# Computational Approaches for Automatic Structural Analysis of Large Bio-molecular Complexes

Zeyun Yu, *Student Member, IEEE,* Chandrajit Bajaj, *Member, IEEE,*

***Abstract*— We present computational approaches for analyzing/interpreting three-dimensional (3D) structures of large bio-molecular complexes at intermediate or moderate resolutions (5Å-20Å). Two problems are addressed: one is the 3D alignment (matching) between segmented subunits and the other is the secondary structure identification within subunits. For the first problem, we propose a fast algorithm to spatially correlate two subunits. A number of applications are discussed and illustrated, including measuring similarities, averaging subunits, refining segmentation, and fitting atomic structures. For the second problem, we propose an efficient algorithm to detect the secondary structures, including both alpha-helices and beta-sheets, of a protein density map. The performance of our approaches is demonstrated on both experimentally reconstructed virus maps and computationally simulated protein density maps.**

***Index Terms*— Structure Analysis, Alignment, Similarity Measure, Segmentation, Secondary Structure Detection, Skeletonization, Cryo-EM Maps, 3D Reconstruction**

## I. INTRODUCTION

COMPUTATIONAL biology has emerged as an increasingly active field in recent years due to the great demands of computational approaches in solving various biological problems. We are now moving from a genomic to a proteomic era, meaning that we are now facing a more challenging task than analyzing the genome sequences: i.e., understanding of how proteins function in and around cells. The 3D structures of proteins are the essential factors that "code" the functions of proteins. Therefore, determining and analyzing the 3D structures of proteins and in particular large biological complexes is extremely important for studying their functions.

X-ray crystallography [1], [2] and nuclear magnetic resonance (NMR) [3], [4] are two primary techniques that have been used to reveal the structures of most existing protein folds seen in the Protein Data Bank (PDB) [5]. While individual proteins or small complexes provide important information, they do not give a full picture of a functional biological complex. Structural determination of large biological complexes (e.g., viruses, ion channels, and the ribosome) therefore offers a more complete structural and functional description of the protein machinery. Knowledge of these structures would provide not only the mechanistic descriptions for how macromolecules interact in an assembly but also clues in developing therapeutic interventions related to diseases. While x-ray crystallography and NMR spectroscopy are quite often restricted to relatively small biological structural units, cryo-electron microscopy (cryo-EM) of single particles, has steadily become a powerful tool in revealing structures of large bio-molecular complexes [6], [7]. The rapid development of this technique has made it possible to resolve the biological complexes at sub-nanometer resolutions (6Å-10Å) [8], [9], [10], [11]. In addition, cryo-EM approach can also capture multiple functional states of a complex, allowing us to directly study the dynamics of an interacting protein machinery.

Signal/Image processing techniques have been extensively used in 3D structure reconstructions at cellular and molecular levels. In particular, most of modern signal/image processing algorithms, including image restoration, noise reduction, contrast enhancement, object detection, alignment, classification, 3D reconstruction, boundary segmentation, skeletonization, and so on, have found fruitful applications in the single particle cryo-EM approach, as will be explained with more details in Section II. Although a number of software for performing 3D reconstructions from cryo-EM data have been made widely available in recent years (e.g., EMAN [12], SPIDER [13], IMAGIC [14]), quantitative and automatic analysis/interpretation of the reconstructed bio-molecular assemblies still remains undeveloped. Current ways for interpreting reconstructed maps depend mainly upon visual inspections with the help of various graphic tools. Due to the large physical size and complexity of the bio-molecular assemblies, however, it is not only tedious and subjective but also very difficult to visually interpret the detailed features/activities of an interacting bio-molecular system. For this reason, automatic structural analysis of large bio-molecular complexes has become increasingly and critically important.

In our previous work [15], we presented an automatic approach of segmenting the reconstructed maps of bio-molecular complexes into dozens to thousands of individual subunits. The automatic segmentation without doubt makes it much easier to interpret the bio-molecular assemblies, allowing us to isolate the subunits and interpret each of them individually without interference from the others. For symmetric structures (such as most viruses), the segmentation also helps eliminate the structural redundancy. In the present paper, we will address two follow-up tasks. The first is to find the spatial alignment and structural similarity between two segmented subunits, based on which a couple of applications will be discussed. The second task that we shall address is the secondary structure identification within a segmented subunit. The demand of locating alpha-helices and beta-sheets becomes increasingly critical as the resolution of the reconstructed maps approaches higher than 10Å [8]. Knowing the exact positions and orienta-

---

Z. Yu and C. Bajaj are with the Computational Visualization Center, Department of Computer Sciences and The Institute of Computational Engineering and Sciences, The University of Texas at Austin, 1 University Station C0500, Austin, Texas 78712, USA. (emails: zeyun.yu@gmail.com, bajaj@cs.utexas.edu).

tions of secondary structures can help us extract the maximum biological information from the reconstructed cryo-EM maps.

The rest of this paper is organized as follows. We start in Section II with a brief introduction to 3D electron microscopy imaging techniques. Then we discuss how the segmented subunits are aligned in 3D space and a number of applications will be shown in Section III. Our algorithm of secondary structure detection is discussed and results on both reconstructed and simulated maps are demonstrated in Section IV. Finally we conclude this paper in Section V.

## II. 3D CRYO-ELECTRON MICROSCOPY IMAGING

Electron microscopy (EM) imaging has been largely used in structural biology to study the activities of cells and organelles. Three-dimensional electron microscopy (3D-EM) imaging plays a unique role in EM for its capability of revealing the three dimensional structures of biological units. The mathematical principle of 3D-EM reconstruction from projection data (experimental EM images) is basically the same as that commonly used in 3D medical imaging (Computed Tomography or CT scan, for example). The major difficulty with EM images is due to the extremely low signal-to-noise ratio (SNR). This is true partly because the electron dose used in EM imaging has to be kept in a very low level (approximately $0.5{\sim}4$ e/$\mathring{A}^2$) in order to reduce the radiation damages of electrons to the specimen. The flash cooling technique, known as cryo-EM, is to quickly cool the samples to liquid nitrogen temperature (about 77K or less) such that the surrounding water does not form crystalline ice. This technique proved to be very successful in preserving the native structures of specimen while reducing the radiation damages [6], [16].

Due to different sample preparation and data collection methods, 3D-EM includes three major techniques: electron crystallography, electron tomography, and single particle method. Electron crystallography [17], similar to x-ray crystallography, can reveal the bio-molecular structures at an atomic resolution. However, the weakness of this technique is that a two-dimensional crystal has to be grown for cryo-EM imaging, which in many cases is hard to do. Electron tomography [18], [19], on the other hand, is a powerful and unique tool for studying 3D ultra-structures of cell organelles or whole cells at a relatively low resolution. Mathematically, the 3D structure of a cell specimen can be reconstructed from a series of 2D projections in different tilt angles. There are several methods to collect the projection data: single-axis tilt, double-axis tilt, and uniform conical tilt, depending on how the specimen is rotated under the fixed camera. However, the tilt angles cannot exceed certain degree (usually $\pm 70^0$) due to the tilt-stage and specimen thickness limitations [20], [21]. For this reason, the reconstructed density maps always have significant distortions in certain regions, commonly known as the missing wedge (pyramid, or cone) problem. Besides this problem, the resolution of this type of reconstructions is often limited in the range of $20\mathring{A} \sim 200\mathring{A}$ because of the limited electron dose used on a single specimen due to the radiation damages [18], [19].

