Revised Manuscript Submitted to Journal of Structural Biology

Automated Segmentation of Molecular Subunits in Electron Cryomicroscopy Density Maps

Matthew L. Baker[1*], Zeyun Yu[2], Wah Chiu [1], Chandrajit Bajaj[2]

[1]National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, [2]The Center of Computational Visualization, Department of Computer Sciences and Institute of Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712

*Corresponding author: Matthew L. Baker

Phone: 713-798-6989

Fax: 713-798-1625

Email: mbaker@bcm.edu

Abstract

Electron cryomicroscopy is capable of imaging large macromolecular machines composed of multiple components. However, it is currently only possible to achieve moderate resolution at which it may be possible to computationally extract the individual components in the machine. In this work, we present application details of an automated method for detecting and segmenting the components of a large machine in an experimentally determined density map. This method is applicable to object with and without symmetry and takes advantage of global and local symmetry if present. We have applied this segmentation algorithm to several electron cryomicroscopy data sets already deposited in EMDB with various complexities, symmetries and resolutions and validated the results using manually segmented density and available structures of the components in the PDB. As such, automated segmentation could become a useful tool for the analysis of the ever increasing number of structures of macromolecular machines derived from electron cryomicroscopy.

Keywords: Segmentation, electron cryomicroscopy, macromolecular complexes, structure

Abbreviations: cryoEM – electron cryomicroscopy, 3D – three-dimensional, RDV – Rice Dwarf Virus

Introduction

Structural biology of macromolecular machines has become a crucial tool for understanding its mechanism of action, and in many instances leads to further function exploration. Large machines usually undergo motion and/or conformational changes in order to carry out a specific biological process (Alberts, 1998; Alberts et al., 1992). These large machines are made up of multiple components, ranging from one molecule repeated several times (e.g. GroEL) to tens of non-equivalent molecules (e.g. ribosome). The study of these machines has become more tractable due to the rapid development of proteomics for identifying and purifying ensembles of macromolecules (Gavin et al., 2002; Sali, 2003).

Despite limitations in resolution, electron cryomicroscopy (also known as cryoEM) can image macromolecular machines. One of the greatest advantages of cryoEM is the ability to analyze whole complexes in near native conditions as well as functionally important states. Even in the absence of high resolution information, analysis of cryoEM structures may reveal significant insight into macromolecular composition, secondary structure, and in some instances, produce pseudo-atomic models of the components (Bottcher et al., 1997; Conway et al., 1997; Zhou et al., 2001).

In analyzing density maps of multi-component machines derived from cryoEM, one of the first steps is to define the boundaries of the individual molecular components (Chiu et al., 2005). These individual components may in fact form higher-ordered structures, such as an asymmetric unit in a virus. Defining the boundary and segmenting the individual components at different resolutions represents a particularly interesting challenge. While at low resolution, individual subunit boundaries may not be resolvable,

making the segmentation nearly impossible. Conversely, at moderate resolutions, the molecular boundary of the components is more likely to be visible. The interactions between the neighbor molecules and the residual noise may, however, result in poor resolvability of the components. Even in the event a high resolution structure is known for one or more components, it is possible that the same components may differ structurally in different functional states.

To date, segmentation has been accomplished mainly by interactively and visually tracing the regions of interest in density maps. This is typically done by examining and defining the regions of interest in a 2D slice of the 3D density map; numerous visualization packages have incorporated these tools such as Amira (TGS, San Diego, CA), IMOD (Boulder Laboratory for 3-D Electron Microscopy of Cells, Boulder, CO) and the SAIL package developed for Iris Explorer (NAG, Dowber Grove, IL.). Such manual segmentation is tedious and often ambiguous even with the most successful visualization interfaces. Despite the prevalence of these tools, few automated methods for segmentation are available (3D watershed method (Volkmann, 2002), normalized graph cut and eigenvector analysis (Frangakis, 2004; Frangakis, 2002), fast marching method (Bajaj et al., 2003; Frangakis et al., 2004; Frangakis et al., 2002; Volkmann, 2002)) and as such, segmentation of volumetric density is still considered one of the most difficult tasks in structure interpretation in complex machines.

In this work, we describe the application of an automated segmentation algorithm (Yu 2005) to three-dimensional density maps derived from single particle cryoEM. This algorithm includes three major computational steps: the detection of critical points of the volumetric density, the detection of global and local symmetry axes if present, and

boundary segmentation of all subunits or domains. The resulting segmentation of the subunits is capable of producing molecular components with high fidelity, requiring only minimal manual refinement.

Methods

CryoEM density maps and X-ray coordinates

To validate the segmentation procedures, we obtained structural data from public structural databases. Individual cryoEM density maps were obtained from EBI's electron microscopy database, EMDB. These density maps include the 9.5 Å resolution P22 mature phage (Jiang et al., 2003) (EMDB ID: 1101), the 23Å resolution P22 tail machine (Tang et al., 2005) (EMDB ID: 1119), 6.8 Å resolution Rice Dwarf Virus (Zhou et al., 2001) (EMDB ID: 1060), 9 Å resolution *E. Coli* 70S ribosome (Valle et al., 2003) (EMDB ID: 1056) and the 6Å, 11.5Å and 25Å resolution GroEL (Ludtke et al., 2004; Ludtke et al., 2001; Sewell et al., 2004) (EMDB IDs: 1081, 1080, 1095 respectively). Corresponding X-ray structures were obtained from the Protein Data Bank (P22 tail spike PDB ID: 1TYU, Rice Dwarf Virus PDB ID: 1UF2, 50S ribosome PDB ID: 1FFK, 30S ribosome PDB ID: 1IBM, GroEL ID: 1GRL).

Segmentation

As described, manual segmentation of three-dimensional density maps can be a tedious and subjective process, especially when the resolution is only marginally high enough to discern the boundaries between subunits. As such, we have developed an automatic and objective computational procedure for asymmetric subunit detection of complexes. This approach is a variant of the well-known fast marching method (Malladi

et al., 1998; Sethian, 1996; Sethian, 1999), in which a contour is initialized from a pre-chosen seed point and allowed to grow until a certain stopping condition is reached. The traditional fast marching method is designed for single object segmentation. In order to segment multiple objects, like the molecular components found in macromolecular machine in a cryoEM density map, a seed for each of the components must be chosen. However, assigning only one seed to each object (i.e. component) may be problematic, as demonstrated previously, where a *re-initialization* scheme was proposed (Yu et al., 2005). This approach consists of three steps: (a) detection of the critical (seed) points; (b) classification of critical points; (c) multi-seeded fast marching method.

