Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2007 April 12; 2007: 236–239. doi:10.1109/ISBI.2007.356832.

AUTOMATIC STRUCTURE INTERPRETATION OF SINGLE PARTICLE CRYO-ELECTRON MICROSCOPY: FROM IMAGES TO PSUEDO-ATOMIC MODELS

Chandrajit L. Bajaj

Department of Computer Sciences & Institute of Computational Engineering and Sciences,
University of Texas at Austin, Austin, Texas 78712

Abstract

Three dimensional Electron Microscopy (EM) and in particular single particle reconstruction using cryo-EM, has rapidly advanced over recent years, such that increasingly several macromolecular complexes can be resolved at subnanometer resolution (6–10 Å). This paper reviews some of the main volumetric image and geometric post-processing steps once a three dimensional EM map (henceforth a 3D map) has been reconstructed from single particle Cryo-EM, as essential steps in an enhanced and automated computational structure interpretation pipeline. In particular the paper addresses automated filtering, critical point calculations, symmetric and non-symmetric molecular domain segmentation, molecular surface selection, curation, and protein secondary structure (α -helices and β -sheets) elucidation from 3D maps.

1. INTRODUCTION

Structural genomics initiatives target solving structures of most existing protein folds by X-ray crystallography [1,2], and nuclear magnetic resonance (NMR) [3,4], such that most of the remaining proteins can be modeled with useful accuracy based on their similarity (homology) to known structures. While structures of individual proteins, or small complexes, such as most of the Protein Data Bank (PDB) [5] entries, provide important information, they do not necessarily yield the “full picture” of a functional biological complex. The study of large macromolecular complexes, such as viruses, ion channels, the ribosome and other molecular machines of various types, offer a more complete structural and functional description of the protein machinery. In recent years, cryo electron microscopy (cryo-EM) of single particles has become the procedure in vogue, with available software packages (e.g., EMAN [6], SPIDER [7], IMAGIC [8]) used to reveal the structure of large complexes at subnanometer resolutions (6–10Å) [9–12]. Nevertheless, the infrastructure for quantitative analysis and visualization of the reconstructed three dimensional EM map (henceforth 3D map) still remains relatively undeveloped, and structure interpretation has become the rate limiting step [13,14]. Determining the individual multiple structural component models (also called structure units) of the macromolecular assembly upto the resolution of the 3D map, yields crucial information of the structural interactions of the assembly, and thereby facilitates the interpretation of the function of the macromolecular machines.

Most current methods for interpreting 3D maps still depend primarily upon manual selection coupled to visual inspection with the help of interactive graphics tools. Due to the large physical size and structural complexity of the biomolecular assemblies, however, mechanical interactive

processing is tedious and subjective. Automatic structural interpretation and analysis of 3D maps of large biomolecular complexes have thus become the preferred avenue of research and development. Automatic segmentation decomposes the 3D maps of biomolecular complexes, into dozens to thousands of individual structure units. Furthermore, with automatic segmentation one is able to isolate the various grouped protein structure units (conformers), and further interpret the ultra-structure of each of the isolated proteins. For symmetric structures, such as most protein capsid shells of viruses [13], the segmentation of the entire capsid into protein structure units (capsomers) also helps eliminate structural representation redundancy.

This paper reviews certain fundamental volumetric image and geometry, computational processing steps once a three dimensional EM map (henceforth a 3D map) has been reconstructed from single particle Cryo-EM, as essential steps in an enhanced and automated computational structure interpretation pipeline[15–17]. In particular the paper addresses automated filtering, critical point calculations, symmetric and non-symmetric molecular domain segmentation, molecular surface selection and curation in section 3, and protein secondary structure (α -helices and β -sheets) elucidation from 3D maps, in section 4. A brief introduction to the single particle Cryo-EM is given in section 2

