Journal of Structural Biology 177 (2012) 410–419

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions

Mirabela Rusu [a,*], Willy Wriggers [b,1]

[a] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St., Houston, TX 77030, USA
[b] Department of Physiology and Biophysics and Institute for Computational Biomedicine, Weill Medical College of Cornell University, 1300 York Ave., New York, NY 10065, USA

ARTICLE INFO

Article history:
Received 26 May 2011
Received in revised form 22 November 2011
Accepted 28 November 2011
Available online 6 December 2011

Keywords:
Cryo-electron microscopy
Intermediate resolution
Extract alpha helices
Annotate alpha helices
Secondary structure elements

ABSTRACT

Cryo-electron microscopy (cryo-EM) enables the imaging of macromolecular complexes in near-native environments at resolutions that often permit the visualization of secondary structure elements. For example, alpha helices frequently show consistent patterns in volumetric maps, exhibiting rod-like structures of high density. Here, we introduce VolTrac (Volume Tracer) – a novel technique for the annotation of alpha-helical density in cryo-EM data sets. VolTrac combines a genetic algorithm and a bidirectional expansion with a tabu search strategy to trace helical regions. Our method takes advantage of the stochastic search by using a genetic algorithm to identify optimal placements for a short cylindrical template, avoiding exploration of already characterized tabu regions. These placements are then utilized as starting positions for the adaptive bidirectional expansion that characterizes the curvature and length of the helical region. The method reliably predicted helices with seven or more residues in experimental and simulated maps at intermediate (4–10 Å) resolution. The observed success rates, ranging from 70.6% to 100%, depended on the map resolution and validation parameters. For successful predictions, the helical axes were located within 2 Å from known helical axes of atomic structures.

1. Introduction

The continuing progress in the field of cryo-electron microscopy (cryo-EM) due to the improvement of the instrumentation, data acquisition, and image processing techniques (Baumeister and Steven, 2000; Frank, 2002) yields increasing numbers of biomolecular systems solved at intermediate to high resolution (Cong and Ludtke, 2010). Our focus here is on the most abundant intermediate-resolution (6–10 Å) reconstructions that exhibit the characteristic density signatures of secondary structure elements.

There are already a number of existing tools for the annotation of such secondary structure elements in cryo-EM maps. HelixHunter is a semi-automatic approach that combines a thresholding-segmentation scheme with an exhaustive search using a short helical template (Jiang et al., 2001). In SSEHunter, a modified template search approach yielded α-helix and β-sheet probabilities for a coarse-grained representation of the map (Baker et al., 2007), which was then manually annotated by secondary structure type. EMatch is a more automated approach that combines a template search with a segmentation-linkage schema (Lasker et al., 2005). All of these template search techniques involve the discrete exploration of the cryo-EM map using a short cylindrical template, which is subject to a relatively coarse angular and translational sampling. More recently, helical regions were also predicted based on density gradient information (Dal Palù et al., 2006); however, the utility of the method was not yet demonstrated on experimental reconstructions. In addition to these algorithmic search approaches, it is still common practice to use manual identification of helical map regions in the modeling work flow. For example, α-helices were flexibly fit into low-resolution cryo-EM maps of transmembrane proteins (Kovacs et al., 2007), and folding topology was modeled from sequence-based secondary structure predictions (Wu et al., 2005; Lindert et al., 2009).

Here, we introduce the VolTrac (Volume Tracer) approach that annotates elongated features corresponding to helical regions in cryo-EM maps. Although significant contributions have already been made by other authors with respect to helix detection, we feel that there is still an opportunity to explore alternative methods. One of our aims was to enable a fully automatic exhaustive search with a novel, quasi-continuous sampling of orientations and translations that visualizes helices on the fly as they are being detected. Inspired by earlier filtering approaches (Dal Palù et al., 2006; Chacón and Wriggers, 2002), we implemented a novel correction of map density variation for enhanced detection of helical densities. Another aim was to detect and follow the curvature of a helix.

* Corresponding author. Present address: Department of Biomedical Engineering, Rutgers, The State University of New Jersey, 599 Taylor Road, Piscataway, NJ 08854, USA.
E-mail address: mirabela.rusu@biomachina.org (M. Rusu).
[1] Permanent address: D. E. Shaw Research, 120 West 45th Street, New York, NY 10036, USA.