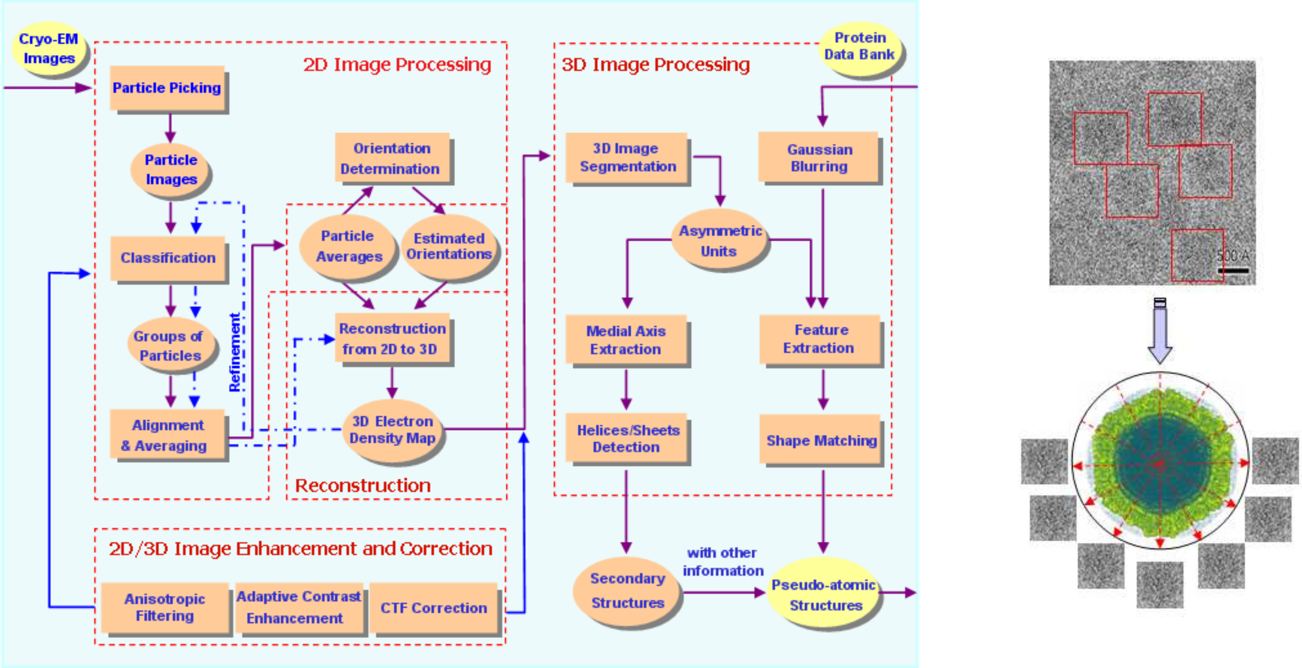
The solution to remedy the above problems is by taking averages of a number of structurally identical particles in random orientations. On one hand, the average of particles in the same orientation (after proper alignment) can improve the signal-to-noise ratio (SNR). On the other hand, particles in different orientations are very likely complementary to each other such that the missing wedge (pyramid, or cone) of one particle could be filled by another. This technique, known as *single particle cryo-EM reconstruction*, can achieve a resolution of about $6\mathring{A}$ [8], [22]. To reduce the radiation damages, each particle is imaged only from one tilt angle and thousands of (or more) particles are used to reconstruct a single 3D structure. Fig. 1(a) shows an overall pipeline of single particle reconstruction and analysis. Each of the steps will be explained below in more details.

### A. Particle Picking

Thousands of (or more for non-symmetric structures) particle images need to be collected in order to reconstruct a 3D density map at moderate to intermediate resolutions ($20\mathring{A} \sim 6\mathring{A}$) [8]. Usually each electron micrograph contains dozens to hundreds of particles, each of which corresponds to one projection of the 3D biological object under investigation. The goal of particle picking is to box out all particles that look reasonably good in both size and shape. Fig. 1(b) shows a small portion of the electron micrograph of the rice dwarf virus [8]. From this picture we can also see how noisy a typical particle image looks. The particles can be boxed out manually and there are some software which provide this function (e.g., EMAN [12]). This method however is subjective and becomes very tedious when tens of thousands of particles need to be picked. Fortunately there are a number of computational approaches for automatic or semi-automatic particle picking. Most of these methods utilize cross-correlation scoring (e.g., [26]) while some of them are based on feature/edge detection (e.g., [27]). Interested readers can refer to a good review paper [28] or a special issue of *Journal of Structural Biology* [29] for various methods on this topic.

### B. Particle Classification and Alignment

Among the whole set of picked particles, some of them may appear the same (subject to in-plane rotations) while some others look completely different. Particle classification is to find all the particle images that have the same appearance but not necessarily in the same orientation, while particles alignment is to align those particles in the same class to the same orientation. The classified and aligned particles can be averaged such that the class average has a significantly improved signal-to-noise ratio (SNR), compared to the original individual particles images. In principle, averaging a number of particle images is the core idea of single particle reconstruction approach, in order to achieve a sub-nanometer resolution from extremely noisy data. Although particle classification and alignment are conceptually different, they are actually dependent upon each other $-$ the alignment is performed on classified particle images but in return, correctly aligned particles also help classify the particle images more accurately. There are reference-based and reference-free methods, both of which are based on the conventional cross-correlation and



(a) Overall pipeline of the single particle reconstruction and analysis (b) Particle images/reconstruction

Fig. 1. Illustrations of single particle cryo-EM technique. (a) The overall pipeline, including particle picking, particle classification/alignment, orientation assignment, 3D reconstruction, and map interpretation. In addition, EM contrast transfer function (CTF) correction [23], anisotropic filtering [24], and adaptive contrast enhancement [25] may also be applied before or after 3D reconstructions, to improve the signal-to-noise ratio. (b) An example of electron micrography images showing particle picking (top) and 3D reconstruction from 2D projections (bottom).

Fourier analysis [21]. More recently, the probability-based methods, such as the maximum likelihood approach [30], [31]. became increasingly popular for their better performance in classifying and aligning particle images.

### C. *Orientation Assignment*

Particle classification and alignment can provide us with a series of class averages corresponding to projections in different orientations. Unfortunately in most cases the particles (henceforth the class averages) appear in randomly-chosen orientations such that we do not have the clue of which orientation each class average corresponds to. Mathematically the relative orientations between two projections of the same 3D object can be determined by the *common line theorem* [7]. However, this method is very time-consuming and the accuracy may not be good enough due to the noise. Another approach for estimating orientations is based on an initial 3D model [12]. The initial model can be built from a representative subset of projections whose orientations are determined by the common line method or randomly-assigned Euler angles. From the initial model, one can compute a series of projections by re-projecting the 3D model in different angles. The class averages can be compared against each of the projections and assigned with the orientation of the best-matched projection. The advantage of the model-based approach is that the orientation assignment is quite straightforward and additionally the particle classification and alignment can be simultaneously conducted during the orientation assignments.

### D. *Reconstruction and Refinement*

The mathematical theorem is well established for 3D reconstruction from 2D projections, given that the orientation of each projection is known. Fig. 1(b) illustrates the 3D reconstruction from a series of projections in different angles. The most popular methods include the direct Fourier space reconstruction [12] and the real-space filtered back projection method [7]. As we mentioned earlier, the reconstructed 3D model can be used as an initial model to refine the particle classification/alignment as well as the orientation assignment. The refined class averages and their orientations are then used to generate a better 3D model. This iterative process is repeated until no significant improvement is observed.

### E. *Map Interpretation*

The reconstructed maps do not convey meaningful information until they are correctly interpreted. The map interpretation is usually performed in two directions as shown in Fig. 1(a). For maps at intermediate resolution ($6\AA - 10\AA$), the secondary structures are visible and detectable. A pseudo-atomic model can be built based on the detected secondary structures and their topological connections [8]. When the resolutions of the reconstructed maps degrade beyond $10\AA$, however, we cannot see the secondary structures with high confidence but we still can build the pseudo-atomic models by fitting a high-resolution PDB structure into the cryo-EM map based only on the density distributions [32]. In both cases, the segmentation of meaningful subunits is very helpful for fast and accurate interpretations [15]. In the following sections, we will present a few computational approaches on automatic structural interpretations but major efforts will be made to solve two problems: 3D structural alignment between subunits and secondary structure identification within subunits.

## III. 3D STRUCTURE ALIGNMENT

The goal of 3D structure alignments is to find the transformation matrix from one subunit to another, such that the two subunits are best matched according to a pre-defined similarity scoring function. In the following we will first discuss how to define the similarity score between two density maps. Then we propose a fast algorithm for aligning two subunits, followed by a couple of applications in structural analysis of cryo-EM maps.

### A. *Similarity Scoring Function*

The traditional similarity scoring function between two scalar maps, denoted by $f$ and $g$, is defined by cross-correlation as follows:

$$S^1_{f,g}(T) = \sum_{i,j,k} f(i,j,k) \times g(T(i,j,k)), \tag{1}$$

where $T$ is the transformation matrix from $f$ to $g$. The 3D structural alignment is to find the best $T$, given two maps $f$ and $g$, such that the above-defined similarity function reaches the highest. It is easy to see that maximizing $S^1_{f,g}(T)$ is equivalent to minimizing the square difference between $f$ and $g$:

$$S^2_{f,g}(T) = \sum_{i,j,k} (f(i,j,k) - g(T(i,j,k)))^2. \tag{2}$$

The advantage of the cross-correlation method is that the fast Fourier transform (FFT) can be employed to speed up the searching in the translational space (three degrees of freedom). As we will see shortly later, the FFT method may not be the best way in our case. Therefore, we shall define the similarity scoring function in real space. However, in order to speed up the searching process, we define the similarity score on a set of critical points, instead of the whole set of voxels. The critical points are those that best capture the features of the molecular density map. In most cases, the critical points can be defined by the local maxima, local minima, and saddle points of the scalar map. If the data is noisy, one may consider preprocessing the map using gradient vector diffusion as discussed in [15]. Our similarity scoring function is hence defined in the following way:

$$S^3_{f,g}(T) = 1 - \frac{\sum_{m=1}^{M} |f(c_m) - g(T(c_m))| + \sum_{n=1}^{N} |f(T^{-1}(d_n)) - g(d_n)|}{\sum_{m=1}^{M} max\{f(c_m),\ g(T(c_m))\} + \sum_{n=1}^{N} max\{f(T^{-1}(d_n)),\ g(d_n)\}}, \tag{3}$$

where $c_m, m = 1, 2, \cdots, M$, are critical points of $f$ and $d_n, n = 1, 2, \cdots, N$, are critical points of $g$. Fig.2 illustrates the idea of this similarity scoring function. There are two major differences between $S^3_{f,g}(T)$ and $S^2_{f,g}(T)$. First, the new similarity scoring function is based only on the critical points. Therefore, the searching for the best $T$ using $S^3_{f,g}(T)$ is much faster than the searching using $S^2_{f,g}(T)$, as the latter one is based on all the voxels of $f$ and $g$. Secondly, the scoring function of $S^3_{f,g}(T)$ is normalized such that the similarity scores are always scaled to the range of [0, 1], where 0 means no similarity and 1 means the highest similarity.

