Identifying Components in 3D Density Maps of Protein Nanomachines by Multi-scale Segmentation

Grigore Pintilie*
Electrical Engineering and Computer Science
MIT
pintilie@mit.edu

Junjie Zhang
Structural & Computational Biology and Molecular Biophysics
Baylor College of Medicine
jz147980@bcm.edu

Wah Chiu
Structural & Computational Biology and Molecular Biophysics
Baylor College of Medicine
wah@bcm.edu

David Gossard
Department of Mechanical Engineering
MIT
gossard@mit.edu

***Abstract*— Segmentation of density maps obtained using cryo-electron microscopy (cryo-EM) is a challenging task, and is typically accomplished by time-intensive interactive methods. The goal of segmentation is to identify the regions inside the density map that correspond to individual components. We present a multi-scale segmentation method for accomplishing this task that requires very little user interaction. The method uses the concept of scale space, which is created by convolution of the input density map with a Gaussian filter. The latter process smoothes the density map. The standard deviation of the Gaussian filter is varied, with smaller values corresponding to finer scales and larger values to coarser scales. Each of the maps at different scales is segmented using the watershed method, which is very efficient, completely automatic, and does not require the specification of seed points. Some detail is lost in the smoothing process. A sharpening process reintroduces detail into the segmentation at the coarsest scale by using the segmentations at the finer scales. We apply the method to simulated density maps, where the exact segmentation (or ground truth) is known, and rigorously evaluate the accuracy of the resulting segmentations.**

I. INTRODUCTION

Cryo-electron microscopy methods yield detailed three-dimensional (3D) density maps of protein nanomachines [1]. The nanomachines consist of multiple components, which are proteins that are in contact with one another. A main task in the analysis of such density maps is to identify the locations and shapes of these proteins, which have sizes on the order of nanometers. This information can give much insight into how nanomachines perform their diverse functions as part of important life processes.

Two ways of identifying the components within a density map is by alignment of known protein structures to the map, for example by template matching, or by segmentation. In this work we focus on segmentation, which assumes no prior knowledge about the structure of the proteins to be detected.

Segmentation has been a widely-studied problem in the field of computer vision, and is in general a hard problem. To make it more amenable, many methods require the input of seed points or contours for each object to be detected. The final segmenting contours are then determined by imposing certain conditions on the resulting shape (e.g. smoothness) while using information in the image being segmented (e.g. pushing the contour towards high gradients). An example of such a method applied to density maps is the level set method [2]. While such methods can produce accurate segmentations, they rely heavily on user input. As such the methods tend to be time intensive, impractical for very large data sets, and subjective. Hence in this work we focus on segmentation methods that do not require the specification of seed points.

The most basic segmentation method that does not require seed points is thresholding [3]. This method isolates regions containing intensity values that are above a chosen threshold. It works well when the objects to be detected contain intensities that stand out against the background. Also, different objects must be well-separated from one another to be detected separately. For density maps of protein nanomachines, this method normally segments the entire nanomachine as a whole, without identifying the components within it. Hence, more sophisticated methods are needed to segment the protein components in such maps.

Another segmentation method that does not require seed points is the watershed method [4]. The map or image to be segmented is taken as a landscape, with the height proportional to the density values throughout the map. Segmented regions center around local maxima, and all points in a region lead to the same local maximum when following the gradient of the density function. In maps with a lot of detail and/or noise, many local maxima are present, and thus the segmentations include many regions. This effect is often referred to as oversegmentation. To address this issue, several approaches have been proposed, for example hierarchical merging of regions [5,6]. These methods introduce extra parameters, which are hard to tune, and their influence on the segmentation accuracy is hard to evaluate in general. The watershed method has already been applied to density maps to segment out protein components with good results [7]. In the latter work, the issue of oversegmentation was addressed by using a variable step size in the segmentation process. However the effect of this parameter on the segmentation accuracy in relation to its effectiveness at reducing oversegmentation was not analyzed.