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.11.029

The VolTrac method combines a genetic algorithm (GA) for quasi-continuous sampling, a bidirectional expansion for following helical curvature and length, and a tabu search strategy for optimizing the exploration. Inspired by Darwinian evolution, GAs optimize a population of solutions allowed to evolve with operators such as mutation and crossover under the pressure of a scoring function (Holland, 1975; Goldberg, 1989). The evolutionary tabu search was introduced earlier for the simultaneous registration of multiple-component crystal structures with the cryo-EM map of their assembly (Rusu and Birmanns, 2010). VolTrac uses a small cylindrical template for which three translations and two rotations are optimized. When sampling the cryo-EM map, the population of cylindrical templates evolves for several generations while maximizing the cross-correlation coefficient. The best scoring template is typically placed within a helical region, aligned to the helical axis. Further processing using a local bidirectional expansion then follows the curvature and determines the length of the helical region. Once identified, the helices are placed into a tabu list to avoid redundant exploration.

Section 2 provides a detailed description of the implementation of the algorithm. In Section 3 we present an extensive validation of VolTrac on simulated and experimental maps with resolutions ranging from 4 to 10 Å. Finally, we describe computational performance, advantages, and limitations in Section 4.

2. Methods

The work flow of VolTrac shown in Fig. 1 A corresponds to the structure of this section. First, a novel local normalization filter for the cryo-EM map is introduced as a pre-processing step. Then, a detailed description of the genetic algorithm (GA), bidirectional expansion, and tabu search strategy is given. In the next step, we present the stop criteria and the post-processing of the helices. Finally, the validation procedure is described.

2.1. Local normalization of the cryo-EM map

A Gaussian-weighted local normalization is applied to the input map prior to launching VolTrac. Such normalization is beneficial because it enhances the appearance of the helices and it equalizes any uneven background density distributions in experimental cryo-EM maps (see Section 3). The filter is used only for helix detection, and no particular physical meaning is attributed to the resulting densities. For each voxel $\mathbf{r}$, the average $\overline{\rho_{em}(\mathbf{r})}$ and the standard deviation $\sigma(\rho_{em}(\mathbf{r}))$ of the densities are computed in the local neighborhood, using weights that follow a Gaussian distribution. The parameter σ_W characterizes the spatial extent of the Gaussian and is given in voxel units. For the maps presented here, σ_W equals 2.5 voxel units for the experimental maps and 1.5 voxel units for simulated maps. Simulated maps do not require leveling of the background, but we observed that locally normalized simulated maps enhance rod-like features, thus promoting the detection of α-helices. Consequently, the local normalization was applied to both experimental and simulated maps. In practical applications the voxel spacing may not always follow the map resolution, so as a rule of thumb we suggest that σ_W should be equivalent to about half the nominal resolution of a map.

The locally normalized densities are then computed according to the formula

$$\rho'_{em}(\mathbf{r}) = \frac{\rho_{em}(\mathbf{r}) - \overline{\rho_{em}(\mathbf{r})}}{\sigma(\rho_{em}(\mathbf{r}))}, \quad (1)$$

where $\rho_{em}(\mathbf{r})$ is the original density at voxel $\mathbf{r}$ and $\rho'_{em}(\mathbf{r})$ represents the locally normalized density. The subtraction of the local mean has the effect of centering the local intensities at zero while the division by sigma normalizes the 'contrast' such that positive densities of the helical rods are at comparable amplitude. Since helices exhibit high densities relative to their surroundings, they will exhibit positive density after filtering. Therefore, for the purpose of helix tracing, we discard the negative densities of the signal.

The local normalization will amplify any exterior noise, so experimental maps that contain outside noise should be thresholded and/or segmented at the molecular surface density level to set exterior densities to zero. Here, the experimental maps were thresholded to the "suggested contour level for viewing the map" given by the EMDB Database (Tagari et al., 2002). Such thresholding does not affect the extraction of α-helices that correspond to higher density regions. No such thresholding or segmentation was applied to the simulated maps, which were created without noise. To remove noise from the filtered maps, a Gaussian blurring using a sigma of 1.5 voxel units was applied in this work after local normalization. Alternatively, a user may apply a more advanced denoising scheme such as the Digital Paths Supervised Variance filter developed by us (Starosolski et al., submitted).