2. 3D CRYO-EM RECONSTRUCTION

Vitrified biological samples are highly sensitive to radiation damage and data must be collected at very low electron doses, resulting in images with a poor signal to noise ratio. Image reconstruction is further complicated by the fact that the images contain a mix of phase-contrast and amplitude-contrast components and must be corrected using the contrast transfer function. Images also suffer from effects of blurring owing to specimen charging and specimen movements during exposure to the electron beam and from asymmetries in the imaging optics (e.g. from astigmatism in the objective lens). Finally, in many cases, the orientations of the particles in the sample are not known precisely and the quality of the reconstruction is limited by the accuracy to which we can estimate these orientations. As a consequence of these and other factors, the best structures determined today from single particle cryoEM have resolutions typically in the range of 6 – 10Å. Figure 1 illustrates the key computational steps in the comprehensive pipeline from 2D images to 3D maps to psuedo-atomic models.

In view of the progress made in biochemical purification of large complexes and the improved resolution of cryo-EM, we expect that the number of macromolecular complex structures solved at subnanometer resolution will continue to increase [18]. While it is not possible to unambiguously determine a full atomic model at this resolution, it is possible to define molecular domains and secondary structure elements, such as α -helices and β -sheets. With the combined help of computational geometry and image (signal) processing techniques, we now describe how we compute the connectivity among these observed secondary structure elements and de-rive the folds of protein domains within the larger complex.

3. STRUCTURE UNIT IDENTIFICATION

3.1. Filtering

Filtering of 3D maps, in the context of structure unit identification, represents the preservation of the map's "critical" features while suppressing the most unimportant details. The critical features of maps could be points, curves, or surfaces in a given 3D map. For segmentation as well as for symmetry detection, we shall only consider critical points. In general, the critical points of a scalar map can be defined by three types: local maxima, local minima, and saddle points. However, the maximal critical points are of direct relevance in capturing the structures, since in a molecular density map, higher density implies existence of more atoms per unit

volume. The maximal critical points can be easily computed from the local maxima of a given 3D map. Since noise is always present in the original maps, a pre-filtering process should be applied. A filter can be either linear or nonlinear. A linear filter (e.g., Gaussian filtering) may suppress some weak features and hence eliminate some critical points. A nonlinear filter [19, 20], however, tends to “flatten” a region, yielding many unwanted critical points.

In an approach we employ[?], we utilize a bilateral pre-filter coupled with an evolution driven anisotropic geometric diffusion PDE (partial differential equation). This has shown significant results in enhancing the feature analysis and modeling of 3D macromolecular density maps.

The PDE model is: $\partial_t \phi - \|\nabla \phi\| \operatorname{div} \left(D^\sigma \frac{\nabla \phi}{\|\nabla \phi\|} \right) = 0$

The efficacy of our method is based on a careful selection of the anisotropic diffusion tensor D^σ based on estimates of the normal and two principal curvatures and curvature directions of a feature isosurface (level-set) in three dimensions. The two subfigures of Figure 3.1 in the left show the performance of this approach on a single slice of the map.

The linear and non-linear filtering techniques often fail because they do not take into the account the features present in the map. An alternative technique of filtering is based on the diffusion of gradient vectors, from which the critical points of the 3D function are extracted. In [21], the authors described a diffusion technique to smooth gradient vector fields. The gradient vectors are represented by Cartesian coordinates and a set of partial differential equations (PDEs) are separately applied to each component of the vectors. The equations are linear or isotropic, and therefore inherit the draw-backs of most linear systems. Another way to diffuse a gradient vector field is based on the polar-coordinate representation of the vectors [22,23]. The drawback of this method is that its somewhat compute intensive.

3.2. Symmetric and Non-Symmetric 3D Segmentation

Segmentation is a way to electronically dissect the 3D map into significant biological structure units, and thereby obtain a clearer view into the molecular machinery’s architectural organization[24]. For instance, it is often helpful to segment an asymmetric subunit out of an icosahedral virus such that the structural interpretation can be conducted only on the asymmetric subunit instead of the entire 3D map without loss of any structural information.