To address the issue of oversegmentation, we use smoothing of the density maps by convolution with a Gaussian filter. This process can greatly reduce the number of local maxima

* This work was supported by NIH grants (PN2EY016525, R01GM079429, P41RR02250) and NSF IIS-0705474.

in the density map. This observation has also been made before in the multi-scale approach for the detection of edges in images [8,9], as well as in the application of the mean-shift method; the latter also typically uses a Gaussian kernel for segmentation of an image [10-12]. In the multi-scale approach, a scale space is obtained by applying a Gaussian filter of increasing widths. The scales range from fine, where the standard deviation of the Gaussian filter is small, to coarse, where the standard deviation of the Gaussian filter is large. The edges that persist throughout the scale space were shown to identify salient edges in the images [9].

In the same spirit, we use the multi-scale approach for the segmentation of density maps. At coarser scales, watershed segmentation yields fewer regions, which in some cases can correspond to individual protein components. However, as a result of the greater degree of smoothing at these coarser scales, the boundaries between regions are distorted and hence less accurate. We use a sharpening operation to increase detail of the segmentation at these coarse scales.

II. METHOD

A. Simulation of density maps

We demonstrate our segmentation method on simulated density maps. A map is simulated using a high-resolution structure obtained from the Protein Data Bank (PDB). The atomic coordinates are first embedded onto a 3D grid, and then the density values on the grid are smoothed by convolution with a Gaussian filter [13]. The standard deviation of the Gaussian filter is specified so as to achieve a desired resolution, which describes how much detail the map contains. We report the resolutions of our simulated maps using the formula $s = 0.187r$, where r is the desired resolution and s is the standard deviation. The latter equation is such that the Fourier transform of the Gaussian filter falls to half its maximum value at the wavenumber $1/r$. This is related to the $FSC_{0.5}$ criterion, which is normally used to determine the resolution of an experimental density map [14].

B. Creating the scale space

A scale space is created for each input density map by further convolution of the map with a Gaussian filter. The scale space parameter, σ, refers to the standard deviation of the Gaussian filter. The resulting maps will represent fine through coarse scales for low to high σ respectively.

C. Segmentation

We use the watershed method described in [15] to segment a density map. This method is very efficient because it only considers each voxel a constant number of times. It involves a sorting stage, where the voxels are sorted by density value, and hence the computational complexity is $O(nlogn)$ where n is the number of voxels in the map. This topological approach was also used for the efficient implementation of the mean-shift segmentation method [16].

For the segmentation of natural images, the height of the landscape is usually taken to be proportional to gradient magnitudes, so as to yield boundaries that fall on high gradient magnitudes. For density maps of protein nanomachines, we take the height to be proportional to the density values, as proposed previously in [7], which yields better results. The density values are considered in decreasing order, since higher densities correspond to the objects to be detected. Thus the centers of each region will fall on high-density values, and the boundaries between the regions will fall on lower density values.

The density maps to be segmented, either simulated or experimental, typically contain non-zero density values outside of the imaged nanomachines. To avoid creating segmentation regions outside the nanomachines, we only consider the density values above a certain threshold. This threshold can be determined by visual inspection, using a visualization program that displays iso-surfaces of a density map interactively, e.g. [17]. The density value used to build the iso-surface can be adjusted until the entire nanomachine appears to be included within the iso-surface, and this density value can be used as the threshold. Alternatively, the threshold can be adaptively chosen if the approximate volume of the entire nanomachine is known, by stopping the segmentation once the combined volume of all the segmentation regions reaches this known volume.

D. Evaluation of segmentation accuracy

In recent work [18], a rigorous measure of segmentation accuracy was used, which we adopt here. The segmentation accuracy is computed using the following formula:

$$|SegmentationAccuracy| = \frac{volume(S \cap P)}{volume(S \cup P)} \qquad (1)$$

In the above equation, $volume(S \cap P)$ is the volume of the intersection of a segmentation region, S, and the actual region occupied by a single protein, P; $volume(S \cup P)$ is the volume of the union of S and P. The segmentation accuracy will be 0 if the segmentation and protein do not overlap at all (the intersection will have 0 volume), and it will be 1 if the segmentation region and the protein overlap exactly (the volumes of the intersection and the union will be the same).