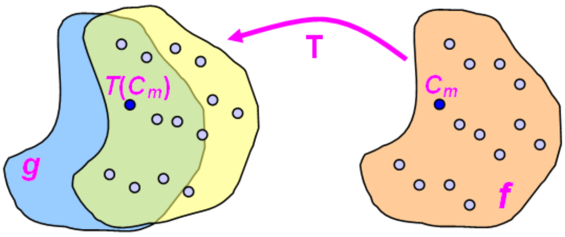


Fig. 2. Illustration of similarity calculation based on critical points.

### B. *3D Alignment Algorithm*

Let us consider two subunits $A$ and $B$, as shown in blue and magenta respectively in Fig. 3, and we want to align subunit $A$ to subunit $B$. In our previous paper [15] we discussed how to automatically detect the local symmetry axes, based on which we could segment each of the subunits. Given the segmented subunits $A$ and $B$ and the symmetry axes, we can transform $A$ to $B$ in four steps:

1) Translate $A$ by $t_0$.
2) Rotate $A$ by $r_0$.
3) Translate $A$ by $t$.
4) Rotate $A$ by $r$.

Since the symmetry axes of both $A$ and $B$ are given, the first two of above transforms, the translation $t_0$ and the rotation $r_0$, are uniquely determined by the symmetry axes. However, the translation $t$ and the rotation $r$ have to be decided based on the similarity scores between the density maps of $A$ and $B$ as discussed in Section III-A. Therefore, the transformation from one subunit to another has two degrees of freedom: one translation and one rotation, as illustrated in Fig. 3.

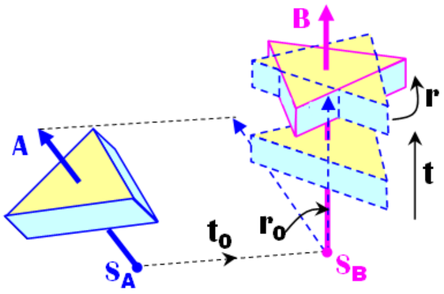
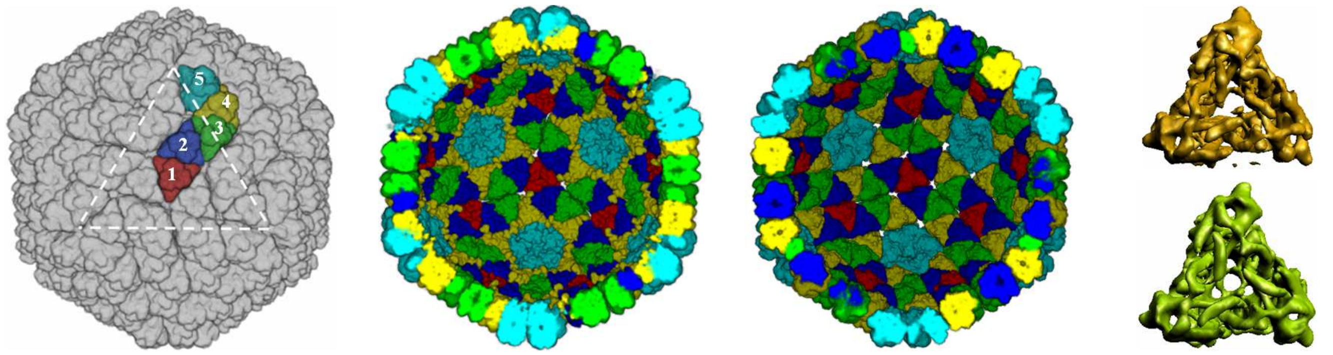


Fig. 3. The alignment between two subunits include translations $t_0$ and $t$, and rotations $r_0$ and $r$. But only $t$ and $r$ are unknown and need to be determined based on the similarity scoring function.

Putting the four transformation matrices together, we have the following matrix that transforms subunit $A$ to subunit $B$:

$$T(t,r) = M_4(r) \times M_3(t) \times M_2(r_0) \times M_1(t_0) \tag{4}$$

The matrices $M_1, M_2, M_3, M_4$ are conventional transformation matrices for translations or rotations. One can easily



(a) Segmented trimers (b) Old segmentation (c) New segmentation (d) Averaged trimer

Fig. 4. Illustrations of segmentation and averaging on rice dwarf virus (RDV). (a) The five independent trimers arranged on the virus capsid (viewed from outside). (b) The segmentation results without subunit alignment (viewed from inside). (c) The segmentation results with subunit alignment showing better consistence between subunits. (d) The averaged trimer (in green) looks less noisy and more symmetric than the original trimer (in golden).

derive the exact expressions for these matrices based on the given information (e.g., symmetry axes with starting/ending points). Substituting Equation 4 into Equation 3, we have the following maximization equation for the 3D alignment between subunits $A$ and $B$, with density functions $f$ and $g$ respectively:

$$\max_{\{t,r\}}\{S^3_{f,g}(t,r) = 1 - \frac{\sum_{m=1}^{M}|f(c_m)-g(T_{t,r}(c_m))|+\sum_{n=1}^{N}|f(T_{t,r}^{-1}(d_n))-g(d_n)|}{\sum_{m=1}^{M} max\{f(c_m),\ g(T_{t,r}(c_m))\}+\sum_{n=1}^{N} max\{f(T_{t,r}^{-1}(d_n)),\ g(d_n)\}}\}, \tag{5}$$

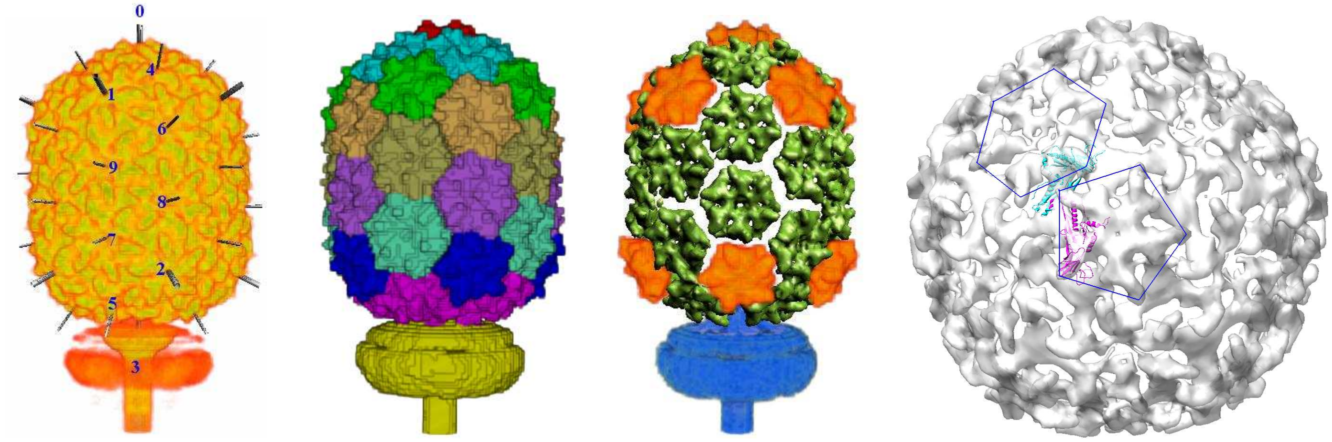
where $T_{t,r}$ is the transformation matrix from $A$ to $B$ as defined in Equation 4. The range of $t$ is user-defined. In our experiments we assumed that $t \in [-10,10]$ in pixel unit. The range of $r$ is from $0^0$ to $360^0$ for general objects. But most of maps we are currently dealing with are the virus structures which have some types of symmetries. Therefore, the segmented subunits usually have n-fold symmetry and the range of $r$ should be from $0^0$ to $\left(\frac{360}{n}\right)^0$.

