,000 genes encoding both soluble and membrane proteins (25–40% of total) [1]. *In vivo* these organise into macromolecular assemblies further increasing the level of structural complexity.

Traditionally protein structures have been solved from the diffraction pattern of 3-D crystals. However, particularly in the case of membrane proteins and macromolecular assemblies, the production of well-ordered crystals is a major bottleneck in structure determination. In recent years, advances in electron cryo-microscopy and high-resolution single particle analysis have developed to the point where they now provide an alternative methodology for high resolution structure determination [2]. Using this approach, images of randomly oriented single particles are aligned computationally (rather than biochemically during the production of 3-D crystals) to reconstruct 3-D structures of proteins and even whole viruses [3]. Modern cryo-electron microscopes are capable of recording structural information to a resolution higher than 2Å^2 (1Å=10^{-10}m) [2]. One of the limiting factors in obtaining 3-D reconstructions to this resolution level is the difficulty of generating a large enough data set ($> 100,000$ particles) which fully samples the 3-D volume at the required resolution and overcomes problems related to the low signal-to-noise (SNR) of cryo-images. Traditionally, particles have been picked manually but this approach is extremely labour intensive (∼1 week for 20,000 particles), and has proved to be a major bottleneck in the 3-D reconstruction process.

This paper describes the development of automatic particle picking software. It was first developed using negatively stained images of the protein ferritin, and has also been tested using cryo-electron micrographs of virus particles.

Cryo electron micrographs are obtained by suspending the purified protein molecules in a thin layer of vitreous ice, which is then imaged in the electron microscope through exposure with a low electron dose. Low dose imaging (10 electrons/Å^2) results in very low contrast electron micrographs,but is necessary in order to minimise beam damage.

Negatively stained images are obtained by dispensing the protein sample onto a thin carbon layer supported

on an electron microscope grid. The bound protein is then washed with buffer prior to the application of a heavy metal stain such as Uranyl acetate. On blotting away the excess stain the remaining solution dries down to form an electron dense meniscus around the protein molecule. The molecular imprint is then imaged at room temperature under conditions that enhance the SNR compared to those obtained by cryo-electron microscopy.

Ferritin, which was used as a test data set, is a protein complex involved in binding iron in a wide range of organisms [4]. In insects ferritin plays an essential role in maintaining iron homeostasis making it of medical and agricultural importance as a potential target against insect disease vectors such as malaria and agricultural pests. Insect ferritin was selected as a test protein as it is smaller than viruses, and allows us to test whether atomic resolution structures can be resolved using single particle analysis. Furthermore crystals diffracting to 2.4Å have been generated providing an independent control for any 3D structures determined.

2 A Correlation-Based Particle Picking Algorithm

A real-space correlation-based particle picking algorithm has been developed. This method was chosen since it can use a normalised correlation function and perform local masking[5]. It can therefore use the rotationally averaged template and mask constructed from the image processing software *IMAGIC* [6].

Prior to running the algorithm, a rotationally averaged particle sum template and a mask are constructed, using the IMAGIC software. The template is constructed by manually selecting a small number of particles, performing translational alignment, averaging, and then rotationally averaging to obtain a circular, symmetric template. The mask is the same size as the template, and has a value 1 where the template data is valid, and 0 otherwise.

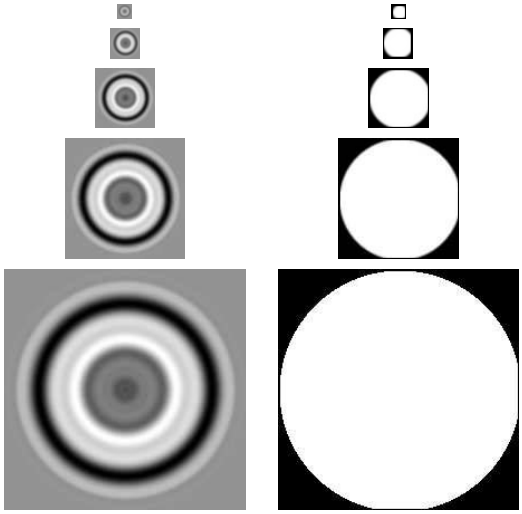
The steps involved in the automatic particle picking algorithm are described as follows:

2.1 Construction of Image Pyramids

The micrographs are generally quite large, of the order of $\sim$ 100 Mb. Therefore, to increase the algorithm's efficiency, an *image pyramid*[7] is constructed. To prevent aliasing, at each level the images are filtered with the Gaussian filter of Equation 1, before being subsampled by a factor of two.

$$G = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

In this manner, the image is progressively halved in size, until one of the image dimensions is less than 1000.