Segmentation, currently is typically carried out either manually [25–29] or semi-automatically [30,31]. Current efforts on the decomposition of an icosahedral map into its structure units still largely rely on manual work with an assistance of a graphical user interface [9,32]. Manual segmentation can be tedious and often subjective even with the help of a sophisticated graphical user interface [27,?].

Automated segmentation is to date, still recognized as one of the most difficult tasks in the field of image processing although various techniques have been proposed for automated or semi-automated segmentation. Commonly used methods include segmentation based on edge detection, region growing and/or region merging, active curve/surface motion and model based segmentation. In particular, two techniques have been discussed in detail within the 3D EM community. One is called *watershed immersion method* [30] and the other is based on normalized graph cut and eigenvector analysis [31]. Recently, Frangakis and Hegerl showed that high-quality segmentation can be achieved by using eigenvector methods that optimize the normalized cut criterion for graph clustering [31]. However, as they themselves recognized in [31], the time needed to perform segmentation of 3-dimensional images is a severe bottleneck.

The 3D map segmentation scheme we employ for symmetric and non-symmetric domains [16,17], is adopted from a variant of the fast marching method [33–35] which is a simplified

version of the level set method [35]. In this method a contour is initialized from a pre-chosen seed point, and the contour is allowed to grow until a certain stopping condition is reached. Traditionally this method is designed for a single domain segmentation. In order to segment multiple structure units, such as the 60T-components of icosahedral virus capsid shells or the 3, 5 or 6- component protein conformers (i.e. trimers, pentamers, hexamers), one has to choose a seed for each of the individual structure units. Here T is known as the triangulation number of the icosahedral representation[]. This requires a pre-processing identification step of all local and global symmetries in the 3D map, discernible at that resolution, to separate out the individual structure units. Details of this symmetry elucidation step, using critical points of the 3D function, and fast cross-correlation analysis are given in [16].

However, assigning only one seed to each structure unit may lead to an incorrect 3D segmentation. To remedy this problem, we have developed an approach based on an idea of “re-initialization” by simply regarding all the critical points as seeds[16]. Every seed initiates one contour and all contours start to grow simultaneously and independently. We further classify the critical points into clusters and merge the growing contours which are initiated by the critical points in the same cluster. Accuracy of the segmentation process is shown in Figure 3.1.

The success of the fast marching based segmentation method relies on the choice of the seed points and the speed function. The number and distribution of the seed points should reflect the dominance of a subregion to be segmented. The same is true for the speed function as well. This leads to the investigation on how to choose these two parameters properly. One way is to estimate the parameters by considering the minimum bounding ellipsoid of the domain of segmentation. This estimation can be done via principal component analysis. The same segmentation method is applicable also to the segmentation of molecular domains which do not exhibit symmetry. This of course still requires an appropriate tuning of the seed points and speed function parameters. The results are promising as shown in Figure 3.

3.3. Stable Computation of Molecular Surfaces

Much of structural interrogation happens via closer examination of the molecular surface of the segmented protein molecule from the virus map. A correct approximation of the molecular surface is therefore very important in order to carry out the subsequent operations. The molecular surface extracted from the segmented map is often flawed due to the presence of extra tunnels, small components and voids. Purification of the molecular surface thus requires a post-processing step that detects these unnecessary small topological artifacts and removes them.

We have developed an algorithm[36] for such topological curation, to yield a faithful description of the molecular surface from a segmented volumetric 3D map. The approach relies on a topology-guided isosurface extraction process. Contour Tree [37] provides a description of the topology of the iso-surface as the isovalue varies. While this description is helpful, it only indicates the number of components of the isosurface. Further enhancement was proposed in [38] to capture the topology of the level set via detection of the Betti numbers of the lower level set for a certain isovalue. Even this enhanced description, which was named Augmented Contour Tree, fails to serve the purpose of extracting a correct molecular surface as it does not provide a handle to locally change the surface to remove the topological artifacts as reflected by the higher values of the first three Betti numbers.