There are a number of optimization techniques to find the $\{t,r\}$ that maximize the similarity score between $A$ and $B$ as defined in Equation 5. We adopted a simple two-level hierarchical method. On the coarse level, the translation variable $t$ is sampled by every one pixel unit from $-10$ to 10 and the rotation variable $r$ is sampled by every $5^0$ from $0^0$ to $\left(\frac{360}{n}\right)^0$ where $n$ is the fold number of the symmetry. For each combination of these $t$ and $r$, we calculate the similarity score and the one that maximizes the scoring function is the solution of Equation 5 on the coarse level, denoted by $\{t^{(1)},r^{(1)}\}$. On the fine level, the sampling is taken within a small range around $t^{(1)}$ and $r^{(1)}$. For the translation, the sampling is taken on the interval $[t^{(1)}-0.9,t^{(1)}+0.9]$ by every 0.1 pixel unit. The sampling for rotation is on the interval $[r^{(1)}-4^0,r^{(1)}+4^0]$ by every $1^0$. Again, the similarity score for each of these samples is calculated and the best one, denoted by $\{t^{(2)},r^{(2)}\}$, gives the final solution of Equation 5. By substituting $\{t^{(2)},r^{(2)}\}$ into Equation 4, we have the transform matrix $T(t^{(2)},r^{(2)})$, which gives the 3D alignment from subunit $A$ to $B$. A couple of applications will be discussed below.

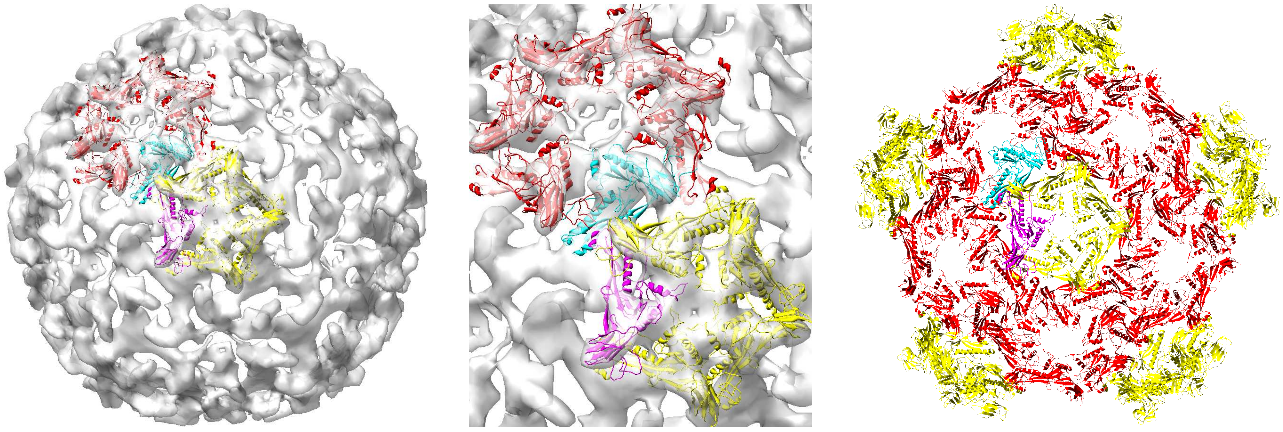
### C. Applications

We will consider two examples in this section. The first one is the rice dwarf virus (RDV), which has a perfect icosahedral symmetry at a resolution of $6.8\AA$ [8]. There are five independent subunits with 3-fold local symmetry, also known as trimers, as shown in Fig.4(a). The second example is the bacteriophage $\phi29$ at a resolution of $15\AA$ [33]. This map has a 5-fold global symmetry imposed along the vertical axis. As shown in Fig. 5(a), there are ten independent subunits including one tail on the bottom (subunit #3), three 5-fold subunits (#0, #1, #2), and six 6-fold subunits (#4 $\sim$ #9).

*1) Segmentation Improvement:* Our previous segmentation algorithm [15] is based on the multi-seeded fast marching method [34]. In this method each subunit is assigned with an initial contour which keeps growing according to a pre-defined speed function and eventually stops on the boundaries between subunits. When the symmetry is considered, some of the subunits are dependent (identical) and hence the contours corresponding to those subunits must grow simultaneously by the same amount in the same way. With this constraint, our previous segmentation [15] incorporated the global symmetry (for example, icosahedral for RDV and 5-fold for $\phi29$) and the local n-fold symmetry of each subunit into the segmentation. However, we did not consider the structural similarity between the independent subunits. In other words, the contours corresponding to independent subunits grew independently. This can sometimes yield noticeable errors especially when the boundaries between subunits are undistinguishable. An example is shown in Fig.4(b). This is the segmentation of the outer layer of the rice dwarf virus (RDV) and five types of trimers are colored differently (see Fig.4(a) for their relatively locations). We can see that some subunits (for example, the yellow ones) occupy more space than the others. To remedy this problem, we can utilize the 3D subunit alignment we discussed in Section III-B, such that the contours within all the subunits grow simultaneously by the same amounts in the properly-aligned directions. The results with this additional constraint are shown in Fig. 4(c). We can see that the subunits in the new results look much more uniform and consistent than those seen in Fig.4(b). However, it should be pointed out that



(a) Density map (b) Segmentation (c) Averaged map (d) Manually-fitted monomers



(e) Fitted hexon/penton (f) Closer view of (e) (g) Pseudo-atomic structure

Fig. 5. Illustrations of segmentation, averaging, and structural fitting on $\phi$29. (a) The reconstructed cryo-EM density map of $\phi$29. The symmetry axes are detected using our automatic method [15] and a total of ten independent subunits are labeled. (b) The subunits are segmented based on our previous segmentation method, combined with the 3D alignment approach discussed in Section III-B. (c) The segmented pentons and hexons can be averaged and the whole structure can be reconstructed from the averaged subunits based on the 3D alignment matrices. (d) Two PDB structures were manually fitted into one monomer of the chosen penton and one monomer of the chosen hexon. Shown here is the view from the top (head) of $\phi$29. (e) We can make copies of the manually-fitted PDB structures to the other monomers of the chosen penton and hexon. (f) A zoomed-in view of (e). (g) We can also make copies of the fitted PDB structure in (d) to the whole capsid and get a pseudo-atomic structure of $\phi$29. Shown here is only the top region of $\phi$29.

the subunits must be segmented before we can align them. Therefore, our previous segmentation algorithm [15] is still useful in providing us with the initial segmentation by which the subunits can be aligned. The segmentation of $\phi$29 is given in Fig. 5(b).

*2) Similarity Measurement:* Another application of our 3D alignment algorithm is the similarity measures between subunits. This is biologically very important because it tells us the structural similarity quantitatively between any two independent subunits. It is computationally straightforward to get the similarity score between two subunits. In fact, it is given by the maximum of the similarity scoring function $S^3_{f,g}(t,r)$ when Equation 5 is optimized. In Table I we give the similarity scores between the five independent trimers as shown and indexed in Fig. 4(a). The scores in bold on the diagonal indicate how symmetric each individual trimer is and they are calculated by Equation 3 where $T$ is the rotation along the related symmetry axis by an amount of $\frac{2\pi}{3}$ (in general, $\frac{2\pi}{n}$, where $n$ is the fold number). Table II shows the similarity scores between the four 5-fold subunits of $\phi$29, where the indexing numbers are given in Fig. 5(a). Since the tail subunit (#3) is obviously different from the other three 5-fold subunits, the similarity scores between the tail and the others are very low. However, the tail still shows a high 5-fold symmetry since it locates exactly on the global 5-fold symmetry axis. In Table III we show the similarity scores between the six 6-fold subunits. The subunit #5 seems less similar to the others because it is close to and very likely “disturbed” by the tail subunit.

TABLE I

SIMILARITY SCORES BETWEEN THE FIVE TRIMERS OF RDV

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ |
|---|---|---|---|---|---|
| $S_1$ | **0.955** | 0.911 | 0.848 | 0.900 | 0.894 |
| $S_2$ | | **0.926** | 0.854 | 0.889 | 0.880 |
| $S_3$ | | | **0.872** | 0.848 | 0.845 |
| $S_4$ | | | | **0.934** | 0.885 |
| $S_5$ | | | | | **0.856** |

*3) Average Map of Subunits:* If the segmented subunits of a large bio-molecular complex (such as viruses) have high sim-

TABLE II

SIMILARITY SCORES BETWEEN THE FOUR PENTONS OF $\phi$29

| | $S_0$ | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|---|
| $S_0$ | **0.991** | 0.948 | 0.949 | 0.344 |
| $S_1$ | | **0.960** | 0.958 | 0.306 |
| $S_2$ | | | **0.961** | 0.314 |
| $S_3$ | | | | **0.991** |

TABLE III

SIMILARITY SCORES BETWEEN THE SIX HEXONS OF $\phi$29

| | $S_4$ | $S_5$ | $S_6$ | $S_7$ | $S_8$ | $S_9$ |
|---|---|---|---|---|---|---|
| $S_4$ | **0.973** | 0.793 | 0.947 | 0.940 | 0.869 | 0.880 |
| $S_5$ | | **0.741** | 0.793 | 0.785 | 0.772 | 0.787 |
| $S_6$ | | | **0.971** | 0.963 | 0.884 | 0.884 |
| $S_7$ | | | | **0.965** | 0.888 | 0.884 |
| $S_8$ | | | | | **0.829** | 0.938 |
| $S_9$ | | | | | | **0.842** |

ilarities, they can be averaged such that the structural analysis of the whole complex can be simplified to the analysis of a single averaged subunit. Apparently averaging several subunits is straightforward if the alignments between any two of them are known. In general, the averaged map has a higher signal-to-noise ratio than each individual subunits, which makes it easier to analyze the structures (e.g., structural fitting or secondary structure identification, as will be discussed shortly later). Fig. 4(d) shows one segmented trimer of RDV (in golden) and the averaged map of the five independent trimers (in green). We can see that the averaged trimer has higher 3-fold symmetry and better signal-to-noise ratio. Fig. 5(c) shows the map constructed from the tail (blue), the averaged penton (orange) and the averaged hexon (green).