(a) (b)

Figure 1: Image pyramids for the (a) template and (b) mask images, constructed from the ferritin data set.

Image pyramids are also constructed for the template and mask images, with the same number of levels as the micrograph. Figure 1 shows the 5-level image pyramids constructed for the template and mask for the ferritin data.

The original mask is a binary image consisting only of the values 0 and 1. However, the construction of the pyramid of mask images smooths the pixel values with the Gaussian filter of Equation (1). This results in pixel values between 0 and 1, particularly around the edges of the mask. Therefore, the mask images can be thought of as *weight* values which scale the contribution of each pixel to the correlation computations.

2.2 Correlation

Computation begins with the lowest resolution (ie, smallest) image, template and mask. The *Normalised Cross Correlation* (NCC) score is computed at each image location (x,y) using Equation (2), resulting in a 2-D array of correlation scores called a *correlation image*.

2.3 Selection of Maxima

Locations where the NCC score is locally maximal are flagged as potential particles. At this stage, there are often a large number of spurious maxima which do not correspond to particles.

2.4 Filtering of Maxima

This step determines which of the local maxima actually correspond to particles, by evaluating the shape of the correlation surface in the vicinity of each maxima.

It was observed that for true particles, the correlation surface consists of a peak surrounded by a trough,

$$NCC(x,y) = \frac{\sum_{(i,j)\in W}^{W} I(x+i,y+j)T(i,j)M(i,j)}{\sqrt{\sum_{(i,j)\in W}^{W} I^2(x+i,y+j)M(i,j)}\sqrt{\sum_{(i,j)\in W}^{W} T^2(i,j)M(i,j)}} \quad (2)$$

where I, T and M are the image, template and mask, W is the correlation window, and $(i,j) \in W$ indicates that a pixel lies within the correlation window.

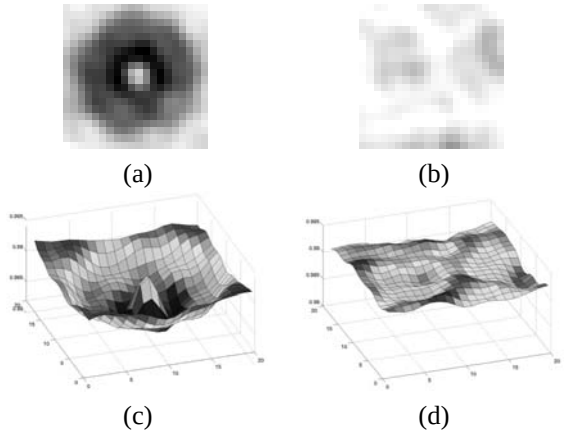


Figure 2: Correlation image and shape of the correlation surface for the ferritin data: (a),(c) in the vicinity of a particle, (b),(d) around a spurious maxima.

while for spurious maxima, the correlation values are more or less flat, as shown in figure 2.

A recursive region-growing algorithm is used to identify valid particles. This algorithm starts with local maxima as seed points and grows outwards in (x,y). For a particle to be valid, the correlation values must drop to a certain value below the seed point, within a given radius range, [*min_radius* – *max_radius*]. If the correlation function drops below this given value before *min_radius* is reached, or still hasn't dropped below the given value when *max_radius* is reached, then the maxima is removed from the set of possible particles.

Once a set of valid particles have been identified, clusters of overlapping points are removed.

2.5 Propagating points through the Image Pyramid

The previous steps compute a set of valid particles using the lowest resolution level of the image pyramid. These points need to be propagated up through the image pyramid through to the full resolution image. This is achieved as follows:

2.5.1 Scaling points to the next level

The particle coordinates are multiplied by two to scale them up to the next higher resolution level of the pyramid.

2.5.2 Refining particle locations

The accuracy of the scaled up particle coordinates is refined by computing the NCC in a small neighbourhood around each point, using the image, template and mask at the current pyramid level. The coordinates of each particle are adjusted to the location of the nearest NCC maxima. If no maxima is present within a close neighbourhood, the point is removed from the set of valid particles.

2.6 Writing output files

The particle locations are output as a ".plt" file (a text format read by the IMAGIC software).

