Automatic Particle Picking Algorithms for High Resolution Single Particle Analysis

Jasmine Banks, Bernard Pailthorpe
Advanced Computational Modelling Centre
University of Queensland
Brisbane 4072, Australia.
{jbanks, bap}@acmc.uq.edu.au

Rosalba Rothnagel, Ben Hankamer
Institute for Molecular Biosciences
University of Queensland
Brisbane 4072, Australia.
{r.rothnagel, b.hankamer}@imb.uq.edu.au

Abstract

As new genome sequencing initiatives are completed, one of the next great challenges of cell biology is the atomic resolution structure determination of the enormous number of proteins they encode. Single particle analysis is a technique which produces 3D structures by computationally aligning high resolution electron microscope images of individual, randomly oriented molecules. One of the limiting factors in producing a high resolution 3D reconstruction is obtaining a large enough representative dataset (∼100,000 particles). Traditionally particles have been picked manually but this is a slow and labour intensive process.

This paper describes two automatic particle picking algorithms, based on correlation and edge detection, which have been shown to be capable of quickly selecting a large number of particles in micrographs. Currently circular and rectangular particles are able to be picked.

1. Introduction

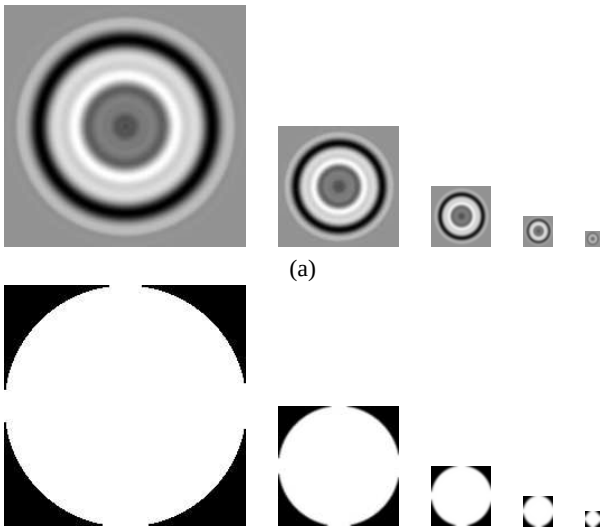
One of the next great challenges of cell biology is the atomic resolution structure determination of the enormous number of proteins encoded in genomes. To date, the Protein Information Resource contains ∼1.9 million protein sequences[10]. This number is increasing rapidly as new genome sequencing initiatives are completed. The human genome project alone identified ∼30,000 genes encoding both soluble and membrane proteins. *In vivo* these organise into macromolecular assemblies, further increasing the level of structural complexity.

Membrane proteins, which are predicted to comprise 25–40% of all encoded proteins[5], form the responsive interface between the cellular and sub-cellular compartments and the outside environment. Their structures are not only of fundamental importance in developing our understanding of molecular cell biology, but are also of immense value in the development of new and highly specific medicines with reduced side effects. In addition, the huge number of macromolecular assemblies are only beginning to be characterised structurally. Consequently, fast-tracking structure determination of membrane proteins, soluble proteins and macromolecular assemblies will underpin future developments in cell biology, structural biology, and proteomics.

Traditionally, protein structures have been solved using crystallography techniques. However, particularly in the case of membrane proteins, the production of well-ordered crystals is a major bottleneck. Therefore, despite their importance, only a small number (80–90) of complete membrane protein structures have been resolved to atomic resolution.

Recent advances in cryo-electron microscopy and single particle analysis have developed to the point where they could potentially provide an alternative methodology for high resolution 3D structure determination[9]. *Cryo-electron microscopy* involves suspending the purified protein molecules in a thin layer of vitreous ice. The suspended particles are imaged in the electron microscope at temperatures of $-170°$C with a low electron dose. Low dose imaging results in very low contrast micrographs, but is necessary to reduce beam damage. The technique of *single particle analysis* produces 3D structures by computationally aligning high resolution electron microscope images of individual, randomly oriented molecules. Modern cryo-electron microscopes are capable of recording structural information to a resolution higher than 2Å(1Å=10^{-10}m). To sample the 3D volume fully at the required resolution, and overcome the low signal-to-noise ratio (SNR) of the images, a large dataset (∼100,000 particles) is required. Particles have been picked manually but this is slow and labour intensive (∼1 week for 20,000 particles) and difficult due to the low SNR of the images.