*4) Semi-Automatic Structural Fitting:* As we mentioned earlier, the structural analysis of large bio-molecular complexes includes two types: structural fitting for moderate resolution maps ($10\AA-20\AA$) and secondary structure analysis for intermediate (sub-nanometer) resolution maps ($5\AA-10\AA$). We shall give details on how to detect the secondary structures in Section IV. In this paragraph, we want to briefly show how to apply the 3D subunit alignment to building a pseudo-atomic model for a cryo-EM map at moderate resolutions. The idea is as follows. First, we choose a related PDB structure and fit it into the monomer ($\frac{1}{n}$ of the n-fold subunit). There are a lot of approaches which can do this computationally (see [32] for a review on various methods). In our test, we simply fit the PDB structure manually into the monomers of the chosen penton and hexon, as shown in Fig. 5(d) in magenta and cyan respectively. Once we have that, the second step is to transform the fitted PDB structure to all other subunits using the transformation matrices obtained from the 3D alignment. This can be done automatically. Fig. 5(e) shows the model for one penton and one hexon only and a closer view of the model is shown in Fig. 5(f). Fig. 5(g) shows the pseudo-atomic model in the top region of $\phi$29.

## IV. SECONDARY STRUCTURE IDENTIFICATION

Proteins typically contains two types of secondary structures: alpha-helices and beta-sheets. There have been some previous work on secondary structure identification. An approach for detecting alpha helices has been described in [35], where the alpha helix is modeled with a cylinder (length and thickness) with Gaussian distribution density function and the cylinder is cross-correlated with the segmented protein map. Since the best solution is achieved by exhaustively searching in translation space (three degrees of freedom) and orientation space (two degrees of freedom), this method is computationally expensive. Another approach, designed for beta sheet detection, was recently proposed in [36]. This method uses a disk (planar) model for beta sheets. It inherits the disadvantage of slow computational speed due to the exhaustive search in both translation and orientation spaces.

In the following, we present a skeleton-based approach for protein secondary structure identifications. Compared to the previous methods, our approach is extremely fast (hundreds of times faster) and yields high accuracy as well. Our skeletonization method is based on the local structure tensor and is segmentation-free.

### A. Local Structure Tensor

Local structure tensor has been used in image processing for solving a number of problems such as anisotropic filtering [37], [24] and motion detection [38]. Given a 3D map $f(x,y,z)$, the gradient tensor is defined as:

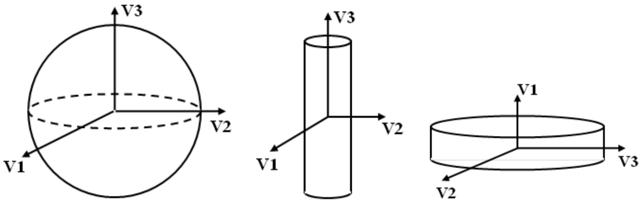
$$G=\begin{pmatrix} f_x^2 & f_xf_y & f_xf_z \\ f_xf_y & f_y^2 & f_yf_z \\ f_xf_z & f_yf_z & f_z^2 \end{pmatrix} \tag{6}$$

This matrix has only one non-zero eigenvalue: $f_x^2+f_y^2+f_z^2$. The corresponding eigenvector of this eigenvalue is exactly the gradient $(f_x, f_y, f_z)$. Therefore, this matrix alone does not give more information than the gradient vector. To make the gradient tensor useful, a spatial average (over the image domain) should be conducted for each of the entries of the gradient tensor, yielding what is called the *local structure tensor*. The averaging is usually based on a Gaussian filter:

$$T_\alpha=\begin{pmatrix} f_x^2*g_\alpha & f_xf_y*g_\alpha & f_xf_z*g_\alpha \\ f_xf_y*g_\alpha & f_y^2*g_\alpha & f_yf_z*g_\alpha \\ f_xf_z*g_\alpha & f_yf_z*g_\alpha & f_z^2*g_\alpha \end{pmatrix} \tag{7}$$

Here $g_\alpha$ is a Gaussian function with standard deviation $\alpha$. The eigenvalues and eigenvectors of the structure tensor $T_\alpha$ 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 [37]. 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 (see Fig.6):

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



(a) Spheres (b) Lines (c) Planes

Fig. 6. Three typical cases for a local structure tensor.

### B. Skeletonization

There have been a number of skeletonization approaches including boundary-based methods [39], [40] and boundary-free methods [41], [42]. While a pre-segmentation is required in boundary-based methods, boundary-free methods do not have this requirement but the skeletons are in general extracted by a skeleton-tracing step. In the following, we present a method to trace the skeletons based on local structure tensors.

*1) Selecting and Classifying Seeds:* As we discussed in Section IV-A, the local structure tensor can distinguish between spherical, linear, and planar structures. The spherical case is usually of no interest in skeletonization. Therefore, we will focus on the other two cases in the following. As we will see in Section IV-C, these two types of local structures exactly correspond to the alpha-helices and beta-sheets of a protein density map.

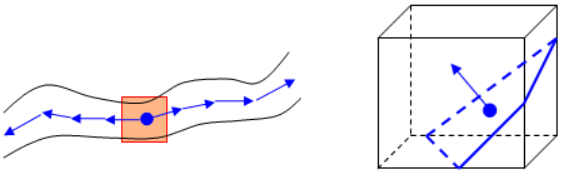
Seeds are the starting points for tracing the skeletons. There are multiple ways to define/detect the seed points, but one thing is common: all the seed points must be on the skeletons. As we will see shortly later, protein density maps look like mountains – high densities (features) correspond to the ridges. Therefore, the skeleton extraction is quite similar to the ridge tracing. For this reason, we choose the maximal critical points of the density maps as the seed points of our skeleton-tracers. The detection of these points was discussed in our previous work [15].

Since we are dealing with two types of skeletons: linear and planar, for which the skeleton-tracings are different, we need to classify each seed point into either linear or planar types before tracing the skeletons. In [37], the authors used three eigenvalues of the structure tensor to distinguish the lines from the planes. However, this criterion does not work well for protein secondary structures because some parts of proteins (e.g., coils, turns) look like helices except they are thinner. Therefore, a better way is to use thicknesses of the secondary structures along three principle axes. The thickness along any direction is defined by the width of the region above a pre-chosen threshold in that direction. Since we know the typical thickness of a helix, the threshold values can actually be determined automatically from the seed points that correspond to the helices based on the initial classification using the simple criterion in [37]. Once we know the thickness information for each seed point, we can classify the seeds into linear and planar according to the following criterion: Let $t_1$, $t_2$, $t_3$ be the thicknesses corresponding to the eigenvectors as shown in Fig.6. The following criterion is used to classify the seeds:

- lines: $\frac{t_1+t_2}{2} > helix_{min}$ and $\frac{t_1+t_2}{2} < helix_{max}$ and $min(\frac{t_1}{t_2}, \frac{t_2}{t_1}) > max(\frac{t_1}{t_3}, \frac{t_2}{t_3})$,
- planes: $t_1 > sheet_{min}$ and $t_1 < sheet_{max}$ and $min(\frac{t_2}{t_3}, \frac{t_3}{t_2}) > max(\frac{t_1}{t_2}, \frac{t_1}{t_3})$,

where $helix_{min}$ and $helix_{max}$ are the minimal and maximal thicknesses of helices, and $sheet_{min}$ and $sheet_{max}$ are the minimal and maximal thicknesses of sheets. All these parameters are provided by users.