Our approach [36] is based on the critical point analysis of the distance function induced by the isosurface. Guided by Contour Tree, an initial isosurface is selected and its critical point structure is analysed. The topological features of the isosurface are the number of components (0-th Betti number), tunnels (1st Betti number) and voids (2nd Betti number). While the number

of connected components and voids are easy to quantify via their geometric attributes, detection and removal of tunnels is difficult. Every tunnel is accompanied by at least one index 2 saddle point of the distance function in the complementary space of the molecule. We therefore collect all the index 2 saddle points and cluster them depending on if they are attached to the same tunnel. The detection of the critical points is done efficiently exploiting the inherent connection of distance function with the Voronoi-Delaunay diagram [39,40]. The augmented stable manifold of each cluster of the critical points then produces an approximation of the tunnels. The detected tunnels can then be ranked via their geometric attributes and the smaller (or insignificant) ones are filled to purify the molecular surface. Figure 4 shows an example of topological purification of the molecular surface extracted from a 3D density map computed from the protein structure pdbid 1A00, taken from the Protein Data Bank.

4. SECONDARY STRUCTURE, TERTIARY FOLD ELUCIDATION

Proteins are polypeptide chains of amino acids. The secondary structures of proteins are made up of a set of helical (or crudely, cylindrical) arrangements of sub-chains called α -helices and mostly planar arrangements of sub-chains called β -sheets. The various structural conformations of sub-groups of helices and sheets, yield the various different tertiary folds [41]. Relatively similar configuration of secondary and tertiary folds arise also in RNA [42].

There are relatively few published algorithms that detect α -helices and β -sheets of proteins from 3D maps at intermediate resolution (6 – 10Å) - *Helix Hunter* by [43], β -sheet detection by [44,45]. These algorithms are very inefficient and suffer primarily due to their exhaustive search paradigm. In both the cases, helices are modeled as cylinders and sheets are modeled as disk like primitives and the algorithm searches over the whole space of the map for best correlation. Recently, we have developed two different algorithmic approaches for helix-sheet identification. In [46], the α -helices and β -sheets are detected by eigen-analysis of the structure tensor of the 3D segmented map of the protein, in selected neighborhoods. The eigen-analysis reveals the line-like or disk-like substructure within the map. It is a boundary-free approach. The second method [47], is a boundary-based approach, and works on the distance function induced by the molecular surface and identifies the helices and sheets by analyzing the index 1 and index 2 critical point (saddle points) structure of this function. Unstable manifolds (which we explain below) of the index 1 critical points reveals the β -sheet and that of index 2 critical points reveals the α -helices. Figure 5 shows the accuracy of the algorithm in detecting the secondary structure of the protein molecules.

4.1. Boundary-free Approach

The skeletonization method, which is key in detecting the secondary structure, is based on eigen-analysis of the local structure tensor and does not require a prior boundary segmentation of the protein 3D map. The local structure tensor has been used in image processing for solving a number of problems such as anisotropic filtering [20,48] and motion detection [49]. Given a 3D map $f(x, y, z)$, the Local Structure Tensor matrix is defined by taking a spatial average of the gradient vectors in a neighborhood over the image domain. The averaging is usually done by a Gaussian filter:

$$T_{\alpha} = \begin{pmatrix} f_x^2 * g_{\alpha} & f_x f_y * g_{\alpha} & f_x f_z * g_{\alpha} \\ f_x f_y * g_{\alpha} & f_y^2 * g_{\alpha} & f_y f_z * g_{\alpha} \\ f_x f_z * g_{\alpha} & f_y f_z * g_{\alpha} & f_z^2 * g_{\alpha} \end{pmatrix} \quad (1)$$