E. The ground truth in simulated maps

In (1), P refers to the region in a density maps occupied by a single protein, this being the ground truth. To obtain this, a density map is simulated for each individual protein from the same structure that was used to generate the density map of the entire nanomachine. Each map is simulated at the same resolution as the map of the entire nanomachine and on an identical grid. However, the region occupied by each protein is not well defined in these maps, because the convolution with a Gaussian filter produces non-zero density values everywhere in the map. On the other hand, the segmentation of the density map of the entire nanomachine includes points with densities only up to a certain threshold, as described previously. The same threshold is thus first applied to each of the protein maps to obtain the volumes of each individual protein.

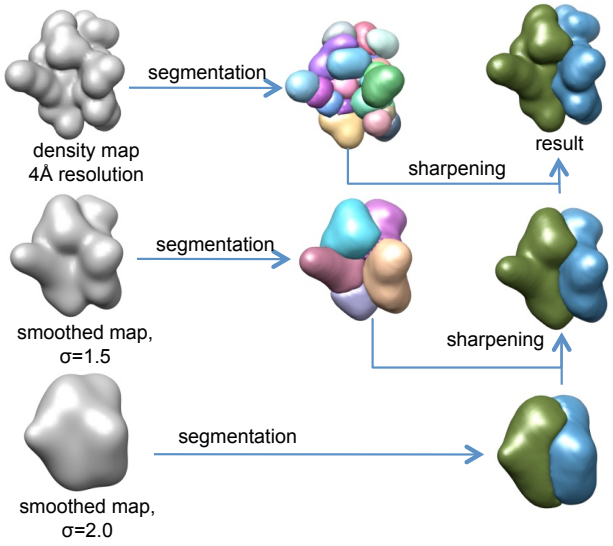

Fig. 1. Illustration of the S^3 method, including smoothing, segmentation, and sharpening operations. The input density map and the smoothed density maps are shown on the left by iso-surfaces. The segmentations are illustrated by smoothed surfaces that enclose each region.

F. Smoothing, segmentation, and sharpening (S^3)

The method presented here involves smoothing, segmentation, and sharpening, and hence we call it S^3. It is illustrated in Fig. 1.

1) Smoothing: The input density map is smoothed by convolution with Gaussian functions of several standard deviations. The coarsest scale, where the largest standard deviation is used, is determined by visual inspection of the segmentation results. At this scale, the segmentation should produce a small number of regions, but not smaller than the number of component proteins that are expected to be in the density map. Also, each region should ideally correspond to a single protein. When the results are such that the combination of two or more regions correspond to a single protein, these regions are merged by user interaction (UI). This step is the only non-automatic part of the process, and involves a small amount of knowledge about the proteins in the density map.

2) Segmentation: The maps at every scale are segmented using the watershed method. This process is completely automatic and only requires the specification of a density threshold, determined as specified above. It is also extremely fast, taking less than a second for density maps with $100 \times 100 \times 100$ voxels.

3) Sharpening: The segmentation at the coarsest scale is sharpened using the segmentation at the next finer scale using a simple overlap rule, as proposed in [19]. The regions at the finer scale are partitioned based on which region at the coarser scale they overlap the most. The regions in each partition are then joined, and the resulting regions replace the regions at the coarser scale.

G. The highest attainable accuracy for a given segmentation

Since in the simulated density maps we analyze here we know the ground truth, we can determine how much the sharpening process could actually improve the segmentation accuracy. Given the segmentation of a density map at any scale, we partition the regions again based on which protein volume they overlap the most. The regions in each partition are merged into a single region, which will correspond as closely to the corresponding protein as the given segmentation will allow. The segmentation accuracy between each of these regions and the corresponding protein is then computed using (1).