The critical points of a scalar map (i.e. cryoEM density map) in general include three types: maximal, minimal, and saddle. Only the maximal critical points are of interest in this approach which represent the high density features in the 3D density map and can be simply computed from the local maxima of a given scalar map. These critical points are regarded as the seed points in the fast marching method (Malladi et al., 1998; Sethian, 1996; Sethian, 1999). In principal, the seed points for the fast marching procedure could be generated by other methods (k-means, random, etc…), however only maximal critical points were considered in this work. Generally, the number of seed points in a map will be much larger than the number of subunits of interest, ranging from hundred to tens-of-thousands depending on the density map itself. As such, each component will be assigned multiple seed points instead of just one.

Complicating the accurate assignment of critical points is noise, which is present in the cryoEM density maps. In visualizing and analyzing cryoEM density maps, a pre-filtering process is generally applied to eliminate noise (Chiu et al., 2005). Linear (e.g.,

Gaussian filtering) or nonlinear (Perona et al., 1990; Weickert, 1998) filters often destroy some weak features and eliminate some critical points. In this work, we have used a method based on *gradient vector diffusion* (Bajaj et al., 2005; Yu et al., 2005), a partial differential equation (PDE)-based technique, that is capable of reducing noise while preserving real density features.

In general, macromolecular machines can have symmetries of different types, including helical (e.g., tobacco mosaic virus), icosahedral (e.g., rice dwarf virus), and n-fold symmetry (e.g., GroEL). In defining the global and local symmetry axes, it is possible to classify the critical points as potential members of each of the components. The most direct method for detecting the symmetry axes simply involves correlating the original map with its rotated map (according to the type of symmetry) and searching the resulting correlation map for the peaks (Masuda et al., 1993). Such an algorithm has been previously implemented and used in studying cryoEM maps of icosahedral viruses (He et al., 2001). However, this method has a very high computational cost, as the time complexity is in the order of $O(NM)$, where N is the number of voxels and M is the number of possible orientations to be searched.

As the aforementioned symmetry detection method scales poorly, we have used an alternative method for the detection of rotational symmetries, given that the n-fold number is known (Bajaj et al., 2005; Yu et al., 2005). N, the number of voxels to be tested, is reduced by restricting the map search only to a subset of the critical points instead of the entire volume. This strategy tremendously reduces the computational cost in detecting the symmetry of a given volumetric map (both globally and locally). Once the symmetry is detected, the critical points can be automatically classified based on their

symmetrically equivalent positions. As the critical points are assigned, equivalent points are assigned membership in the symmetrically related components. However, critical points in different subunits are classified with different memberships; the same membership is assigned to critical points only within one subunit.

Once all of the critical points (seeds) are indexed and assigned memberships, the traditional fast marching method (Malladi et al., 1998; Sethian, 1996; Sethian, 1999) can be used with the following modifications. First, each object may consist of a number of seeds instead of just one. Secondly, since each seed initiates a marching contour and all contours start to grow simultaneously and independently, each seed (and accordingly, the marching contour) must be attached with a membership index based on the classification of seeds. Once a voxel is conquered by a marching contour, it should be assigned with the same index of the marching contour. Thirdly, two marching contours with the same index should merge into one when they meet, while two marching contours with different indexes should stop on their common boundaries. This idea is known as multi-seeded fast marching method and has been used elsewhere (Bajaj et al., 2003; Sifakis et al., 2001).

All automated segmentation was performed using either the standalone segmentation programs or a graphical version incorporated into VolRover, a freely available volume rendering and feature analysis program developed at the Computational Visualization Center (CVC) at the University of Texas at Austin (download at http://ccvweb.csres.utexas.edu/software/applications.php). For the described automated segmentation of the aforementioned density maps, only the global symmetry, density range (density values corresponding to structure) and map were input by the user.

Validation of Automated Segmentation

Individual subunits identified from automated segmentation were fit with their corresponding high-resolution crystal structures and/or manually segmented subunits using FOLDHUNTER (Jiang et al., 2001). FOLDHUNTER simulates the coordinate data at the same resolution as density map in question, which is then followed by an exhaustive cross-correlation based search. As such, assessments of the fits were obtained by examining the cross correlation value for each fit. The correlation values range from 0 to 1, where 1 indicates the two density maps are identical and a value of 0 reflects no similarity of structures.

Results

To assess the fidelity of the automated segmentation routine, the aforementioned cryoEM datasets were subjected to segmentation and analysis. Each of these data sets has a corresponding high resolution crystal structure and/or previously manually segmented monomeric subunit from the cryoEM density map making it possible to provide both qualitative and quantitative measure of accuracy. Additionally, these data sets represent a wide range of complexities, symmetries and resolutions typically obtained by current cryoEM experiments.

Segmentation of the Mature Bacteriophage P22

Bacteriophage P22 (Fig. 1A) is a prototypical polyhedral virus, composed of a single capsid protein, gp5 arranged on a T=7 icosahedral lattice surrounding a dsDNA genome. Each asymmetric unit in P22 contains seven unique copies of gp5, six of which are about a local six-fold axis and one about the icosahedral five-fold axis.

The first round of segmentation successfully extracted the protein capsid layer from the dsDNA genome (Fig. 1B). Based on the identified 5-3-2 icosahedral symmetry and the local 6-fold symmetry (Fig 1C), the second round of segmentation extracted the asymmetric unit, containing one gp5 molecule in the penton and six gp5 molecules in the hexon within an asymmetric unit (Fig. 1D-F). The final round of segmentation segmented out each individual gp5 monomer from the asymmetric unit (Fig. 1G,H).

No atomic resolution structure of gp5 is known, however, manual segmentations of the gp5 monomers in the asymmetric unit are available (Jiang et al., 2003). When compared to the manually segmented gp5 monomer (Fig. 1I), the automatically segmented gp5 monomers from the local 6-fold axis have a cross correlation value of 0.79 with the corresponding manually segmented monomer, while the five-fold gp5 monomer has a cross-correlation value of 0.72 to the corresponding manually segmented gp5 subunit. Some minor differences can be seen in small features in the capsid floor.

Segmentation of Rice Dwarf Virus

Like bacteriophage P22, Rice Dwarf Virus (RDV) is an icosahedral virus (Fig. 2A). However, RDV contains two capsid layers encapsulating a dsRNA genome. The outer capsid layer contains trimers of the P8 protein arranged on a T=13 icosahedral lattice. As such, each asymmetric unit contains 13 P8 proteins, or 4 1/3 unique trimers. The inner capsid layer contains 60 copies of the P3 protein dimer arranged on a T=1 icosahedral lattice. The double layer capsid of RDV has been solved to 6.8Å resolution using cryoEM (Zhou et al., 2001) and subsequently by X-ray crystallography (Nakagawa et al., 2003), and as such both manually segmented and X-ray structures are available for the individual RDV capsid proteins.