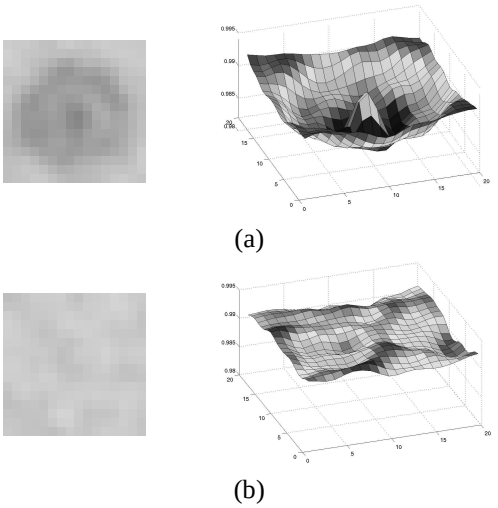
This paper describes two automatic particle picking algorithms, based on correlation and edge detection. The algorithms have been tested with both negatively stained



(a)

(b)

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



(a)

(b)

Figure 2. example of pixel data and shape of the correlation surface: (a) in the vicinity of a particle (b) around a spurious maxima, from the ferritin data set.

(high contrast) and cryo (low contrast) micrographs.

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 local masking[6].

A rotationally averaged particle sum and a binary mask were constructed, using the IMAGIC software[4]. The template was constructed by manually selecting a number of particles, performing translational alignment, averaging, and then rotationally averaging to obtain a circular, symmetric template. The constructed mask is the same size as the template, and has the value 255 where the template data is valid, and 0 otherwise.

2.1. Construction of Image Pyramids

The micrographs are sampled finely ($\sim$0.9Å per pixel), consequently the digitised images are generally quite large, for example, the test dataset images of the protein ferritin are of size 8718×13071 pixels, with a template of size 216×216 pixels. The amount of computation can be dramatically reduced by performing particle picking using a lower resolution image, template and mask. Therefore, image pyramids are constructed, where each level is constructed by smoothing the previous level with a Gaussian filter (to prevent aliasing), and then sub-sampling by a factor of two. In this manner, the micrograph image dimensions are progressively halved until one of the image dimensions is less than 1000 pixels.

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

The original full-sized mask is a binary image consisting only of the values 0 and 255. However, the construction of the pyramid smooths the pixel values, resulting in pixel values between 0 and 255, 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 (1), resulting in a 2-D array of scores called a *correlation image*.

2.3. Selection of Maxima

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

2.4. Filtering of Maxima

This step determines which of the local maxima correspond to particles, by examining the shape of the correlation surface in the vicinity of each maxima. It was observed that for particles, the correlation surface consists of a peak surrounded by a trough, 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 max-

$$\mathrm{NCC}(x,y) = \frac{\sum_{(i,j)\in W}^{W} I(x+i,y+j)T(i,j)M(i,j)}{\sqrt{\sum_{(i,j)\in W}^{W} I^2(x+i,y+j)M(i,j)\sum_{(i,j)\in W}^{W} T^2(i,j)M(i,j)}} \quad (1)$$

where NCC = Normalised Cross Correlation score, (x,y) = image location, I = image, T = template, and M = mask, and (i,j) are indicies into the pixel window, W.

ima at locations (x,y) as seed points and then grows outwards in an 8-connected manner[3]. For a particle to be valid, the correlation values must drop a certain value below the seed point, $intensity\_drop$, within a given radius range, min_radius to max_radius. If the correlation function drops more than $intensity\_drop$ before min_radius is reached, still hasn't dropped by $intensity\_drop$ when max_radius is reached, for every point around the centre, then the location is removed from the set of possible particles.

Once a set of valid particles has been identified, distance between particle centres are computed, and clusters of overlapping particles removed.

2.5. Propagating Particles to the Highest Resolution

The previous steps identify a set of particles using the lowest resolution level of the pyramid. These locations may be propagated up through the image pyramid to the full resolution image. This is a two step process. First, the particle coordinates are multiplied by two to scale them up to the next higher resolution level of the pyramid.