III. RESULTS

The S^3 segmentation method was applied to 4 simulated density maps. All of the maps were simulated to a resolution of 4Å. The simplest structure (shown in in Fig. 1) contains only two small components closely interacting with each other. Because this structure is very small, a very fine grid spacing of 0.2Å was possible, yielding a map of $\sim 100 \times 100 \times 100$ voxels. The other three density maps are of full-fledged nanomachines (Fig. 2). The Thermsome and GroEL+GroES are chaperones that help misfolded proteins attain their functional forms. They have barrel-like shapes in which the misfolded proteins bind, and they consist of 16 and 21 proteins respectively. The ribosome is also an extremely important nanomachine: it transcribes RNA into proteins. For these larger structures, the grid spacing was 2Å, also yielding maps of $\sim 100 \times 100 \times 100$ voxels.

For the density map shown in Fig. 1 and for the density map of the Thermosome, the segmentation at the coarsest scale produces a single region for each protein, and hence the segmentation was achieved with very little effort. For the GroEL+GroES and Ribosome nanomachines, the segmentation at the coarsest scale produced on average two regions per protein, and joining these two regions by UI to form a single region corresponding to each protein also did not require a great deal of effort.

The segmentation accuracies for all four density maps at each scale are plotted in Fig. 3. For the small structure, the accuracy of the segmentation is quite low at the coarsest scale, but rises to $\sim 97\%$ when sharpened to the finest scale. The highest attainable accuracy given the segmentation at each scale is the same as the accuracy obtained after sharpening, showing that in such a simple case the S^3 method can produce an almost perfect segmentation. For the larger nanomachines however, the segmentation accuracies were somewhat lower, especially when many protein components are present. The lower accuracies could be attributed in part to the larger grid spacing used for the density maps of these larger structures, which introduces some discretization error. Despite this limitation, the S^3 method is able to segment out each protein component successfully, with accuracies of $\sim 73\%$ and higher.

IV. CONCLUSIONS

We have presented a multi-scale segmentation method for the segmentation of protein components in density maps. The method requires very little effort on the part of the user. This is an improvement over previous methods that relied on seed

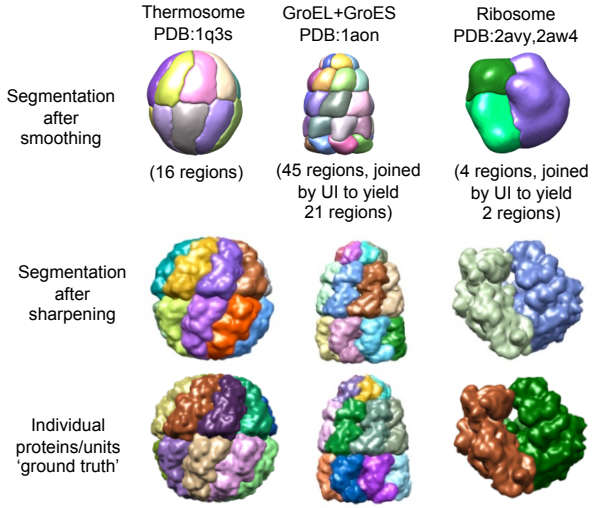

Fig. 2. Three simulated density maps of nanomachines that the S^3 method was applied to. The resulting segmentations visually match the ground truth very well.

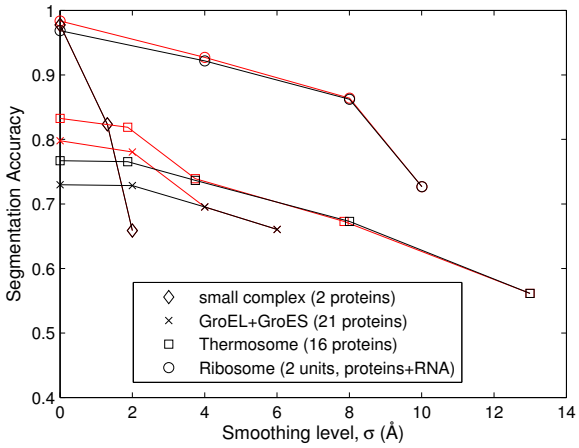

Fig. 3. Average segmentation accuracies for 4 density maps. The accuracy for each protein component is computed using (1), and an average is then calculated over all the proteins in each density map. This average is plotted at all the scales considered. The scale where $\sigma = 0$ refers to the input density map. The average highest attainable accuracy given the segmentation of a density map at each scale is plotted in red.