Automated segmentation of the virus correctly separates the two capsid shells, identifying the symmetry and capsomeres in each shell (Fig. 2B,C). A single P8 trimer had a 0.74 correlation value compared to the manually segmented trimer (Fig. 2D). Correlation values to the R-trimer determined from X-ray crystallography were very similar, with a correlation value of 0.85 for the automatically segmented trimer and 0.79 for the manually segmented trimer. Other trimers in the asymmetric unit from the automated segmentation also had similar correlation values with those from the crystal structure, ranging from 0.85 to 0.95 with an average correlation value of 0.89. Similar variations in the trimers of RDV from the cryoEM reconstruction were previously noted and might reflect symmetry mismatches between the two capsid layers (Zhou et al., 2001).

Further segmentation of the trimer to its constituent P8 monomers yielded correlation values of 0.8 and 0.84 with the manually segmented cryoEM density map and the X-ray structure of P8, respectively (Fig. 2C,E). As with the P8 trimer, visual assessment of the segmentation appeared to capture all the features in the X-ray structure; only slight deviations between the automated segmentation and X-ray structure could be seen near molecular boundaries. As such, the automated segmentation of a P8 trimer and a P8 monomer proved to be better than manual segmentation methods. Similar results to the P8 segmentation were obtained with the inner capsid protein, P3 which is a structurally non-identical dimer within the asymmetric unit (data not shown).

Segmentation of GroEL

GroEL, which contains 14 identical subunits with D7 symmetry, is a chaperonin responsible for protein folding in bacteria. While several atomic resolution structures for

the entire GroEL complex are known (51 GroEL records in the PDB), it has also been the subject of many structural studies by cryoEM (Ludtke et al., 2004; Ludtke et al., 2001; Ranson et al., 2001; Saibil et al., 2002). Currently, the EBI EMDB houses eleven GroEL structures determined by cryoEM to resolutions between 6Å and 25Å resolution. Of these structures, three GroEL structures, including the highest resolution (6Å), the lowest resolution (25Å) and an intermediate resolution (11.5Å) were subject to automated segmentation (Fig. 3). In both the 6Å and 11.5Å resolution structures, automated segmentation revealed 14 monomers (Fig. 3B,F). At 6Å resolution, the automatically segmented GroEL monomer had a correlation value of 0.76 with the 1GRL X-ray structure (Braig et al., 1994) (Fig. 3C,D), while at 11.5Å resolution the GroEL segmented monomer had a correlation value of 0.75, (Fig 3G,H), similar to the values for the manually segmented monomers. The lack of a better correlation score likely reflects authentic differences between the crystal and cryoEM structures, as previously noted (Ludtke et al., 2004; Ludtke et al., 2001; Ranson et al., 2001). In the 25Å resolution GroEL density map, automated segmentation revealed seven segments; each segment contained two GroEL monomers, one from each of the two heptameric rings (Fig. 3H). Fitting a GroEL monomer from the crystal structure to one subunit of the segmented 25Å resolution cryoEM map of GroEL yielded a correlation value of 0.52. Each segmented subunit contained density equivalent to two GroEL monomers, and as such fitting of a GroEL monomer from the X-ray structure resulted in the relatively low correlation score. However, fitting of two GroEL subunits, adjacent monomers from each heptameric ring, resulted in a correlation value of 0.76. This score, which is similar to the higher resolution automated segmentation of GroEL emphasizes that the segmented subunit in

the 25Å resolution GroEL map does indeed contain density corresponding to two monomers.

Segmentation of P22 Tail Machine

While GroEL and aforementioned viruses are composed of a single structural protein, RDV has only one structural protein per capsid layer, The P22 tail machine is a complex of five proteins, each with different symmetry types, and thus represents a more complex segmentation. As with the other tests structures, the 23Å resolution cryoEM structure of the P22 tail machine was subjected to automated segmentation and analyzed by comparing the segments to known crystal structures (Fig.4). The automated segmentation revealed the five main sub-complexes of the tail machine. While each of these segments has multiple copies of the same proteins, (e.g. six-fold in the gp10 region), automated segmentation was incapable of further segmenting the sub-complexes into the individual protein subunits, likely due to the relatively limited resolution of the map. Nevertheless, fitting of the high-resolution tail spike protein trimer (1TYU) to the tail spike density segment produced a correlation score of 0.76, similar to results from the other aforementioned segmented density maps. Again, the imperfect score may reflect authentic differences between the map and crystal structure or may be related to the intermediate resolution.

Segmentation of 70S Ribosome

Differing from the aforementioned examples, the *E. Coli* 70S ribosome (Fig. 5A) contains no symmetry. The 70S ribosome, which translates mRNA into polypeptide chains, contains 2 distinct halves, the 30S and 50S subunits, each composed of protein

and RNA. Previous attempts at segmenting intermediate resolution cryoEM ribosome data have targeted the isolation of the RNA component from the protein component (Spahn et al., 2000). In this work, rather than segmenting based on the RNA and protein components, automated segmentation was used to define the 50S and 30S halves. The automatically segmented 50S and 30S halves (Fig. 5B) from 70S ribosome density map had correlation values of 0.63 and 0.73 with the 1FFK (Ban et al., 2000) and 1IBM (Ogle et al., 2001), the 50S and 30S X-ray structures from *H. marismourti* and *T. thermophilus*, respectively. As can be seen in Figure 5, the automated segmentation of the 70S into its 50S and 30S halves is qualitatively very good. It appears that errors in segmentation occur primarily in small features near the boundaries between the two ribosomal halves. However, in this case, it is not possible to determine if these differences are indeed biologically significant, structural differences between the species of the map and crystal coordinates or errors in the cryoEM reconstruction.

After establishing the boundaries of 30S and 50S halves, two more rounds of segmentation were done using the automated algorithm. In the 30S, automated segmentation of the RNA from protein was carried out (Fig. 4C). As in separating the two 30S and 50S subunits from the 70S ribosome, the automated segmentation was able to delineate between the protein and RNA with reasonable accuracy (correlation value of 0.66, zoomed in view of the segmented RNA is shown in Fig. 4I). Comparison of the segmented protein and RNA (Fig. 4D,G respectively) with the density simulated from the X-ray structure of the 30S subunit (Fig. 4E,H respectively) illustrates the ability of the automated segmentation routine to correctly extract the complex densities. Further automated segmentation of the individual protein constituents in the protein half of the

previous 30S segmentation was done, resulting in the complete identification of approximately six of the twenty subunits (three subunits are shown in Fig. 4F). However, segmentation of the tRNA from the RNA component was not possible.