For display purposes, the software can also save the original image with markers super-imposed at particle locations. The user can specify either circular or square markers, as well as the marker size.

3 Implementation and Results

The algorithm has been implemented in C++ on a 64 processor SGI Origin machine running IRIX. Parallel programming constructs were used to take advantage of the availability of the multiple processors.

The particle picking software has been tested on a dataset of negatively stained ferritin images. Figure 3 shows an example of an image from the ferritin series and the obtained results. The software appeared to select most of the particles, and returned few false positives. It was also able to discriminate between single and aggregated particles.

Cryo images are typically of lower SNR and pose more of a challenge to automatic particle picking algorithms. Figure 4 shows a cryo image of a virus and the obtained results. Despite the low contrast of the image, a large number of particles were detected.

A successful particle picking algorithm must achieve the following goals:

- select a high proportion ($\geq$ 90%) of particles present in a micrograph
- keep the number of false picks as low as possible.

The developed algorithm appears to have achieved these goals with the tested negatively stained ferritin and cryo virus images.

Ground truth results would typically be obtained by a person manually selecting particles in a micrograph. The results of particle picking algorithms could then

compared with the manually selected points. However, this may still not yield a true indication of an algorithm's success, since there is often a wide discrepancy in the particles selected by different people, especially for low contrast images.

A solution to this problem could be that a large number of people manually select particles in the test images, and a "certainty" score computed for each particle, based on the proportion of people who selected that particle.

More "weight" would then be given to an algorithm's success if it selected mostly particles with a higher certainty score, rather than obscure or dubious particles which may be false. In addition if an algorithm selects, for example, 100% of points with a certainty score of 1.0, and say 50% of points with a certainty score of 0.5, then the algorithm has performed as well as a human particle picker.

Currently, discussions are underway among particle picking groups worldwide about constructing ground truth datasets which will then be made available [8].

The software input currently is run from the command line and reads the required parameter values from a text file. These parameters include:

Path to image pyramid: If an image pyramid was already constructed by a previous run of the program, then it can be read in from saved files, reducing computation time. If the image pyramid has not yet been constructed this parameter is set to "none".

Correlation threshold: Maxima having correlation scores below this threshold are not considered as possible particle locations.

Local maxima radius: Radius used to test if a point's correlation score is locally maximal.

Grow threshold: Minimum difference between a peak and its surrounding trough in the correlation score.

Min and max radius: Minimum and maximum radius of the peaks in the correlation score.

4 Conclusions and Further Work

A correlation-based algorithm for automatic particle picking in electron micrograph images has been developed. Despite this work being in its early stages, results are encouraging. The algorithm has been shown to be capable of selecting most of the particles, with few false positives, with both negatively stained and cryo images of virus particles.

The algorithm will be extended to be able to select particles differing widely in size, shape and symmetry, as new types of proteins and viruses are analysed. In particular, routines will have to be developed to handle non-symmetric particles.

Some techniques which will be considered in order to extend the software include:

- constructing a global template for correlation by rotationally averaging all possible particle views and orientations.
- using edge detection techniques to complement the existing correlation routines.

In addition, to make the software more user-friendly, a user interface which will enable users to adjust parameter values will be developed.

5 Acknowledgements

This project was conducted with funding and facilities provided by the Queensland Parallel Supercomputing Foundation.

We would like to thank Paul Young and Chang Yi Huang for providing the ferritin samples.

References

[1] D. Jones. Do transmembrane protein superfolds exist? *Federation of European Biochemical Societies (FEBS) Letters*, 423:281–285, 1998.

[2] M. van Heel, B. Gowen, R. Matadeen, E. Orlova, R. Finn, T. Pape, D. Cohn, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution. *Quaterly Reviews of Biophysics*, 33(4):307–369, 2000.

[3] Z. Zhou, M. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Liu, and W. Chiu. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Structural Biology*, 8(10):868–873, 2001.

[4] Y. Ha, D. Shi, G. Small, E. Theil, and N. Allewell. Crystal structure of bullfrog M ferritin at 2.8 åresolution: analysis of subunit interactions and the binuclear metal center. *Journal of Biological Inorganic Chemistry*, 4(3):243–256, 1999.

[5] A. Roseman. Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy*, 94:225–236, 2003.

[6] Image Science. IMAGIC-5 advanced scientific image processing. http://www.imagescience.de/imagic/welcome.htm. visited on 20/08/2003.