Next, the accuracy of the scaled up particle locations is improved 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 then adjusted to the coordinates of the nearest NCC maxima. If no maxima is present within a close neighbourhood, the point is removed from the set of valid particles.

The process is repeated until the particle coordinates are propagated up to the highest resolution image.

3. An Edge-Based Particle Picking Algorithm

Edge detection based particle picking algorithms first perform edge detection on the micrographs, then locate particle shapes in the edge image.

3.1. Pyramid Generation

To reduce the amount of computation required, an image pyramid is constructed for the micrograph image, in a similar manner as for the correlation algorithm.

3.2. Edge Detection

Edge detection algorithms are applied to the lowest level of the image pyramid. Both the Laplacian of Gaussian (LOG) and Canny edge detectors have been implemented. [2, 3]. The output of the edge detection stage consists of a 2D binary edge image, where "1" denotes the presence of an edge. The Canny edge detector additionally outputs an an edge direction image.

3.3. Particle Selection in Edge Images

Next, the edge image needs to be interpreted to find edge arrangements that correspond to particles.

3.3.1 Contour Following. The first technique implemented involved following edge contours to determine if they are roughly circular in shape. This is most suited to the unbroken contours produced by the LOG algorithm.

A recursive region growing algorithm is used to follow connected edge pixels. When an edge pixel is encountered, the edge is followed by growing outwards in an 8-connected manner. Once a pixel has been visited, it is flagged as already belonging to a contour, so that it is not processed again. The edge following process determines the extent of the contour, and estimates the centre of a particle it may represent by averaging the (x,y) coordinates of all edge pixels it comprises. If a contour's extent is greater than a valid particle size, or if it touches an image border, it is removed from further consideration.

Next, it is determined whether the contour is roughly circular. A simple test used is to estimate the minimum and maximum radii, min_r and max_r, and to compute the eccentricity, $e = min\_r/max\_r$. A value of e close to 1.0 indicates a close to circular shape, while a small e indicates a highly elliptical shape. If min_r, max_r and e all fall within given limits, then the contour is accepted as representing a circular particle.

3.3.2 The Hough Transform. Hough transform based techniques[3] are better suited to situations where edges denoting a particle shape may be fragmented into several contours. A parameter space called the accumulator array is used, where the number of dimensions equals the number of parameters defining the particles. Every location in the accumulator array is initialised to zero. Each edge pixel increments locations in the accumulator array, corresponding to sets of particle parameters, for all particles which this edge pixel could possibly belong to. After all edge pixels have been processed, local maxima in the accumulator array indicate likely sets of parameters corresponding to particles.

Circle detection using the Hough transform requires a three dimensional accumulator array, in which the dimensions correspond to the radius, r and the centre (a,b) of

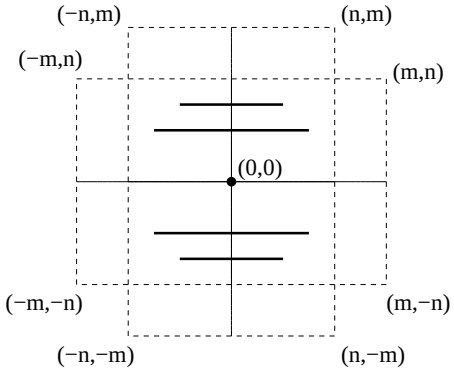


Figure 3. Solid lines indicate possible rectangle centre locations, for an edge pixel located at the origin. Dashed lines indicate particle extents.

circles. Given an edge pixel at location (x, y), all possible (a, b, r) configurations are computed from the equation for a circle, $r^2 = (x - a)^2 + (y - b)^2$, and these locations in the accumulator array incremented. The size of the accumulator array and amount of computation required can be reduced by considering only radii in the possible range for particles. At the completion of the Hough transform process, local maxima in the accumulator array indicate the parameters (a, b, r) of detected circles, where (a, b) are the particle coordinates.

Rectangle detection was based on a modified version of the Hough transform[11]. A 4 dimensional accumulator array was used, where the dimensions are centre location (a, b), and rectangle width w and height h. As the number of dimensions of the accumulator array increases, the amount of computation required increases considerably. However this can be kept to a minimum if the variations in w and h are small.