Here g_{α} is a Gaussian function with standard deviation α . The eigenvalues and eigenvectors of the structure tensor T_{α} indicate the overall distribution of the gradient vectors within the local window, similar to the well-known principal component analysis (PCA). Three typical

structures can be characterized based on the eigenvalues [48]. Let the eigenvalues be $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Then we have the following cases

- Spheres: $\lambda_1 \approx \lambda_2 \approx \lambda_3 > 0$.
- Lines: $\lambda_1 \approx \lambda_2 \gg \lambda_3 \approx 0$.
- Planes: $\lambda_1 \gg \lambda_2 \approx \lambda_3 \approx 0$.

Using the eigen-analysis of the local structure tensor, the possibility of a voxel falling into the linear or planar portion of the skeleton is computed. Further, the thickness along the direction of the eigen-vectors filters the portions which actually correspond to α -helices and β -sheets. Figure 6 shows the performance of this algorithm in terms of accuracy on a blurred atomic X-ray structure.

4.2. Boundary-based Approach

The boundary-based approach works on analyzing the distance function induced by the molecular surface in three dimensions. The critical points of the distance function are of four types - maxima, minima, index 1 saddle and index 2 saddle. All these four types of critical points can be identified from the Voronoi-Delaunay diagram of the point sample of the molecular surface [51]. Every critical point has an associated structure, called stable/unstable-manifold. These stable/unstable manifolds are the collection of points in three dimension that have similar gradient vector field of the distance function and converge to the critical point if they are allowed to move in the steepest ascent/descent of the distance function [52]. The unstable manifold of an index 1 saddle point is two dimensional, whereas the unstable manifold of an index 2 saddle point is one dimensional. This leads to the annotated skeletonization algorithm, which computes the unstable manifold of the index 1 and 2 saddle points of the distance function close to the medial axis and thus dissects the medial axis into linear or planar sub-features. The linear portion of the skeleton, with proper thickness, corresponds to the α -helices and the planar portions correspond to the β -sheets.

5. CONCLUSIONS

In view of the progress made in biochemical purification of large complexes and the improved resolution of cryo-EM, it is all but certain that the number of macromolecular complex structures solved at subnanometer resolution will continue to increase. As demonstrated in the previous sections, with the development and application of volumetric image processing and computational geometry algorithms, it is possible to automatically elucidate the ultrastructure of protein constituents of macromolecules, namely, the protein's secondary structure elements, such as helices and sheets, from 3D reconstructed maps. With this information in hand, it is also conceivable that one may now link the structural informatics with other biophysical and biochemical informatics, to better interpret the functional mechanism of biological machines and their components.

Acknowledgments

This research was supported in part by NSF grants IIS-0325550, CNS-0540033 and grants from the NIH 0P20 RR020647, R01 GM074258, R01-GM073087. The implementation of all of the above structural interpretation algorithms, as well as the pictures used here, have been made using our in-house image processing and visualization software tool, called VolRover. The VolRover program is in the public domain and can be freely downloaded from our center's software website: (<http://www.ices.utexas.edu/CVC/software/>).

6. REFERENCES

1. Woolfson, MM. An Introduction to X-ray Crystallography. Cambridge University Press; Jan. 1997
2. "Online course on protein crystallography," <http://www.structmed.cimr.cam.ac.uk/course.html>.