[7] T. Lindeberg. *Scale-space theory in computer vision.* Kluwer, 1994.

[8] Center for Integrative Molecular Biosciences, The Scripps Institute. *MultiDisciplinary Workshop on Automatic Particle Selection for Cryo-Electron Microscopy*, La Jolla, California, Apr 2003.

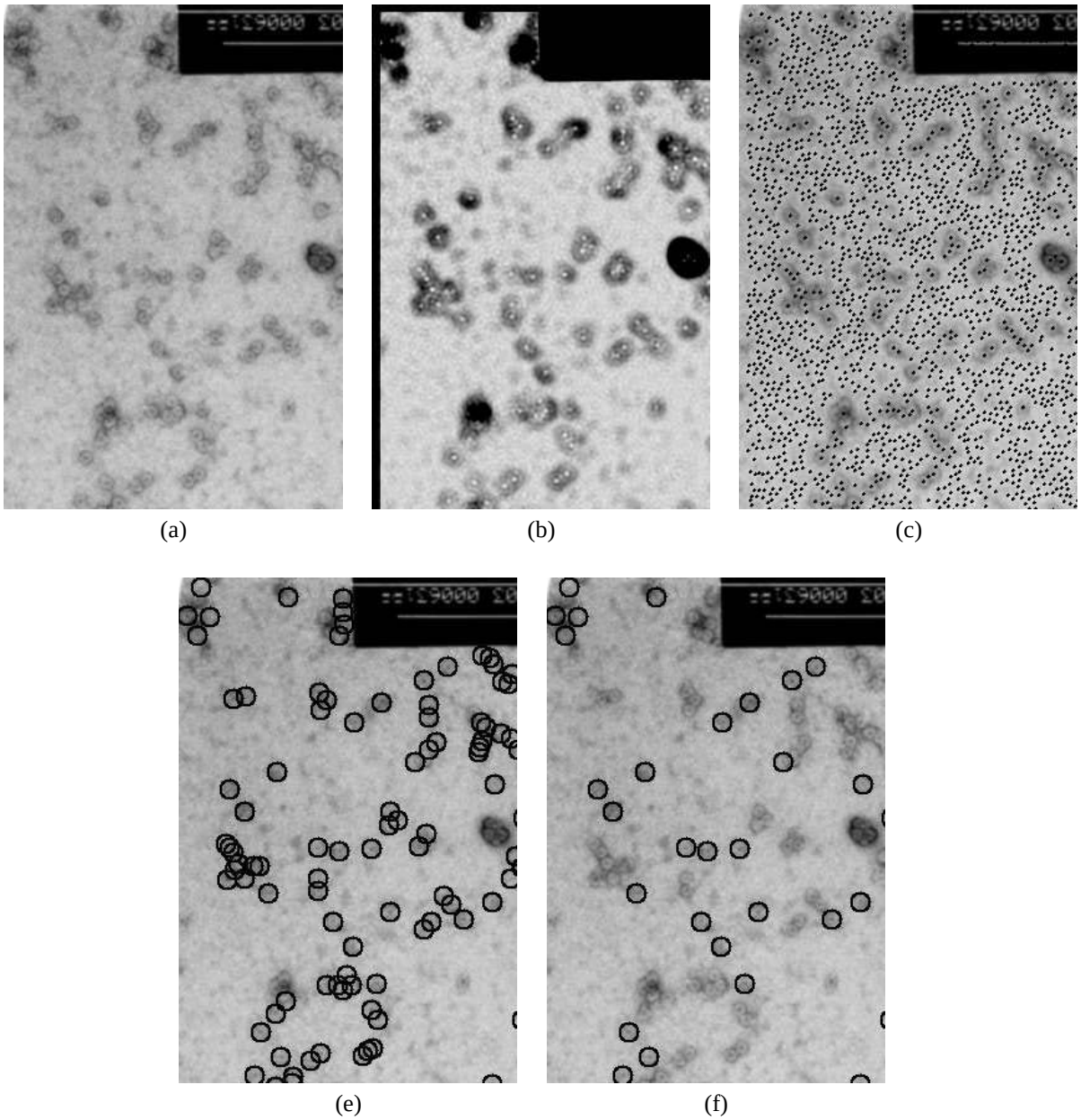


Figure 3: Particle picking results for an image from the ferritin data set: (a) portion of a negatively stained micrograph (b) normalised cross correlation image (c) local maxima (d) selected particles after filtering (e) selected particles after clusters are removed.