Given an edge pixel (x, y), all possible centre locations for this pixel, as shown in Figure 3, are incremented in the accumulator array. The shape also needs to be rotated by the edge orientation, which is obtained as an output of the Canny edge detection process.

Combined circle and rectangle detection has been implemented for images containing both circles and rectangles. The first stage of the process detects circles. The edge pixels comprising the circles then need not be considered for rectangle detection, thus saving processing time. Furthermore, centres of rectangular particles cannot occur within a distance of min_radius from the circle edges, therefore these regions can also be removed from consideration as possible rectangle locations.

3.4. Propagating Particles to Highest Resolution

As with the correlation algorithm, the particle coordinates may be propagated up to the highest resolution image level of the pyramid. This is again a two step process. Particle coordinates are first of all multiplied by two to scale them up to the next level of the pyramid. In the next higher resolution image, edge detection and particle identification only need be performed in a small neighbourhood around each particle.

The process may be repeated until the particle coordinates are propagated up to the highest image.

4. Particle Picking Results and Discussion

The algorithms were initially tested with a set of negative stained ferritin images. Figure 4 shows a region from one image, and particles picked using the correlation and edge detection algorithms. Figure 5 shows results obtained with a test cryo image of a virus. Cryo images tend to be more of a challenge than negatively stained images due to the reduced contrast.

Testing was also carried out using a test data set of Keyhole Limpet Hemocyanin (KLH)[7]. The particles are cylindrical in shape, resulting in circular and rectangular views of the particle in the micrographs. Figure 6 shows the results of particle picking using both correlation, and edge detection followed by the combined Hough circle and rectangle detection method.

The algorithms were shown to be capable of selecting a large number of particles in micrographs, with few false positives. For structural biologists to make use of these algorithms, a suitable interface needs to be developed. A Graphical User Interface (GUI) has been developed for the correlation algorithm. The GUI has been implemented in C++ using wxWindows, and assists with parameter selection, display of results, and allows a small number of missed/erroneous particles to be added/deleted. Using this software with test data sets, it was possible to select a large number of particles in a few hours, which would have formerly taken weeks of work.

The edge detection algorithm will also need to be incorporated into this user interface. Furthermore, particle detection algorithms will also need to be written to detect differently shaped and oriented particles. One method could be to use the generic Hough transform which could potentially detect a wide variety of particles based on a reference-table for each particle shape silhouette[1], or it may be possible to use techniques such as neural networks.

The particle coordinates are output in a form designed to be input in to the IMAGIC package. The IMAGIC software is then used to align particles, compute class sums, determine their orientation, and produce the final 3D model of the protein molecule.

The presented algorithms locate particles in a low resolution image and then propagate them to the highest resolution image. In many cases, the extra computation involved in accurately propagating the particles to the high resolution