3. Ferentz AE, Wagner G. NMR spectroscopy : a multifaceted approach to macromolecular structure. *Quarterly Reviews of Biophysics* 2000;33:29–65. [PubMed: 11075388]
4. Guntert P. Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics* 1998;31:145–237. [PubMed: 9794034]
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Research* 2000;28:235–242. [PubMed: 10592235]
6. Ludtke SJ, Baldwin PR, Chiu W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *Journal of Structural Biology* 1999;128:82–97. [PubMed: 10600563]
7. Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, Leith A. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *Journal of Structural Biology* 1996;116:190–199. [PubMed: 8742743]
8. Van Heel M, Harauz G, Orlova EV, Schmidt R, Schatz M. A new generation of the IMAGIC image processing system. *Journal of Structural Biology* 1996;116:17–24. [PubMed: 8742718]
9. Zhou ZH, Baker ML, Jiang W, Dougherty M, Jakana J, Dong G, Lu G, Chiu W. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Structural Biology* 2001;8(10):868–873.
10. Jiang W, Li Z, Baker ML, Prevelige PE, Chiu W. Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolution. *Nature Structural Biology* 2003;10(2):131–135.
11. Zhou ZH, Dougherty M, Jakana J, He J, Rixon F, Chiu W. Seeing the herpesvirus capsid at 8.5 angstrom. *Science* 2000;288:877–880. [PubMed: 10797014]
12. Matadeen R, Patwardhan A, Gowen B, Orlova EV, Pape T, Cuff M, Mueller F, Brimacombe R, van Heel M. The e. coli large ribosomal subunit at 7.5 angstrom resolution. *Structure* 1999;7:1575–1583. [PubMed: 10647188]
13. Baker TS, Olson NH, Fuller SD. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and Molecular Biology Reviews* 1999;63(4):862–922. [PubMed: 10585969]
14. van Heel M, Gowen B, Matadeen R, Orlova EV, Finn R, Pape T, Cohen D, Stark H, Schmidt R, Schatz M, Patwardhan A. Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews Biophysics* 2000;33(4):307–369.
15. Murphy, RF.; Meijering, E.; Danuser, G., editors. *IEEE Transactions on Image Processing: Special Issue on Molecular and Cellular Bioimaging*. Vol. 14. 2005.
16. Yu Z, Bajaj C. Automatic ultra-structure segmentation of reconstructed cryo-em maps of icosahedral viruses. *IEEE Transactions on Image Processing* 2005;14(9):1324–1337. [PubMed: 16190468]
17. Baker M, Yu Z, Chiu W, Bajaj C. Automated segmentation of molecular subunits in electron cryomicroscopy density maps. *Journal of Structural Biology* 2006;156(3):432–441. [PubMed: 16908194]
18. Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu Rev Biophys Biomol Struct* 2002;31:303–319. [PubMed: 11988472]
19. Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Machine Intell* 1990;12:629–639.
20. Weickert, J. *ECMI Series*. Teubner: Stuttgart; 1998. *Anisotropic Diffusion In Image Processing*.
21. Xu C, Prince JL. Snakes, shapes, and gradient vector flow. *IEEE Trans Image Processing* 1998;7(3):359–369.
22. Yu, Z.; Bajaj, C. Anisotropic vector diffusion in image smoothing. *Proceeding of the 9th IEEE ICIP*; 2002. p. 828-831.
23. Yu, Z.; Bajaj, C. Image segmentation using gradient vector diffusion and region merging. *International Conference on Pattern Recognition*; 2002. p. 941-944.
24. Ellis RJ. Macromolecular crowding: obvious but un-derappreciated. *Trends Biochem Sci* 2001;26(10):597–604. [PubMed: 11590012]
25. Hessler D, Young SJ, Ellisman MH. A flexible environment for the visualization of three-dimensional biological structures. *Journal of Structural Biology* 1996;116(1):113–119. [PubMed: 8742732]
26. Kremer JR, Mastronarde DN, McIntosh JR. Computer visualization of three-dimensional image data using IMOD. *J Struct Biol* 1996;116:71–76. [PubMed: 8742726]