*2) Tracing Skeletons:* The line-tracer is one-dimensional and hence is much easier than the plane-tracer. To trace a line structure, we start from the seed in two opposite directions, and follow the principle axis, defined by the eigenvector corresponding to the minimum eigenvalue of the local structure tensor (see Fig.7(a)). To trace a planar structure, we use the well-known marching cube method [43]. We start from the seed point and compute the plane that is perpendicular to the eigenvector corresponding to the maximal eigenvalue ($v_1$ in Fig.6(c)). The plane partitions all eight vertices of the cell containing the seed point into two classes: positive and negative. In addition, the plane intersects with some of the twelve edges of the cell. Both the classification of the vertices and the intersection information with the edges can be used in the marching cube algorithm, yielding a polygon (a list of triangles) representing the skeletons within the current cell (see Fig.7(b)). Then we move to the neighboring cells with which the detected skeleton (polygon) intersects. For the example in Fig.7(b), we need to check four neighboring cells (back, front, right, and bottom). For each of those new cells, the “checking” point (similar to the seeds) is calculated as the center of the existing intersecting points between the already-detected skeletons and the new cell. The new polygon within the new cell is extracted using the idea as explained above, based on the “checking” point and the new structure tensor around it. This process is repeated until certain stopping criterion is reached. The plane-tracer outputs a triangular mesh of skeletons. The stopping criteria for both line-tracer and plane-tracer are similar: the tracing process terminates whenever no new cells satisfy the criteria as discussed in Section IV-B.1.



(a) Line-tracer (b) Plane-tracer

Fig. 7. Skeleton-tracers. (a) Tracing 1D skeletons. (b) Tracing 2D skeletons (the marching cube method [43]).

*3) Merging Skeleton:* For each of the seed points, one can extract a curve or a surface, corresponding to a helix or sheet, respectively. However, quite often we see more than one seed points corresponding to the same secondary structure. One example is illustrated in Fig.8(a) and a closer view is shown in Fig.8(b). In order to have only one curve/surface for each helix/sheet, we need to merge the superfluous skeletons corresponding to the same helix/sheet. Besides eliminating the redundancy, the merging process can also give balanced skeletons because the new skeletons are the average of previously redundant ones. Fig.8(c) shows the skeleton after merging. It is worth pointing out that the topological ambiguity problem seen in the original marching cube method [43] can be resolved by an improved algorithm as discussed in [44].

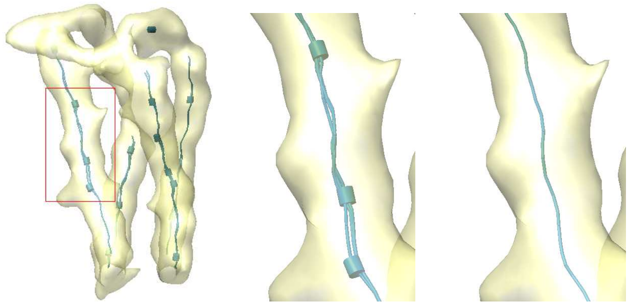

(a) Traced skeleton (b) Closer view (c) Merged skeleton

Fig. 8. Skeleton extraction. (a) The initial skeletons traced by the line-tracer. (b) A zoomed-in view of the rectangular region as shown in (a). The thick "dots" are seed points. (c) The skeletons after merging

### *C. Applications: Protein Secondary Structure Detection*

Once the skeletons are extracted, it is quite straightforward to locate the secondary structures in a protein density map. It is generally enough to represent beta-sheets using the extracted triangular mesh. However, if the resolution of the given maps is higher than $5\AA$, it is possible to see the individual beta-strands. In this case, we can apply the line-tracer to extract the strands and get a better model of the beta-sheets. As for alpha-helices, we can easily build a cylinder model for each helix based on the extracted skeletons.

We have tested our skeleton-based secondary structure identification approach on a large number of simulated and experimentally-reconstructed protein density maps. Due to the space limit, however, we will only show a few examples here. Fig. 9 illustrates the performance of our approach on the Gaussian blurred maps of two X-ray atomic structures. The first example is cytochrome c' (PDBID = 1bbh). The blurred map at $8\AA$ and detected skeletons are shown in Fig.8. From the skeletons, four alpha helices are detected as shown in Fig.9(a). To demonstrate the accuracy of our approach, the skeletons and detected helices are compared with the PDB structure in Fig.9(b) and (c). Another example is the blurred map of rat CD4 (PDBID = 1cid) as seen in Fig.9(d). The detected skeletons (sheets) are shown in Fig.9(e) and compared with PDB structures in Fig.9(f).

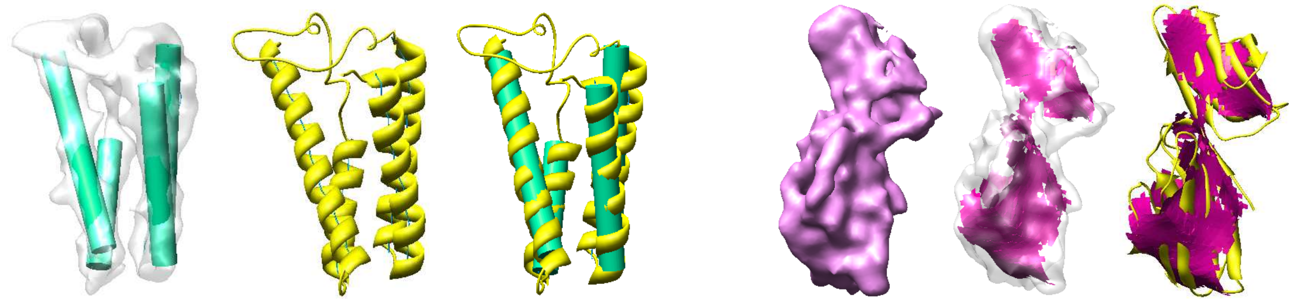
We also tested our approach on two simulated maps containing both alpha-helices and beta-sheets. Fig.10(a) shows the blurred map of the triose phosphate isomerase from chicken muscle (PDBID = 1tim) at $8\AA$. The detected alpha-helices and beta-sheets are shown in Fig.10(b) with density map and in Fig.10(c) with PDB structure. We can see that most helices/sheets agree very well with the PDB structure except that one small alpha-helix that is immediately adjacent to a long helix is missed (indicated by a red arrow). This result also agrees with that seen in [35]. Fig.10(d) and (e) give another view of (b) and (c), respectively. In Fig.11(a) we show a more complicated simulated map at $8\AA$, blurred from the PDB structure of the bluetongue virus VP7 (PDBID = 1bvp). This map contains a total of 27 alpha-helices in the lower domain and a few beta-sheets in the upper domain. The detected alpha-helices and beta-sheets are shown in Fig.11(b) with density map and in Fig.11(c) with PDB structure. All 27 alpha-helices and major portions of beta-sheets are correctly identified and agree very well with the PDB structure. Although two small alpha-helices (indicated by red arrows) are misidentified due to a couple of turns getting very close to each other, our results show better performance than the method proposed in [35]. Fig.11(d) shows another view of (c). A closer view of the skeletons together with the PDB structure is shown in Fig.11(e).

We also demonstrate our approach on a 3D cryo-EM map reconstructed from experimental samples of the rice dwarf virus (RDV) [8]. Fig.12(a) shows the reconstructed 3D map (left), the segmented double capsid layers (top-right), and segmentation of the outer layer (bottom-right). The segmented P3 protein (from inner layer) and P8 protein (from outer layer) are shown in Fig.12(b). The detected alpha-helices and beta-sheets for both proteins are shown in Fig.12(c).

In addition to the high accuracy, the speed of our approach is another advantage compared to the previous methods [35], [36]. The helix-hunter used in [35] may take up to one hour for the map such as P8 or P3 proteins of RDV, to detect only the alpha-helices. Our method takes only 3 seconds for P8 protein and 10 seconds for P3 proteins on a typical Linux workstation with single processor, to detect both alpha-helices and beta-sheets. We do not have the timings for sheet-miner as proposed in [36]. However, we believe that the sheet-miner of [36] should be as slow as, if not slower than, the helix-hunter of [35] because both methods used exhaustive searching schemes in the translational and orientational spaces.