Discussion

In general, automated segmentation performed well, accurately identifying the monomeric subunits and defining their boundaries in a variety of different types of cryoEM density maps and at different resolutions (Table 1). Additionally, this automated segmentation routine isolates individual subunits from the entire complex at once. In manual segmentation, it is likely that multiple segments need to be generated, each requiring several rounds of refinement. While automated segmentation has an initial cost, up to 10 hours (SGI Onyx2 with a 400 MHz MIPS R12000) for large maps (~720^3 voxels), an iterative manual segmentation could far exceed this amount of time depending on number of subunits to be isolated. In such, this automated segmentation algorithm is a useful tool in eliminating or reducing the tedious and subjective process of manual subunit identification.

Based on both quantitative and qualitative assessments of the examples shown here, the automated segmentation routine is equivalent to or exceeds the segmentations done manually. From the tested data sets, it also appears that global and local symmetry has little effect on the quality of segmentation. The reported correlation values reflect the relative fidelity at which the automated algorithm functions. While these values are quite good, they are not perfect. This could be due to several factors. First, the manually segmented subunits, as already described, are subject to user bias. Errors in manual segmentation may result in less than optimal correlation values. Furthermore, in the

ribosome example, the X-ray structure and the cryoEM density map are from two different organisms. Differences between these ribosomes may manifest as structural differences and thus effect the overall correlation of the subunits to the corresponding structure. Additionally, when comparing with the X-ray structure, errors could arise from not only intrinsic errors, but may also reflect subtle differences between the cryoEM map and X-ray structure as in the case of GroEL (Ludtke et al., 2004; Ludtke et al., 2001). While the nature of these differences cannot be accurately assigned, it should be noted that the automated method performed at least as good as or better than manual segmentation in our four test data sets.

For the most part, the errors seen in the automated segmentation were small and close to the boundary or edges regions between neighboring subunits in the density map. These regions, in general, are the most difficult to segment and most prone to user bias. Additionally, the quality of the map and possible resolution may vary in such regions and thus pose a problem during classification in our procedure. These errors may not be negligible and indeed reflect details germane to the interpretation of the density map. As such, automated segmentation provides a good first step in the subunit identification and segmentation, which could be further augmented by manual segmentation utilizing prior biochemical or structural knowledge.

In complex cryoEM structures, like the ribosome and viruses with multiple capsid layers, the progressive segmentation of individual pieces is most likely optimal in segmenting the individual components. By creating a hierarchical segmentation pattern, from the most distinguishable to least distinguishable features, it is possible to assess the fidelity and/or quality of the automated routine. In the case of the 70S ribosome, the

individual 30S and 50S subunits, as well as the RNA component were segmented. Only a small portion of the 30S proteins were correctly and completely isolated. However, this level of subunit segmentation is on par with the original citation. As such, a hierarchical search allowed for a maximal extraction of information from the map and is likely the best approach when dealing with complicated, low-symmetry macromolecular machines.

A potential limitation to the segmentation of individual components is the resolvability of individual densities in the map. Based on the GroEL example, this automated segmentation routine appears to be as effective in identifying individual subunits at 11.5Å resolution as at 6.5Å resolution. At lower resolution, automated segmentation preformed nearly as well, although it did identify two GroEL subunits per segment (one GroEL monomer from each ring). Additionally, the 23Å resolution P22 tail machine was correctly segmented into its constituent components, although it was not possible to segment out the individual proteins from the five individual segments. In both of these cases, the cryoEM density had sufficient resolution to identify the oligomeric sub-complexes, but insufficient resolution to correctly segment the individual protein subunits. As such, segmentation is possible at nearly any resolution provided that individual subunits or complexes with the macromolecular complex can be identified. It is important to note here that any form of segmentation is capable of defining structural subunits, however the fidelity of the segmentation is governed by the ability to resolve features within the density map itself. In this respect, the overall accuracy of segmentation, including the automated segmentation routine described here, is ultimately limited by map quality and resolution.

Conclusion

This work has demonstrated the feasibility and accuracy of an automated mass density segmentation routine in both low and intermediate resolution cryoEM density maps. While small errors persist in the segmentation, automated segmentation could provide at a minimum a good initial step at segmenting an entire macromolecular complex, and thereby eliminating the tedious and subjective nature of manual segmentation. In practice, an experienced investigator should look at the segmented map and fine tune it with other knowledge about the components of interest. Therefore, the proposed technique should not be viewed a complete replacement of human intervention in the discovery process. Nevertheless, as the number of low to intermediate resolution cryoEM structures grow, such automated techniques shall become invaluable in the rapid structural analysis of complex structures.

Acknowledgments

We would like to thank Dr. Steve Ludtke and Dr. Donghua Chen for providing the GroEL density maps and Dr. Wen Jiang for his helpful discussions. The research of authors MLB and WC was supported by NSF (EIA-0325004) and NIH (P41RR02250 and P20RR020647). The research of authors ZY and CB was supported by NSF (EIA-0325550, CNS-0540033) and NIH (P20 RR020647, R01GM074258, R01GM073087).

References

Alberts, B., 1998. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. Cell 92, 291-294.

Alberts, B., Miake-Lye, R., 1992. Unscrambling the puzzle of biological machines: The importance of the details. Cell 68, 415-420.

Bajaj, C., Yu, Z.,2005. Geometric processing of reconstructed 3d maps of molecular complexes. In Handbook of computational molecular biology, S. Aluru, ed. (Chapman & Hall/CRC Press).

Bajaj, C., Yu, Z., Auer, M., 2003. Volumetric feature extraction and visualization of tomographic molecular imaging. J. Struct. Biol. 144, 132-143.

Ban, N., Nissen, P., Hansen, J., Moore, P. B., Steitz, T. A., 2000. The complete atomic structure of the large ribosomal subunit at 2.4 a resolution. Science 289, 905-920.

Bottcher, B., Wynne, S. A., Crowther, R. A., 1997. Determination of the fold of the core protein of hepatitis b virus by electron cryomicroscopy. Nature 386, 88-91.

Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D. C., Joachimiak, A., Horwich, A. L., Sigler, P. B., 1994. The crystal structure of the bacterial chaperonin groel at 2.8 Å. Nature 371, 578-586.

Chiu, W., Baker, M. L., Jiang, W., Dougherty, M., Schmid, M. F., 2005. Electron cryomicroscopy of biological machines at subnanometer resolution. Structure (Camb) 13, 363-372.

Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J., Steven, A. C., 1997. Visualization of a 4-helix bundle in the hepatitis b virus capsid by cryo-electron microscopy. Nature 386, 91-94.

Frangakis, A. S., Forster, F., 2004. Computational exploration of structural information from cryo-electron tomograms. Curr. Opin. Struct. Biol. 14, 325-331.

Frangakis, A. S., Hegerl, R., 2002. Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis. J. Struct. Biol. 138, 105-113.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141-147.

He, J., Schmid, M. F., Zhou, Z. H., Rixon, F., Chiu, W., 2001. Finding and using local symmetry in identifying lower domain movements in hexon subunits of the herpes simplex virus type 1 b capsid. J. Mol. Biol. 309, 903-914.

Jiang, W., Baker, M. L., Ludtke, S. J., Chiu, W., 2001. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. J. Mol. Biol. 308, 1033-1044.

Jiang, W., Li, Z., Baker, M. L., Prevelige, P. E., Chiu, W., 2003. Coat protein fold and maturation transition of bacteriophage p22 seen at subnanometer resolution. Nat. Struct. Biol. 10, 131-135.

Ludtke, S. J., Chen, D. H., Song, J. L., Chuang, D. T., Chiu, W., 2004. Seeing groel at 6 a resolution by single particle electron cryomicroscopy. Structure (Camb) 12, 1129-1136.

Ludtke, S. J., Jakana, J., Song, J. L., Chuang, D. T., Chiu, W., 2001. A 11.5 a single particle reconstruction of groel using eman. J. Mol. Biol. 314, 253-262.

Malladi, R., Sethian, J. A.,1998. A real-time algorithm for medical shape recovery. In Proceedings of international conference on computer vision), Vol, pp. 304-310.

Masuda, T., Yamamoto, K., Yamada, H., 1993. Detection of partial symmetry using correlation with rotated-reflected images. Pattern Recognition 26, 1245-1253.

Nakagawa, A., Miyazaki, N., Taka, J., Naitow, H., Ogawa, A., Fujimoto, Z., Mizuno, H., Higashi, T., Watanabe, Y., Omura, T., Cheng, R. H., Tsukihara, T., 2003. The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. Structure (Camb) 11, 1227-1238.

Ogle, J. M., Brodersen, D. E., Clemons, W. M., Jr., Tarry, M. J., Carter, A. P., Ramakrishnan, V., 2001. Recognition of cognate transfer rna by the 30s ribosomal subunit. Science 292, 897-902.

Perona, P., Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Trans on Pattern Analysis and Machine Intelligence 12, 629-639.

Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B., Fenton, W. A., Horwich, A. L., Saibil, H. R., 2001. Atp-bound states of groel captured by cryo-electron microscopy. Cell 107, 869-879.

Saibil, H. R., Ranson, N. A., 2002. The chaperonin folding machine. Trends Biochem. Sci. 27, 627-632.

Sali, A., 2003. NIH workshop on structural proteomics of biological complexes. Structure (Camb) 11, 1043-1047.

Sethian, J. A., 1996. A fast marching level set method for monotonically advancing fronts. Proc. Nat. Acad. Sci. USA 93, 1591-1595.

Sethian, J. A., 1999. Level set methods and fast marching methods (2nd edition) (Cambridge University Press).

Sewell, B. T., Best, R. B., Chen, S., Roseman, A. M., Farr, G. W., Horwich, A. L., Saibil, H. R., 2004. A mutant chaperonin with rearranged inter-ring electrostatic contacts and temperature-sensitive dissociation. Nat. Struct. Mol. Biol. 11, 1128-1133.

Sifakis, E., Tziritas, G., 2001. Moving object localization using a multi-label fast marching algorithm. Signal Processing: Image Communication 16, 963-976.

Spahn, C. M., Penczek, P. A., Leith, A., Frank, J., 2000. A method for differentiating proteins from nucleic acids in intermediate-resolution density maps: Cryo-electron microscopy defines the quaternary structure of the escherichia coli 70s ribosome. Structure 8, 937-948.

Tang, L., Marion, W. R., Cingolani, G., Prevelige, P. E., Johnson, J. E., 2005. Three-dimensional structure of the bacteriophage p22 tail machine. Embo J. 24, 2087-2095.

Valle, M., Zavialov, A., Li, W., Stagg, S. M., Sengupta, J., Nielsen, R. C., Nissen, P., Harvey, S. C., Ehrenberg, M., Frank, J., 2003. Incorporation of aminoacyl-trna into the ribosome as seen by cryo-electron microscopy. Nat. Struct. Biol. 10, 899-906.

Volkmann, N., 2002. A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. J. Struct. Biol. 138, 123-129.

Weickert, J.,1998. Anisotropic diffusion in image processing (ECMI Series, Teubner, Stuttgart, ISBN 3-519-02606-6).

Yu, Z., Bajaj, C., 2005. Automatic ultrastructure segmentation of reconstructed cryo-em maps of icosahedral viruses. IEEE Transactions on Image Processing 14, 1324-1337.

Zhou, Z. H., Baker, M. L., Jiang, W., Dougherty, M., Jakana, J., Dong, G., Lu, G., Chiu, W., 2001. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. Nat. Struct. Biol. 8, 868-873.

Table 1. The results of the automated segmentation procedure are summarized. The subunit represents the monomeric subunit obtained from automated segmentation. The model represents the high resolution structure or the previously segmented monomer. The correlation value represents the cross correlation value (0-1) between the subunit and model.