27. Li Y, Leith A, Frank J. Tinkerbelle—a tool for interactive segmentation of 3d data. *Journal of Structural Biology* 1997;120(3):266–275.
28. McEwen BF, Marko M. Three-dimensional electron microscopy and its application to mitosis research. *Methods Cell Biol* 1999;61:81–111. [PubMed: 9891310]
29. Harlow ML, Ress D, Stoschek A, Marshall RM, McMahan UJ. The architecture of active zone material at the frog's neuromuscular junction. *Nature* 2001;409:479 – 484. [PubMed: 11206537]
30. Volkman N. A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *Journal of Structural Biology* 2002;138(1–2):123–129. [PubMed: 12160708]
31. Frangakis AS, Hegerl R. Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis. *Journal of Structural Biology* 2002;138(1–2):105–113. [PubMed: 12160706]
32. Jiang W, Li Z, Baker ML, Prevelige PE Jr, Chiu W. Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nat Struct Biol*, in press. 2003
33. Malladi, R.; Sethian, JA. A real-time algorithm for medical shape recovery. *Proc. of Int'l Conf. on Computer Vision*; 1998. p. 304–310.
34. Sethian JA. A marching level set method for monotonically advancing fronts. *Proc Natl Acad Sci* 1996;93(4):1591–1595. [PubMed: 11607632]
35. Sethian, JA. *Level Set Methods and Fast Marching Methods*. 2. Cambridge University Press; 1999.
36. Bajaj, C.; Gillette, A.; Goswami, S. Workshop on Topology Based Methods in Visualization (TopoInViz 07). 2006. Topology based selection and curation of level sets.
37. Hamish, Carr; Jack, Snoeyink; Ulrike, Axen. Symposium on Discrete Algorithms. 2000. Computing contour trees in all dimensions; p. 918–926.
38. Pascucci V, Cole-McLaughlin K. Parallel computation of the topology of level sets. *Algorithmica* 2003;38(2):249–268.
39. Giesen, J.; John, M. The flow complex: a data structure for geometric modeling. *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms*; 2003. p. 285–294.
40. Dey, TK.; Giesen, J.; Goswami, S. Shape segmentation and matching with flow discretization. In: Dehne, F.; Sack, J-R.; Smid, M., editors. *Proc. Workshop Algorithms Data Structures (WADS 03)*; Berlin, Germany. 2003. p. 25–36. LNCS 2748
41. Branden, C.; Tooze, J. *An Introduction to Protein Structure*. Garland Publishing Inc; New York: 1999.
42. Schneider B, Morveki Z, Berman H. RNA conformational classes. *Nucleic Acids Research* 2004;32(5):1666–1677. [PubMed: 15016910]
43. Wen J, Baker M, Ludtke S, Chiu W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J of Mol Bio* 2001;308:1033–1044. [PubMed: 11352589]
44. Kong Y, Ma J. A structural-informatics approach for mining β - sheets: locating sheets in intermediate-resolution density maps. *J of Mol Bio* 2003;332:399–413. [PubMed: 12948490]
45. Kong Y, Zhang X, Baker TS, Ma J. A structural-informatics approach for tracing β -sheets: building pseudo- α traces for β -strands in intermediate-resolution density maps. *Journal of Molecular Biology* 2004;339:117–130. [PubMed: 15123425]
46. Yu, Z.; Bajaj, C. Tech Rep TR-06-05. Department of Computer Sciences; UT Austin: 2006. Computational approaches for automatic structural analysis of large bio-molecular complexes.
47. Bajaj C, Goswami S. Automatic fold and structural motif elucidation from 3D EM maps of macromolecules. *ICVGIP 2006, LNCS 4338* 2006:264–275.
48. Fernandez JJ, Li S. An improved algorithm for anisotropic nonlinear diffusion for denoising cryotomograms. *Journal of Structural Biology* 2003;144:152–161. [PubMed: 14643218]
49. Khne, G.; Weickert, J.; Schuster, O.; Richter, S. A tensor-driven active contour model for moving object segmentation. *Proceedings of IEEE International Conference on Image Processing*; 2001. p. 73–76.
50. Jiang W, Baker ML, Ludtke SJ, Chiu W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *Journal of Molecular Biology* 2001;308:1033–1044. [PubMed: 11352589]
51. Siersma D. Voronoi diagrams and morse theory of the distance function. 1999

52. Matsumoto, Y. An Introduction to Morse Theory. Hudson, K.; Saito, M., translators. American Mathematical Society; 2002.