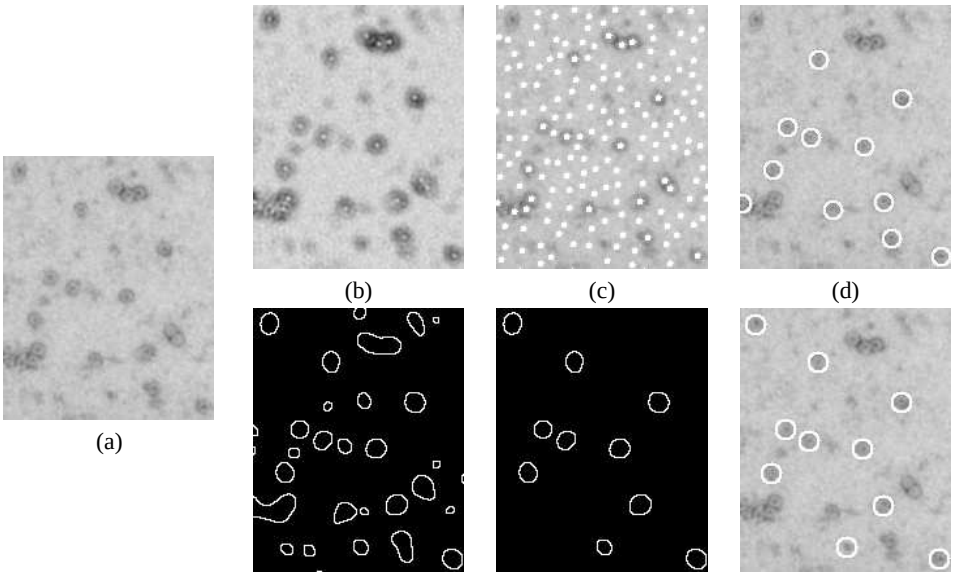


Figure 4. Results obtained using negatively stained ferritin: (a) small section of micrograph (b) correlation scores (c) correlation peaks (d) picked particles using the correlation algorithm (e) edge detection using the LOG filter (f) contours corresponding to particles (g) picked particles using the edge detection algorithm.

image may be unnecessary. This is because the IMAGIC software, which is designed to work with particles picked by a human, includes a particle alignment procedure.

5. Conclusions

Automatic particle detection in electron micrographs will be an important component of a high-throughput pipeline to fast track 3D structure determination of membrane proteins and macromolecular assemblies.

Further work will include extending the user interface to incorporate the edge detection algorithm, and extending the particle picking algorithms to detect differently shaped and oriented particles. Techniques for noise removal need to be considered. One such technique is the bilateral filter. This non-linear filter can smooth noise while preserving edge features[8].

At present, cryo electron micrographs of the test protein ferritin are being imaged. Successful particle picking and 3D reconstruction from this data will prove the concept that protein structures can be determined to atomic resolution using cryo electron microscopy and single particle analysis.

6. Acknowledgements

This project was conducted with funding and facilities provided by the Queensland Parallel Supercomputing Foundation, and is funded by the Australian Research Council Discovery Grant, “High resolution single particle analysis of biological macromolecules”.

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

References

[1] D. Ballard and C. Brown. *Computer Vision*. Prentice Hall, 1982.

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 679–698, 1986.

[3] R. Gonzalez and R. Woods. *Digital Image Processing*. Addison Wesley, 1992.

[4] Image Science. imagic-5 image processing. http://www.imagescience.de/imagic/.

[5] A. Kanapin, *et al.* Mouse proteome analysis. *Genome Resource*, 13:1335–1344, 2003.

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

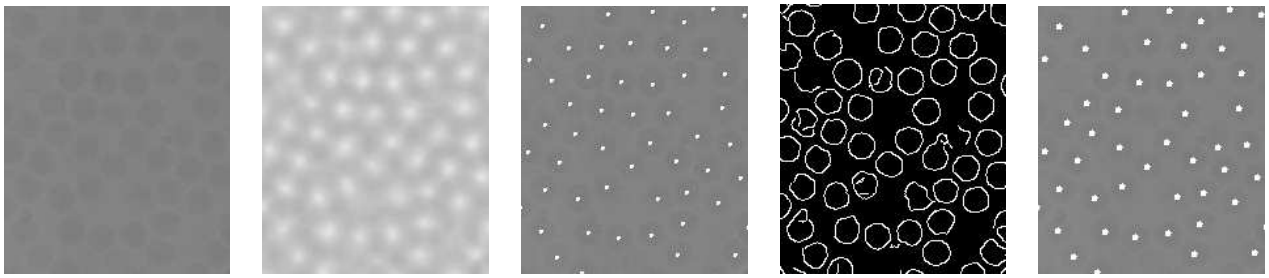
[7] The Scripps Institute. Annotated image datasets. http://ami.scripps.edu/prtl_data/.

[8] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *IEEE International Conference in Computer Vision*, pages 59–66, 1998.

[9] M. van Heel, *et al.* Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics*, 33(4):307–369, 2000.