2.2. Genetic algorithm

Inspired by biological evolution, GAs use genetic operators such as mutation and crossover to optimize a fitness function in an iterative optimization (Goldberg, 1989). GAs consider a population

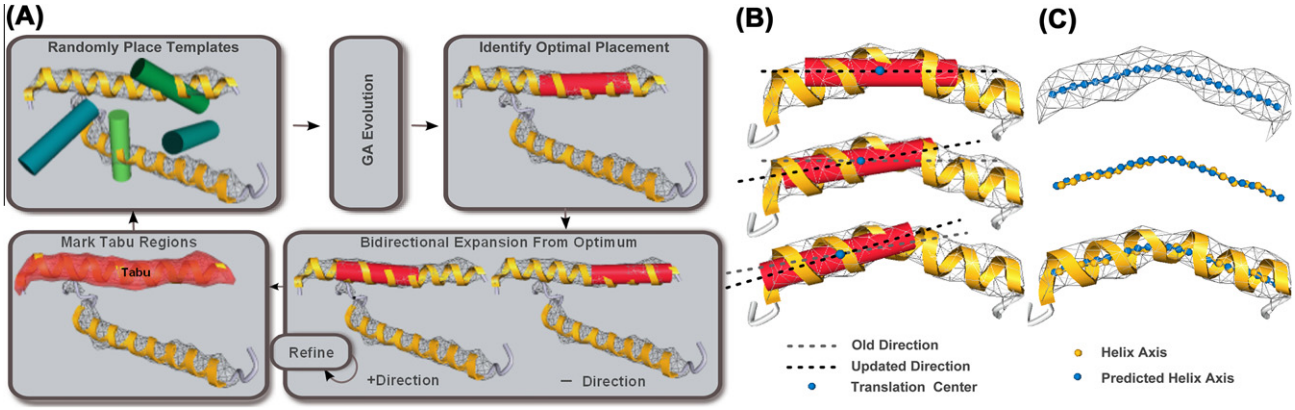

Fig. 1. (A) VolTrac work flow: A random initial population of cylindrical templates is allowed to evolve for several generations. The best scoring template is then used for the bidirectional expansion. The annotated region is included in the tabu list. A new GA run is executed, starting from new random distributions. (B) During the bidirectional expansion, the axis of the region is updated, allowing the template to follow the curvature of the helix. (C) Top: The predicted helix is described by the translation centers obtained in the bidirectional expansion. Middle: Comparison between the axes of the predicted helix and known helix. Bottom: The axis of the known helix is obtained by averaging four consecutive alpha carbons of the atomic structure. All molecular graphics in this paper were generated with Sculptor (Birmanns et al., 2011).

of candidate solutions and allow it to evolve over several generations according to an elitist scheme based on the principle of survival of the fittest. One reason for using the GA optimization for VolTrac was that it allowed the approach to be integrated into our interactive molecular graphics software Sculptor (Birmanns et al., 2011) to visualize helices in real time as they were identified. Another important reason for the GA was the possibility of supporting a quasi-continuous representation of translations and rotations.

In VolTrac, each individual in the population represents a cylindrical template (radius = 2 Å, length = 20 Å) with the three translational and two rotational degrees of freedom as the free optimization parameters (the irrelevant rotation about the cylinder's main axis is ignored). These five degrees of freedom are encoded by four parameters [x,y,z; r_i], where x, y, and z represent the three dimensional translations and r_i is an index of the list of angles that uniformly sample the rotational degrees of freedom using an angular step size of 1°. The angular sampling thus approaches a continuum, in contrast to earlier template convolution techniques that reported orientational steps of up to 15°. The fitness of each individual in the population is then estimated based on a cross-correlation coefficient that samples the cryo-EM map within the template cylinder mask. This coefficient is calculated according to Eq. (1) in Wriggers (2012), where ρ_{calc} corresponds to the cylinder mask projected to the map lattice. As mentioned above, we correlate this template with a locally normalized map, an approach that is similar to normalizing the correlation locally (Roseman, 2000).

Two genetic operators are considered during the GA evolution (see Rusu and Birmanns, 2010, for implementation details). The mutation modifies the transformation of the templates, allowing them to sample the cryo-EM map at different placements or with various orientations. Large mutations enable the template to explore the map, while small mutations have the effect of a more localized refinement. The mutation operator modifies randomly picked individuals by applying variations that follow a Cauchy distribution:

$$C(\beta, x) = \beta/(\pi \cdot (\beta^2 + x^2)) \tag{2}$$

where $\beta = 0.05$ corresponds to the standard deviation. Compared to a Gaussian, the Cauchy distribution is also biased to small variations, but it creates larger deviations with higher probability, thereby promoting a better exploration of the search space. For example, when the mutation operator is applied, a new template is generated according to the parameters of an individual template randomly selected from the population. The new template only differs from the original by one randomly chosen parameter following a Cauchy distribution. The new template will be slightly different but typically close to the original, thereby allowing a refinement of the placement.