points being specified, a much harder process that involves more knowledge about the components to be segmented. Some of the segmentations achieved with the presented method were very accurate, however lower accuracies were obtained for density maps with more protein components. We plan to research ways to further improve the method so as to achieve higher accuracies. Moreover we plan to apply the method to a wider range of nanomachines and especially to experimental density maps, where its use will aid in further understanding how these entities perform their complex functions.

ACKNOWLEDGMENT

The authors thank Prof. Jonathan King in the Dept. of Biology at MIT for biological insights related to the problem. Molecular graphics images were produced using the Chimera package from the Computer Graphics Laboratory, University of California, San Francisco (supported by NIH P41 RR-01081).

REFERENCES

[1] S.J. Ludtke, P.R. Baldwin, and W. Chiu, EMAN: semiautomated software for high-resolution single-particle reconstructions, *Journal of structural biology*, vol. 128, Dec. 1999, pp. 82-97.

[2] M.L. Baker, Z. Yu, W. Chiu, and C. Bajaj, Automated segmentation of molecular subunits in electron cryomicroscopy density maps, *Journal of Structural Biology*, vol. 156, Dec. 2006, pp. 432-441.

[3] L.G. Shapiro and G.C. Stockman, *Computer Vision*, Prentice Hall, 2002.

[4] Beucher, S. and Lantuejoul, C, Use of watersheds in contour detection, Rennes, France: 1979.

[5] L. Najman and M. Schmitt, Geodesic Saliency of Watershed Contours and Hierarchical Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, 1996, pp. 1163-1173.

[6] B. Marcotegui, S. Beucher, and C. De Morphologie Mathmatique, Fast implementation of waterfall based on graphs, *Volume 30 of Computational Imaging and Vision*, vol. 30, 2005, pp. 177–186.

[7] N. Volkmann, A novel three-dimensional variant of the watershed transform for segmentation of electron density maps, *Journal of Structural Biology*, vol. 138, 2002, pp. 123-129.

[8] [8] A. Witkin, Scale-space filtering: A new approach to multi-scale description, *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, 1984, pp. 150-153.

[9] P. Perona and J. Malik, Scale-space and edge detection using anisotropic diffusion, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, 1990, pp. 629-639.

[10] K. Fukunaga and L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, *Information Theory, IEEE Transactions on*, vol. 21, 1975, pp. 32-40.

[11] Yizong Cheng, Mean shift, mode seeking, and clustering, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, 1995, pp. 790-799.

[12] D. Comaniciu and P. Meer, Mean shift: a robust approach toward feature space analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, 2002, pp. 603-619.

[13] W. Wriggers and S. Birmanns, Using situs for flexible and rigid-body fitting of multiresolution single-molecule data, *Journal of Structural Biology*, vol. 133, pp. 193-202.

[14] M. van Heel and M. Schatz, Fourier shell correlation threshold criteria, *Journal of Structural Biology*, vol. 151, Sep. 2005, pp. 250-62.

[15] L. Vincent and P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, 1991, pp. 583-598.

[16] S. Paris and F. Durand, A Topological Approach to Hierarchical Segmentation using Mean Shift, *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1-8.

[17] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin, UCSF Chimera–a visualization system for exploratory research and analysis, *Journal of computational chemistry*, vol. 25, Oct. 2004, pp. 1605-12.

[18] E. Garduno, M. Wong-Barnum, N. Volkmann, and M.H. Ellisman, Segmentation of electron tomographic data sets using fuzzy set theory principles, *Journal of Structural Biology*, vol. 162, Jun. 2008, pp. 368-379.

[19] D. DeCarlo and A. Santella, Stylization and abstraction of photographs, *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, San Antonio, Texas: ACM, 2002, pp. 769-776.