Subunit	Model	Correlation
Bacteriophage P22, gp5 6-fold	Manual segmentation, gp5 6-fold	0.79
Bacteriophage P22, gp5 5-fold	Manual segmentation, gp5 5-fold	0.72
RDV P8 trimer	Manual segmentation, P8 trimer	0.74
RDV P8 trimer	X-ray structure, P8 trimer	0.85
RDV P8 monomer	Manual segmentation, P8 monomer	0.80
RDV P8 monomer	X-ray structure, P8 monomer	0.84
GroEL monomer, 6Å resolution	X-ray structure, GroEL monomer	0.76
GroEL monomer, 11.5Å resolution	X-ray structure, GroEL monomer	0.75
GroEL dimer, 25Å resolution	X-ray structure, GroEL monomer	0.52
GroEL dimer, 25Å resolution	X-ray structure, GroEL dimer	0.76
P22 tail machine	X-ray structure, P22 trimeric tail spike	0.76
70S ribosome, 50S subunit	X-ray structure, 50S subunit	0.63
70S ribosome, 30S subunit	X-ray structure, 30S subunit	0.73
30S subunit, RNA	X-ray structure, 30S subunit RNA	0.66

Figures

Figure 1 Segmentation of 9.5Å resolution structure of the mature bacteriophage P22. A Volume rendering of the P22 density map is shown in (A, EMDB ID: 1101). Detected global icosahedral symmetry and local 6-fold symmetry axes are shown in (B). Two segmented capsomeres are shown in (C). A capsomere at the icosahedral five-fold, with five identical copies of gp5, is shown in maroon. The second capsomere, in cyan, is located at the local six-fold and consists of 6 unique gp5 subunits. Segmented gp5 subunits about the icosahedral 5-fold axis are shown in (D). A single averaged gp5 subunit from the five-fold capsomere is shown in (E). Segmented gp5 subunits about the local 6-fold axis are shown in (F). A single averaged monomer from the local six-fold axis capsomere is shown in (G). A comparison of the average gp5 subunits about the local six-fold axis from manual segmentation (blue mesh) and the automated segmentation (grey density) are shown in (H).

Figure 2 Segmentation of 6.8Å resolution structure of rice dwarf virus (RDV). A volume rendering of the RDV density map is shown in (A, EMDB ID: 1060). Segmentation of the outer capsid layer asymmetric unit is shown in (B). Each color represents one of the 4 1/3 unique trimers. A segmented trimer (grey) and the corresponding segmented P8 monomer (orange) are shown in (C). A comparison of the manually segmented trimer (blue mesh) and automatically segmented trimer (grey density) are shown in (D). The automatically segmented P8 monomer (grey density) is shown in comparison to the corresponding X-ray structure of RDV P8 (E).

Figure 3 Segmentation of GroEL at different resolutions. The original density maps of the 6 Å (A, EMDB ID: 1080), 11.5Å (E, EMDB ID: 1081) and 25Å (I, EMDB ID: 1095)

resolution GroEL structures are shown in side views. The automated segmentation of GroEL at the three resolution are shown for the 6 Å (B), 11.5Å (F) and 25Å (J) resolution GroEL structures. A single automatically segmented subunit is shown for the 6 Å (C), 11.5Å (G) and 25Å (K) resolution GroEL structures. Fitting of the GroEL X-ray structure to the 6Å resolution, 11.5Å and 25Å resolution automatically segmented subunit from GroEL are shown in (D,H, L), respectively.

Figure 4 Segmentation of the 23Å resolution bacteriophage P22 tail machine. (A) Top and side views of the P22 tail machine are shown (EMDB ID:1119). (B) Top and side views of the automatically segmented P22 tail machine. Individual components are colored (tail spike in pink, portal in blue, gp4 in green, gp10 in magenta and gp26 in cyan). (C) A single tail spike, indicated with an arrow in (B), is shown fitted with the corresponding trimeric tail spike X-ray structure (1TYU).

Figure 5 Segmentation of a cryoEM map of no symmetry at 10Å resolution. (A) The reconstructed cryoEM maps of 70S ribosome aminoacyl-tRNA from E. coli complex (EMDB ID: 1056). (B) 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 (C). Comparisons of automatic segmentation and the X-ray structure of the 30S subunit are shown in (D-J). The segmented density for the protein (D) and RNA (G) is shown next to the protein (red, E) and RNA (red, H) from the X-ray structure rendered at ~10Å resolution. A third round of automated segmentation on the protein component of the 30S subunit successfully revealed several individual components (mesh, F). No additional segmentation was done on the RNA component of the 30S subunit. The X-ray structure

of the protein components and RNA are also shown superimposed on the segmented density in (F,J). The arrows in (D,G) indicate the regions depicted in (F,I).