The crossover operator enables the exchange of information between GA template individuals. New transformations are identified by swapping the translations and rotations of selected templates. We used a combination of crossover schemes where [x,y,z; r_i] were either swapped at one or multiple points, or modified using arithmetic operations (Rusu and Birmanns, 2010). The crossover operator not only affords efficient exploration of the cryo-EM map, but it is also particularly beneficial in the case of bundles of parallel helices where the orientation is conserved.

Initially, the cylindrical template population randomly samples the cryo-EM map outside of any tabu regions (Fig. 1A, top left). This population is then allowed to evolve under selective pressure for several generations until an optimal solution is found (a solution is considered optimal when no further improvements are achieved over several generations).

2.3. Bidirectional expansion

The template with the optimal placement usually covers part of a helical region, aligned with its main axis (e.g., Fig. 1A, top right), but it does not capture the full length of the helix or its curvature. Starting from this optimal placement and using the template's main axis as an indicative direction, a bidirectional expansion is performed to determine curvature and length of the helix.

The bidirectional expansion is performed in two steps, using an 8 Å-long cylinder with a radius of 1 Å. First, a local refinement of the translations and rotations is performed at the current placement of the template. In the second step, the template is translated in one and then the other direction along the axis of the optimal solution (Fig. 1A, bottom right). These two steps are iterated along the axis of the helical region until the score at the current position falls below a certain percentage of the initial score. By default, this limiting score threshold is set to 70% of the initial (highest) score computed within the current region. The iterative annotation is based on the assumption that the short template should maintain a rather constant score when moved inside a helical region. Therefore, as the template reaches the end of the region, the score decreases considerably and the expansion is stopped.

We note that the score threshold acts as an adjustable tolerance for deviations from the ideal rod shape due to experimental noise or reconstruction artifacts. In this work, we used a relatively stringent 80% threshold for idealized (simulated) maps and more permissive thresholds of 55–70% for experimental maps. The threshold levels were determined based on the observed performance of the algorithm (see Sections 3 and 4).

Atomic structures of proteins exhibit both straight and bent helices. Therefore, VolTrac considers the general case in which helices may be curved. At each translation of the template, the orientation is subject to refinement, following the curvature of the region (see Fig. 1B). The translation center is stored as an axis point of the predicted helix (Fig. 1C top). This predicted axis closely follows the known axis of the atomic structure (Fig. 1C bottom), as is evident from a side-by-side comparison (Fig. 1C center). The parameterization of the cylinder length and radius was chosen so that a linear point density of ~1.5 Å was achieved in order to approximately match the point spacing of the known axis. The predicted axes are exported in PDB format to be visualized in an external program. We recommend our molecular graphics program Sculptor (Birmanns et al., 2011), where exported helices can be rendered directly using tube or ribbon representations.

2.4. Tabu search

Once a helical region is characterized by the bidirectional expansion, it is appended to the tabu list and eliminated from further exploration. A tabu region is defined about each translation center of the template, i.e., the axis of the predicted helix. The radius of each exclusion sphere is set to 6 Å to generate overlapping spheres for adjacent axis points, marking the entire length and width of the helix. During the evolution, the templates are not allowed to be placed within such tabu regions, i.e., their centers may not be closer than 6 Å to the points in the tabu list. This strategy prevents the algorithm from revisiting occupied regions, thereby promoting more efficient exploration of other helical regions in the map.

The previously described GA and bidirectional expansion steps identify one helical region at a time and therefore need to be iterated several times until all helical regions are identified. Each such iteration starts with a new random population of short cylindrical templates, while the tabu regions are preserved between iterations.

2.5. Stop criteria

The algorithm is iterated until a stop criterion is met. Without loss of generality, it can be assumed that the number of helices to be identified, denoted by N, is known either from prior structural data or sequence-based secondary structure prediction algorithms. For a given N, the algorithm will stop exploring the cryo-EM map when it has identified $3 \cdot N$ helices. More than N helices are investigated to allow for some imperfect ranking of the results, thereby yielding a better exploration of the search space. If the number of helices is not known a prior, the algorithm is stopped once the map has been extensively explored as assessed by a coverage rate, when it is impossible to place more templates into the map due to the tabu regions, or when only short helices are annotated for multiple consecutive iterations. In this case, more than $3 \cdot N$ helices may be identified.