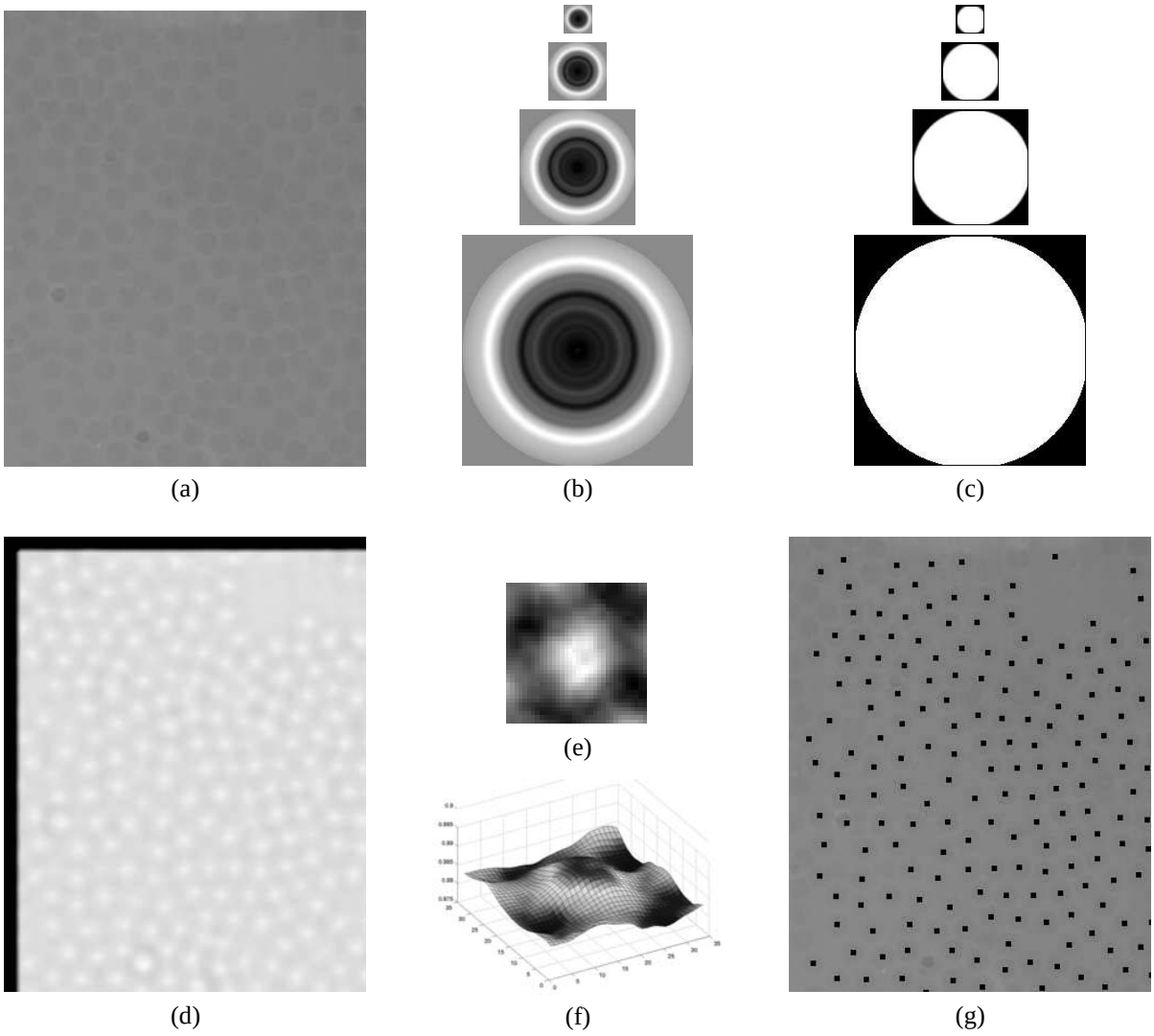


Figure 4: Particle picking results for the virus image: (a) portion of a cryo micrograph (b) 4-level template pyramid (c) 4-level mask pyramid (d) normalised cross correlation image (e) correlation image in the vicinity of a particle (f) correlation surface in the vicinity of a particle (g) selected particles after filtering and local clusters are removed. The correlation surface in the vicinity of a particle is different to that of the ferritin images in that it forms a wide peak.