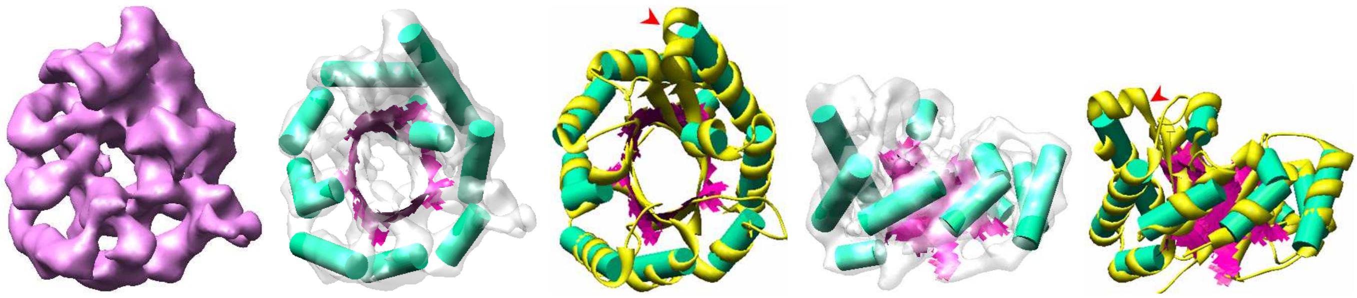
## V. CONCLUSIONS

In this paper we presented several computational approaches to automatically analyze structures of cryo-EM maps. In particular, we proposed fast methods to align (match) two segmented subunits in 3D space, and to identify secondary structures of a protein density map. The 3D alignment algorithm provides us with a set of transformation matrices which proved to be very useful in a number of applications, including similarity measurements, segmentation improvements, subunit averaging, and structural fitting between PDB structures and cryo-EM maps. Our skeleton-based method for secondary structure identifications is extremely fast (usually hundreds of times faster) compared to the existing methods [35], [36]. While the existing methods can only detect either helices or sheets, our approach can detect both types of secondary structures with high accuracies. When combined with our previous segmentation method [15], the approaches presented here can be employed to automatically interpret a wide range of bio-molecular structures especially those reconstructed by the single particle cryo-EM technique as introduced in Section II.



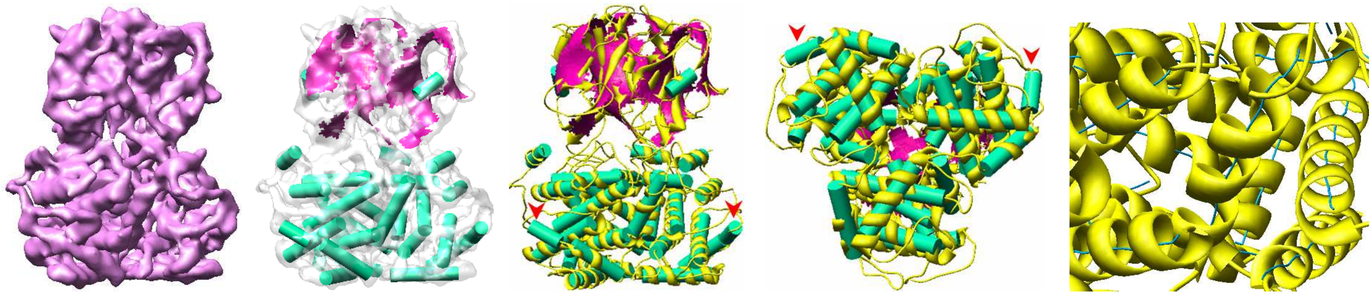
(a) Detected helices (b) Skeletons (c) Helices with PDB (d) Blurred map (e) Skeletons (f) Sheets

Fig. 9. Examples on two blurred maps of x-ray crystal structures. (a) The alpha-helices detected from the blurred map of cytochrome c' (PDBID = 1bbh). The blurred map at 8$\mathring{A}$ and extracted skeletons were shown in Fig.8. (b) The skeletons detected from the blurred map and compared with the PDB structure. (c) The detected helices are compared with the PDB structures. (d) The blurred map of rat CD4 (PDBID = 1cid) at 8$\mathring{A}$. (e) The skeletons (corresponding to beta-sheets) detected from the blurred map. (f) The sheets are compared with the PDB structures.



(a) Blurred map (b) Helix-sheet (map) (c) Helix-sheet (PDB) (d) Another view of (b) (e) Another view of (c)

Fig. 10. Secondary structure identification on the triose phosphate isomerase from chicken muscle (PDBID = 1tim). (a) The blurred maps at 8$\mathring{A}$ from the x-ray crystal structure. (b) The alpha-helices (green) and beta-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 alpha-helix that is immediately adjacent to a long helix is missed (indicated by red arrow). This result also agrees with that seen in [35]. (d) Another view of (b). (e) Another view of (c).



(a) Blurred map (b) Helix-sheet (map) (c) Helix-sheet (PDB) (d) Another view of (c) (e) Zoomed-in skeletons

Fig. 11. Secondary structure identification on the bluetongue virus VP7 (PDBID = 1bvp). (a) The blurred maps at 8$\mathring{A}$ from the x-ray crystal structure. (b) The alpha-helices (green) and beta-sheets (pink) detected using our method. (c) The detected helices/sheets are compared with the PDB structures. All alpha-helices and major portions of beta-sheets are correctly identified and agree very well with the PDB structure. Although two small alpha-helices (indicated by red arrows) are misidentified due to a couple of turns running into each other, our results show better performance than the method proposed in [35]. (d) Another view of (c). (e) A closer view of the detected skeletons together with the PDB structure. The chosen region is roughly in the rectangular area right below the center of (c).

## ACKNOWLEDGEMENTS

This work was supported in part by NSF grants INT-9987409, ACI-022003, EIA-0325550, and grants from the NIH 0P20 RR020647, R01 GM074258. We are grateful to Dr. Wah Chiu at Baylor College of Medicine, for helpful discussions related to this project and for providing us the reconstructed 3D map of the rice dwarf virus. We are also thankful to Dr. Tim Baker at The University of California at San Diego, for sharing the $\phi$29 cryo-EM density map and for providing the manually-fitted PDB structures as shown in Fig.5(d).

## REFERENCES

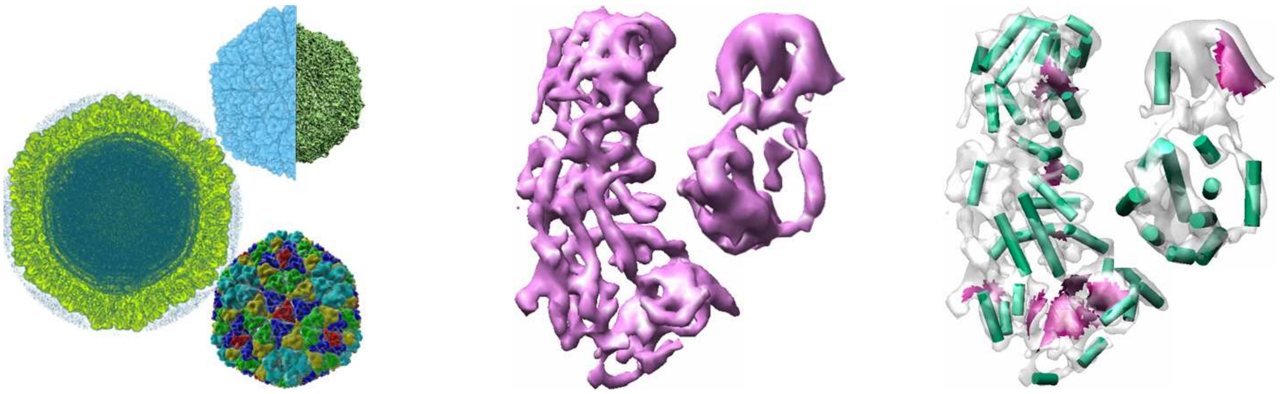
[1] M. Woolfson, *An Introduction to X-ray Crystallography*. Cambridge University Press, January 1997.

[2] “Online course on protein crystallography,” *http://www-structmed.cimr.cam.ac.uk/course.html*.

[3] A. E. Ferentz and G. Wagner, “NMR spectroscopy : a multifaceted approach to macromolecular structure,” *Quarterly Reviews of Biophysics*, vol. 33, pp. 29–65, 2000.

[4] P. Guntert, “Structure calculation of biological macromolecules from NMR data,” *Quarterly Reviews of Biophysics*, vol. 31, pp. 145–237, 1998.

[5] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, “The protein data bank,” *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.



(a) Reconstructed maps and segmentation (b) Segmented proteins (c) Detected helices and sheets

Fig. 12. The example of a real cryo-EM reconstructed map. (a) The reconstructed map of the rice dwarf virus (left), the segmented double capsid layers (top-right), and the segmentation of the outer layer (bottom-right). (b) The segmented P3 protein (left) and P8 protein (top-right). (c) The detected alpha-helices (green) and beta-sheets (pink) from the P3 and P8 proteins.

[6] T. S. Baker, N. H. Olson, and S. D. Fuller, "Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs," *Microbiology and Molecular Biology Reviews*, vol. 63, no. 4, pp. 862–922, 1999.

[7] M. van Heel, B. Gowen, R. Matadeen, E. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan, "Single-particle electron cryo-microscopy: towards atomic resolution," *Quarterly Reviews Biophysics*, vol. 33, no. 4, pp. 307–369, 2000.

[8] Z. H. Zhou, M. L. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Lu, and W. Chiu, "Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus," *Nature Structural Biology*, vol. 8, no. 10, pp. 868–873, 2001.

[9] W. Jiang, Z. Li, M. L. Baker, P. E. Prevelige, and W. Chiu, "Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolution," *Nature Structural Biology*, vol. 10, no. 2, pp. 131–135, 2003.