Fig. 1.

A comprehensive computational pipeline of structure interpretation and psuedo-atomic model generation starting from 2D cryo-EM images of virus particles. The pipeline consists of several computational steps, only some of them are described in this review.

Fig. 2.

Filtering, symmetry detection and segmentation applied to a cryo-EM 3D map of the Rice Dwarf Virus (RDV). Starting from the input 3D map, the signal to noise ratio is enhanced and the inner capsid, outer capsid and the genomic core are separated and the symmetry axes of the icosahedral outer capsid are automatically detected. The bottom row shows further segmentation of the outer capsid into asymmetric structure units, trimeric structure units and eventually into the monomeric constituent proteins (P8).

Fig. 3. Segmentation of a cryoEM map of no symmetry at 10Å resolution. (left) The reconstructed cryoEM maps of 70S ribosome aminoacyl-tRNA from E. coli complex. (middle) Segmentation of 70S ribosome map into 30S and 50S subunits. The 30S subunit is represented by yellow density while the 50S subunit is shown in blue. The 30S subunit is further segmented in protein (purple) and RNA (yellow) in (right)

Fig. 4.

Topological curation of molecular surface from a simulated density map of pdbid 1A00. The red surface triangles show the mouth of the detected tunnels. This molecule has a large tunnel as shown in the leftmost picture and four smaller tunnels as shown in the second picture from the left. These four tunnels, although smaller than the biggest tunnel, are biologically crucial as they provide the pathway for other binding molecules. The two rightmost figures show the two small tunnels which are merely caused by the artifacts of isosurface generation, and hence are removed.

Fig. 5.

Results of secondary structure elucidation from 3D maps. (a) Molecular surface display of the Bluetongue Virus capsid protein extracted from an $8\text{\AA}$ resolution map. (b) The detected helices are displayed as green cylinders and the detected sheets are displayed as pink patches. (c) The detected helices and sheets are superimposed with a ribbon diagram visualization of the crystal structure model of the Bluetongue Virus capsid protein (1BVP) from the Protein Data Bank. (d) same as (c) but viewed along a local three-fold axis from the bottom of the trimer. (f–h) shows another example where the capsid protein is automatically segmented from a reconstructed 3D map at $7\text{\AA}$, of the Rice Dwarf Virus (RDV). The monomeric constituent protein P8 of the outer capsid layer has in total 16 helices and 3 sheets. The current algorithm detects 12 helices and 1 sheet.

Fig. 6.

Secondary structure identification on the triose phosphate isomerase from chicken muscle (PDBID = 1tim). (a) The blurred maps at $8\text{\AA}$ from the X-ray crystal structure. (b) The α -helices (green) and β -sheets (pink) detected using our method. (c) The detected helices/sheets are compared with the PDB structures. We can see that most helices/sheets agree very well with the PDB structure except that one small α -helix that is immediately adjacent to a long helix is missed (indicated by red arrow). This result also agrees with that seen in [50]. (d) Side view of (b). (e) Side view of (c).

Fig. 7.

Secondary structure elucidation algorithm for Insulin Receptor Tyrosine Kinase Domain with pdbid: 1IRK. (a) volume rendering of its blurred 3D Map at $8\text{\AA}$ resolution (b) surface rendering of the protein's molecular surface (c) unstable manifolds of index 1 (yellow) and index 2 (red) saddle points and a closeup of the boxed portion (d) detected sheet (red) and the axes of the cylinders that correspond to the helices (e) combined display of the computed helices (green cylinders) and sheets (red patches) and the ribbon diagram of the secondary structure, as documented in Protein Data Bank.