Figure 1

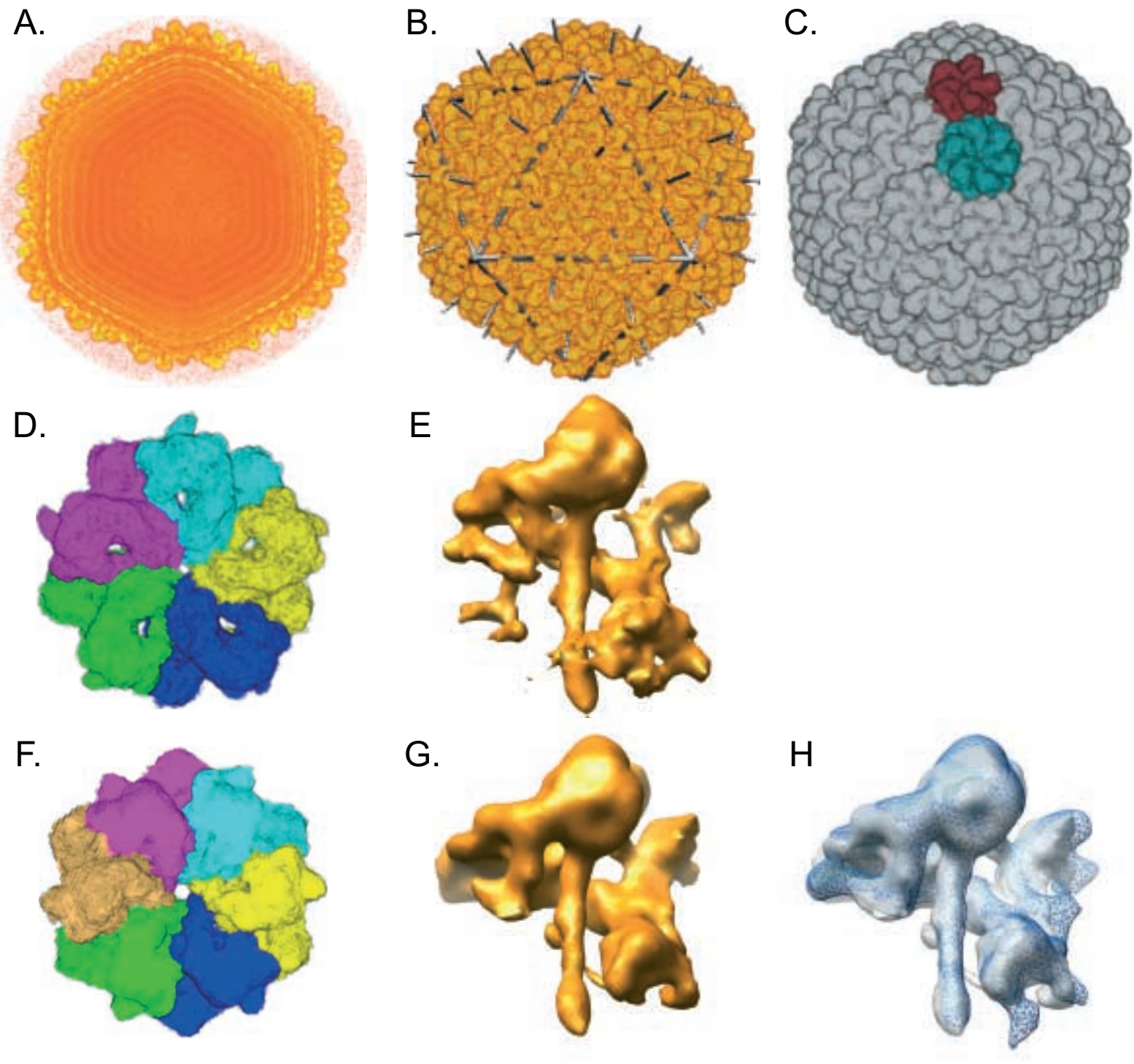

Figure 2

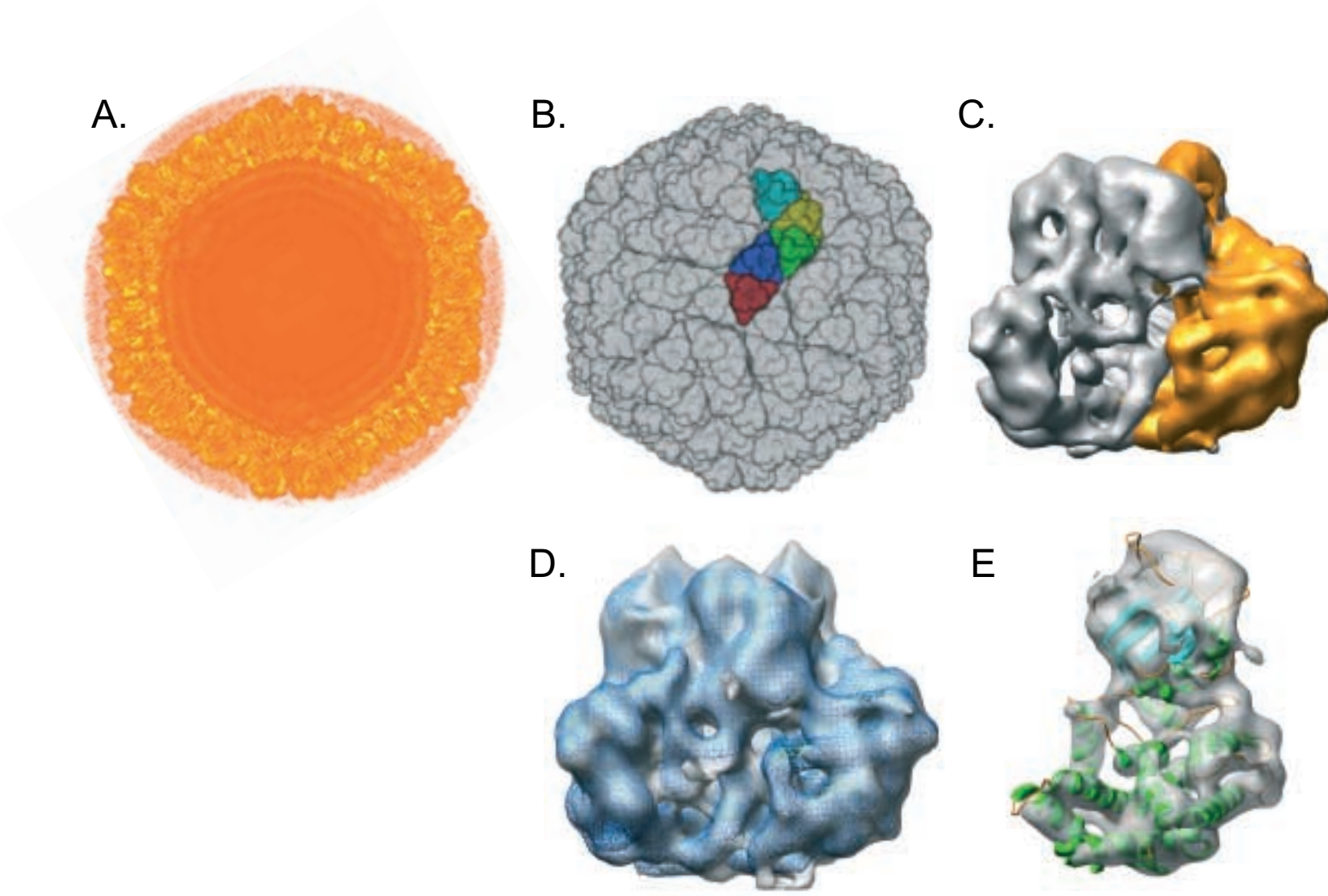

Figure 3