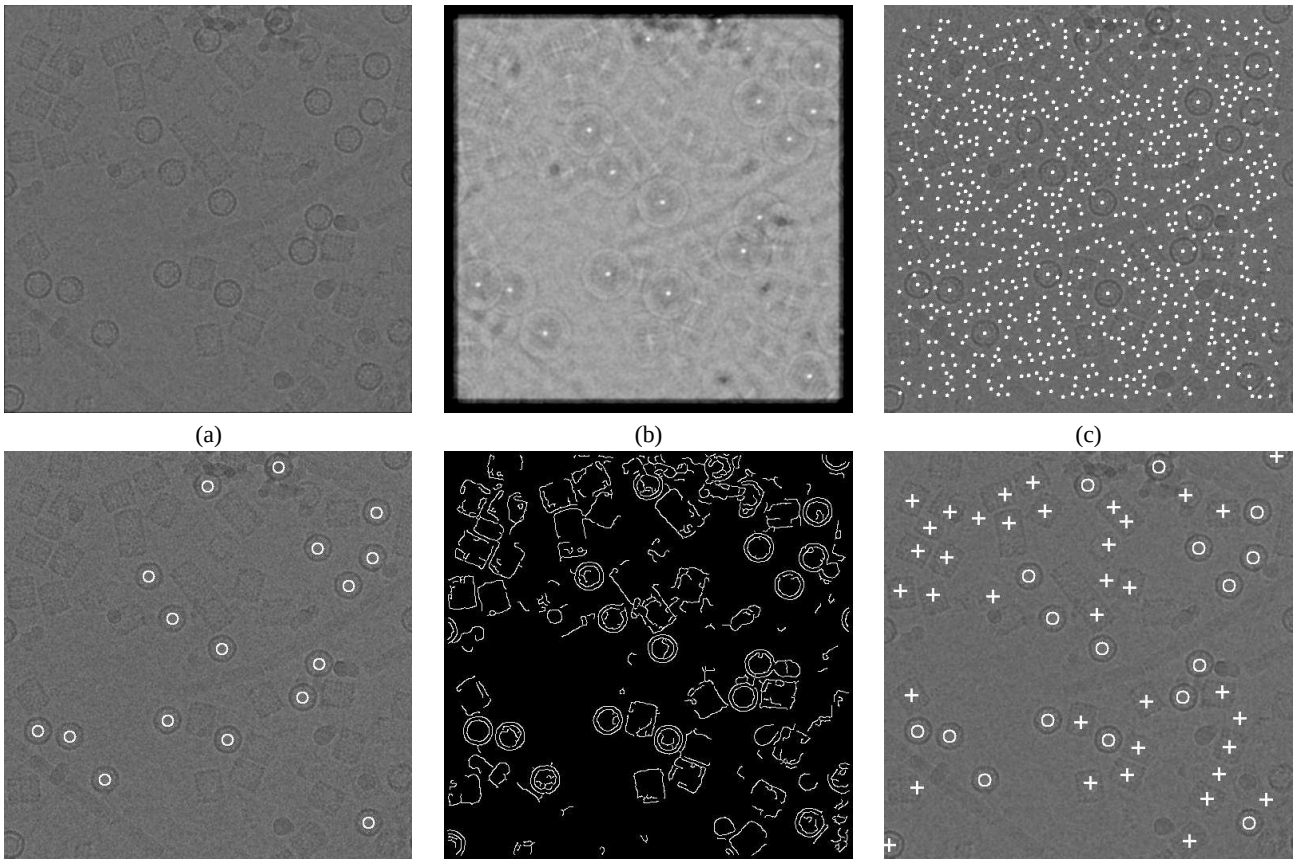
[10] C. Wu, *et al.* The protein information resource. *Nucleic Acids Resource*, pages 345–347, 2003.

[11] Y. Zhu, B. Carragher, F. Mouche, and C. Potter. Automatic particle detection through efficient hough transforms. *IEEE Trans. on Medical Imaging*, 22(9):1053–1062, 2003.



(a) (b) (c) (d) (e)

Figure 5. Results obtained using test cryo micrograph of a virus: (a) small region of virus image (b) correlation scores (c) picked particles using the correlation algorithm (d) edge detection using the Canny edge detector (e) contours corresponding to particles (f) picked particles using edge detection followed by circle detection with the Hough transform.



(a) (b) (c)

(d) (e) (f)

Figure 6. Results obtained using Keyhole Limpet Hemocyanin dataset: (a) micrograph (b) correlation scores (c) maxima in correlation array (d) picked particles using the correlation algorithm (e) edge detection using Canny edge detector (f) picked particles using edge detection followed by combined circle and rectangle detection with the Hough transform.