[10] Z. Zhou, M. Dougherty, J. Jakana, J. He, F. Rixon, and W. Chiu, "Seeing the herpesvirus capsid at 8.5 angstrom," *Science*, vol. 288, pp. 877–880, 2000.

[11] R. Matadeen, A. Patwardhan, B. G. E. Orlova, T. Pape, M. Cuff, F. Mueller, and R. B. M. van Heel, "The e. coli large ribosomal subunit at 7.5 angstrom resolution," *Structure*, vol. 7, pp. 1575–1583, 1999.

[12] S. Ludtke, P. Baldwin, and W. Chiu, "EMAN: semiautomated software for high-resolution single-particle reconstructions," *Journal of Structural Biology*, vol. 128, pp. 82–97.

[13] J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith, "SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields," *Journal of Structural Biology*, vol. 116, pp. 190–199.

[14] M. V. Heel, G. Harauz, E. Orlova, R. Schmidt, and M. Schatz, "A new generation of the IMAGIC image processing system."

[15] Z. Yu and C. Bajaj, "Automatic ultra-structure segmentation of reconstructed cryo-em maps of icosahedral viruses," *IEEE Transactions on Image Processing: Special Issue on Molecular and Cellular Bioimaging*, vol. 14, no. 9, pp. 1324–1337, 2005.

[16] W. Chiu, "What does electron cryomicroscopy provide that x-ray crystallography and nmr spectroscopy cannot?" *Ann. Rev. Biophys. Biomol. Struct.*, vol. 22.

[17] L. Amos, R.Henderson, and P.N.Unwin, "Three-dimensional structure determination by electron microscopy of two-dimensional crystals," *Prog. Biophys. Mol. Biol*, vol. 39, pp. 183–231, 1982.

[18] B. McEwen and M. Marko, "The emergence of electron tomography as an important tool for investigating cellular ultrastructure," *The Journal of Histochemistry & Cytochemistry*, vol. 49, no. 5, pp. 553–563, 2001.

[19] A. Koster, R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz, and W. Baumeister, "Perspectives of molecular and cellular electron tomography," *Journal of Structural Biology*, vol. 120, pp. 276–308, 1997.

[20] P. Penczek, M. Marko, K. Buttle, and J. Frank, "Double-tilt electron tomography," *Ultramicroscopy*, vol. 60, pp. 393–410, 1995.

[21] J. Frank, *Three-Dimensional Electron Microscope of Macromolecular Assemblies*. San Diego: Academic Press, 1996.

[22] S. Ludtke, D. Chen, J. Song, D. Chuang, and W. Chiu, "Seeing groel at 6-angstrom resolution by single particle electron cryomicroscopy," *Structure*, vol. 12, pp. 1129–1136, 2004.

[23] J.Fernandez, J.R.Sanjurjo, and J.Carazo, "A spectral estimation approach to contrast transfer function detection in electron microscopy," *Ultramicroscopy*, vol. 68, pp. 267–295, 1997.

[24] J. Weickert, *Anisotropic Diffusion In Image Processing*. ECMI Series, Teubner, Stuttgart, ISBN 3-519-02606-6, 1998.

[25] Z. Yu and C. Bajaj, "A fast and adaptive algorithm for image contrast enhancement," in *Proceedings of 2004 IEEE International Conference on Image Processing*, 2004, pp. 1001–1004.

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

[27] Z. Yu and C. Bajaj, "Detecting circular and rectangular particles based on geometric feature detection in electron micrographs," *Journal of Structural Biology*, vol. 145, no. 1-2, pp. 168–180, 2004.

[28] W. Nicholson and R. Glaeser, "Review: automatic particle detection in electron microscopy," *Journal of Structural Biology*, vol. 133, no. 2-3, pp. 90–101, 2001.

[29] C. Potter, Y. Zhu, and B. Carragher, "Automated particle selection for cryo-electron microscopy," *Journal of Structural Biology*, vol. 145, no. 1-2, pp. 168–180, 2004.

[30] F. Sigworth, "A maximum-likelihood approach to single-particle image refinement," *Journal of Structural Biology*, vol. 122, pp. 328–339, 1998.

[31] M. Sjors, M. Valle, R. Nuez, C. Sorzano, R. Marabini, G. Herman, and J. Carazo, "Maximun likelihood multi-reference refinement for eletron microscopy images," *Journal Molecular Biology*, vol. 348, pp. 139–149, 2005.

[32] W. Wriggers and P. Chacon, "Modeling tricks and fitting techniques for multiresolution structures," *Structure*, vol. 9, pp. 779–788, 2001.

[33] M. Morais, K. Choi, J. Koti, P. Chipman, D. Anderson, and M. Rossmann, "Conservation of the capsid structure in tailed dsdna phage the pseudoatomic structure of phi29," *Molecular Cell*, vol. 18, pp. 149–159, 2005.

[34] E. Sifakis and G. Tziritas, "Moving object localization using a multi-label fast marching algorithm," *Signal Processing: Image Communication*, vol. 16, no. 10, pp. 963–976, 2001.

[35] J. Wen, M. Baker, S. Ludtke, and W. Chiu, "Bridging the information gap: computational tools for intermediate resolution structure interpretation," *Journal of Molecular Biology*, vol. 308, pp. 1033–1044, 2001.

[36] Y. Kong and J. Ma, "A structural-informatics approach for mining b-sheets: locating sheets in intermediate-resolution density maps," *Journal of Molecular Biology*, vol. 332, pp. 399–413, 2003.

[37] J.-J. Fernandez and S. Li, "An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms," *Journal of Structural Biology*, vol. 144, pp. 152–161, 2003.

[38] G. Khne, J. Weickert, O. Schuster, and S. Richter, "A tensor-driven active contour model for moving object segmentation," in *Proceedings of IEEE International Conference on Image Processing*, 2001, pp. 73–76.

[39] L. Lam, S. Lee, and C. Suen, "Thinning methodologies - a compresensive survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14(9), pp. 869–885, 1992.

[40] R. Ogniewicz and O. Kubler, "Hierarchic voronoi skeletons," *Pattern Recognition*, vol. 28, no. 3, pp. 343–359, 1995.
[41] J. Jang and K. Hong, "A pseudo-distance map for the segmentation-free skeletonization of gray-scale images," in *Proc. Int'l Conf. Computer Vision*, 2001, pp. 18–23.
[42] Z. Yu and C. Bajaj, "A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion," in *Proc. Int'l Conf. Computer Vision and Pattern Recognition*, 2004, pp. 415–420.
[43] W. Lorensen and H. E. Cline, "Marching cubes: a high resolution 3D surface construction algorithm," *Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.
[44] A. Lopes and K. Brodlie, "Improving the robustness and accuracy of the marching cube algorithm for isosurfacing," *IEEE Trans. Visualization and Computer Graphics*, vol. 9, no. 1, pp. 16–29, 2003.

**Zeyun Yu** (S'03) received the B.S. degree in Mathematics from Peking University, Beijing, China, in 1996 and the M.S. degree in Pattern Recognition and Machine Intelligence from Chinese Academy of Sciences, Beijing, China, in 1999. He is currently a Ph.D. candidate in Computer Science at The University of Texas at Austin.

His research interests include image processing, pattern recognition, geometric/physical modelling for visualization, computational biology and bioinformatics. His Ph.D. dissertation focuses on automatic structural interpretations of reconstructed cryo-electron microscopy maps of large molecular complexes.

He is a student member of IEEE and ACM. He is also a member of IEEE Engineering in Medicine and Biology Society.

**Chandrajit Bajaj** (M'84) graduated from the Indian Institute of Technology, Delhi with a Bachelor's Degree in Electrical Engineering, in 1980 and received his M.S. and Ph.D. degrees in Computer Sciences from Cornell University, in 1983, and 1984 respectively.

He is currently the Computational Applied Mathematics Chair in Visualization Professor of computer sciences at the University of Texas at Austin, as well as the director of the Center for Computational Visualization, in the Institute for Computational and Engineering Sciences (ICES). Prior to the University of Texas, He was a professor of computer sciences at Purdue University and director of the Purdue center for image analysis and visualization. His research areas span Image Processing, Geometric Modeling, Computer Graphics, Visualization, and Computational Mathematics. His current research topics include denoising, reconstruction and compression algorithms for volumetric and time-dependent imaging; as well as data structures that support multi-resolution finite element approximations of large geometries and multiple function fields. He has been developing integrated approaches to computational modeling, simulations, mathematical analysis and interrogative visualization, especially for dynamic bio-medical phenomena.

He has over 190 publications, has written one book and edited three other books in his area of expertise. He is on the editorial boards for the International Journal of Computational Geometry and Applications, the ACM Transactions on Graphics, and ACM Computing Surveys. He is on numerous national and international conference committees and has served as a scientific consultant to national labs and industry.