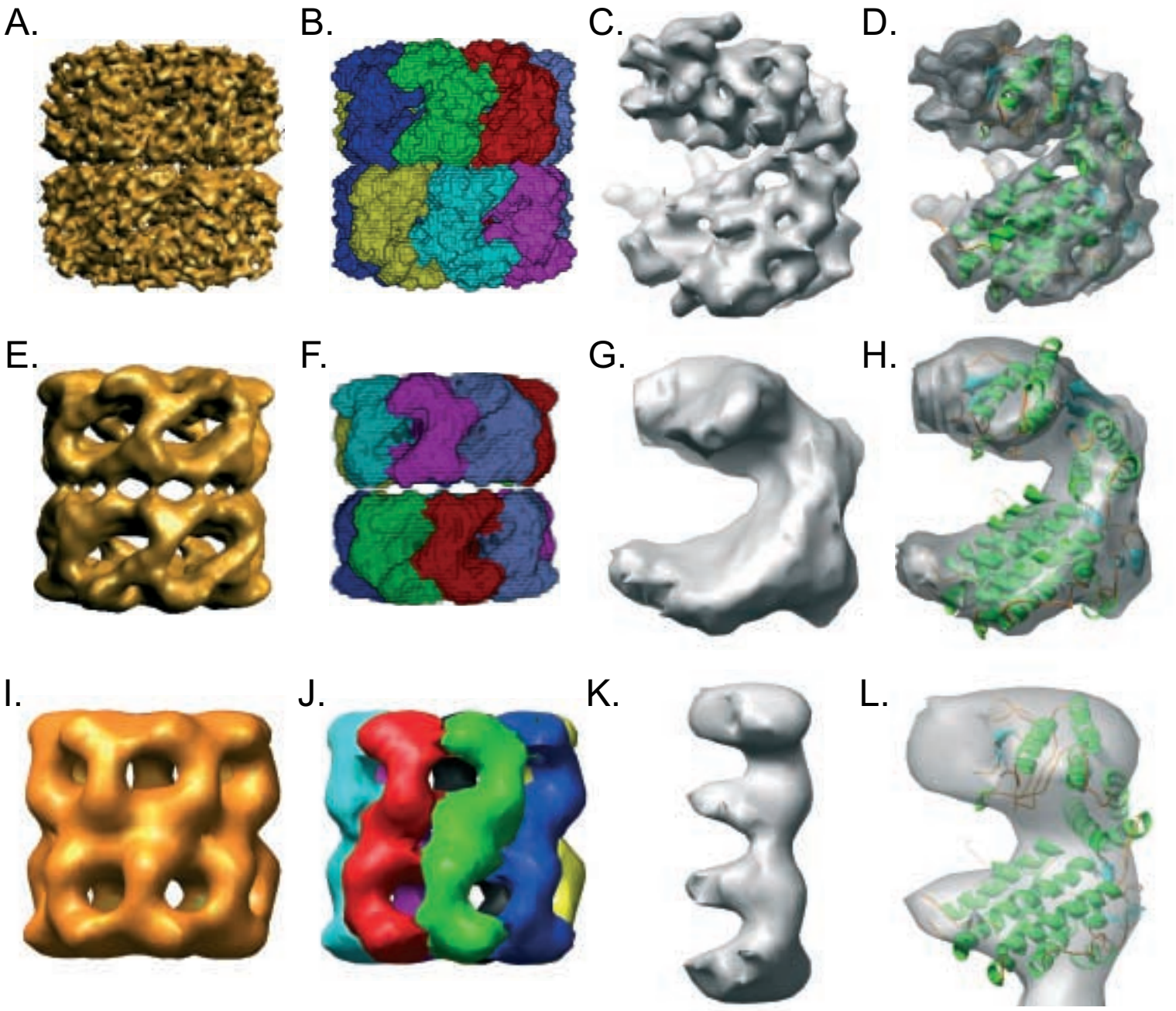

Figure 4

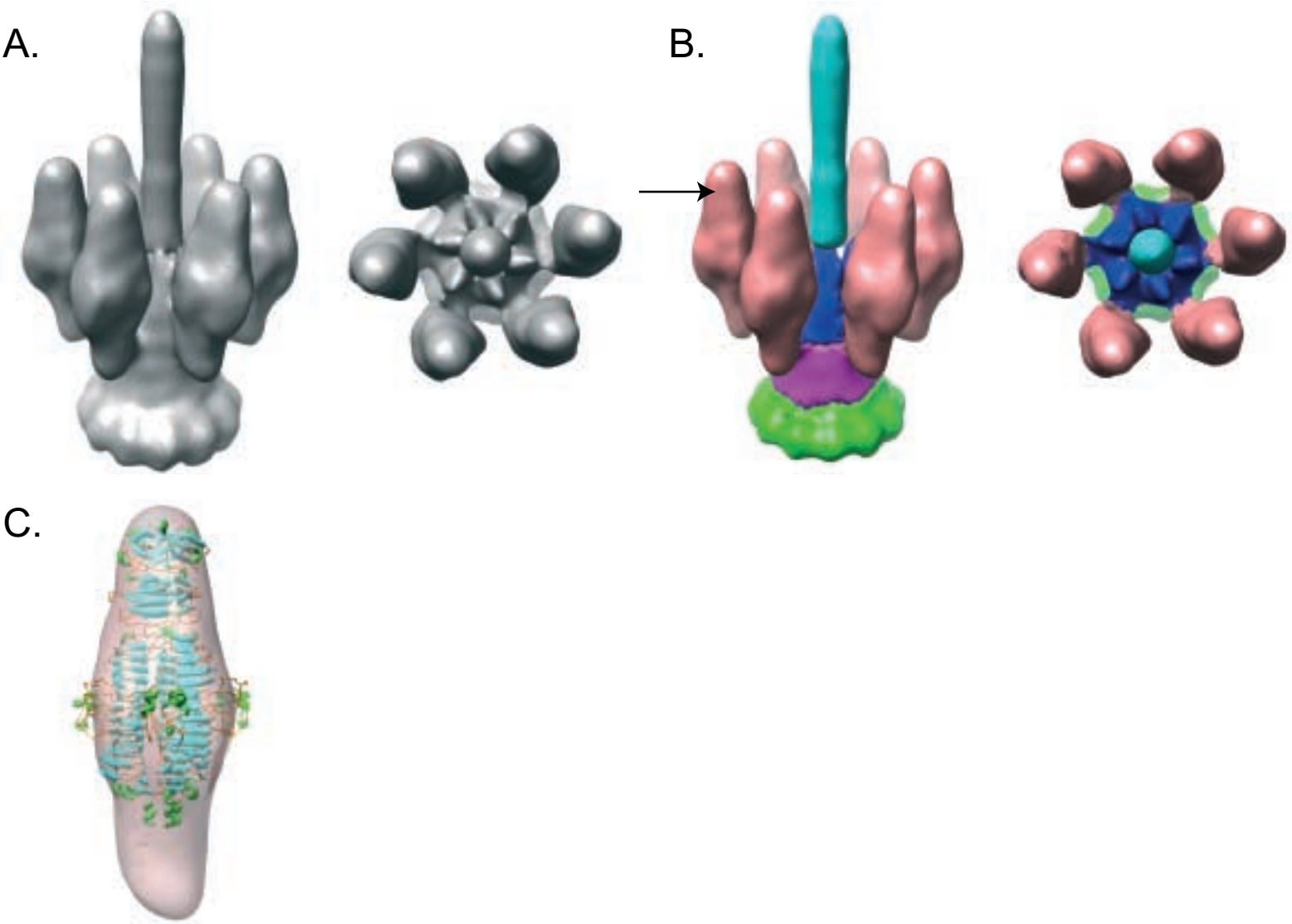

Figure 5

Click here to download high resolution image