2.6. Ranking of predicted helices

The outcome of the algorithm is a list of helices described by the translation centers of the template during the bidirectional expansion (Fig. 1C). To facilitate the exploration of the results, the list is sorted in a post-processing step using a correlation-weighted length

$$L_{Helix} = \frac{[\overline{CC_{Expansion}}]^2 \cdot [CC_{Interior}]^2}{[CC_{Exterior}]^2} \cdot Len, \quad (3)$$

where *Len* represents the length of the helix (computed as the sum of distances between adjacent points on the axis of the predicted helix). The squared correlations in the fraction emphasize the scoring function relative to the length of the helix: $\overline{CC_{Expansion}}$ is the mean normalized cross-correlation measured by the short cylindrical template during the bidirectional expansion. $CC_{Interior}$ is the normalized cross-correlation of the predicted axis. $CC_{Exterior}$ is the normalized cross-correlation of a 5 Å radius cylinder following the predicted axis. The heuristic L_{Helix} is high for long helices that have high density around the axis and low density at the exterior.

2.7. Validation

We designed a range of tests on simulated and experimental maps to assess the sensitivity and accuracy of the predictions afforded by VolTrac. Experimental maps were selected from cases where closely matching atomic structures were available from X-ray crystallography. We created simulated maps by low-pass filtering of atomic structures using Sculptor (Birmanns et al., 2011). Resolution values presented here correspond to the Situs convention and they are smaller by a factor of 1.282 compared to EMAN (Ludtke et al., 1999); see Section 4 of Wriggers (2012). Anecdotal evidence suggests that the Situs resolution values are close to reported resolution values of experimental maps, whereas EMAN convention, used in many earlier helix detection publications, is close to the crystallographic resolution which tends to give somewhat larger direct space values.

The *N* top-ranked solutions (where *N* represents the known number of helices) were selected for validation, and their performances were quantified using point-based and helix-based measures. Point-based measures compare the points on the predicted axes with points on the known axes (resulting from averaging coordinates of four consecutive alpha carbons in the crystal structure):

- The point sensitivity (pSe) is defined as the percentage of known points that were correctly predicted by VolTrac.
- The point positive predictive value (pPPV) is defined as the percentage of predicted points that correspond to known axis points.
- The root mean square deviation (RMSD) of corresponding predicted and known axis points quantifies the separation of the axes.

For these point measures (pSe, pPPV, and RMSD), a predicted point is considered to match a known point if they are found within 4 Å of each other. This tolerance was set in order to accommodate experimental maps that show minor conformational differences from the crystal structure (Wriggers et al., 2000).

Helix-based measures directly compare the geometric properties of the predicted and known helices:

- The Δ Turns value is defined as the number of mismatched helical turns between two helices.
- The helix sensitivity (hSe) is the percentage of known helices detected among the top N predicted helices. For this measure, a known helix is considered to a true positive (TP), i.e., detected by VolTrac, if a predicted helix at least partially overlaps and aligns with its axis. For example, a reported hSe of 70.6%, where $N = 17$, implies that 12 out of the 17 helices of the crystal structure were found among the top 17 predicted solutions. Typically, other known helices are identified as well, but they are ranked lower in the list of solutions and not considered for the hSe value.

3. Results

This section is organized as follows: first, the local normalization of densities is demonstrated using an experimental map of the chaperonin GroEL (Ludtke et al., 2004); then, a typical helix extraction outcome is shown for an idealized (simulated) 8 Å resolution GroEL map (Braig et al., 1995). To assess the performance of VolTrac more systematically, we performed a series of tests using simulated maps and six experimental maps at variable resolutions.

3.1. Local normalization

Experimental cryo-EM volumes may suffer from uneven density distributions due to conformational disorder and alignment artifacts, with higher density exhibited at the core and a lower density at the surface. For example, a GroEL map solved at 6 Å-resolution (Ludtke et al., 2004) shows high densities in the equatorial domain, whereas the apical domain appears weaker (Fig. 2A, B). Therefore, to normalize the features across the map, the Gaussian-weighted local normalization was applied (see Section 2). The resulting filtered volume (Fig. 2C, D) shows a more uniform distribution of density, bringing the equatorial and apical domains to a comparable level (indicated by arrows in Fig. 2B, D). This balanced distribution of filtered densities was conserved at different isosurface values, as observed by visual inspection in a molecular graphics program (data not shown).

3.2. Application example

To demonstrate a typical VolTrac application, a GroEL monomer (from PDB ID: 1OEL, (Braig et al., 1995) was low-pass filtered to 8 Å resolution with a voxel size of 2 Å (Fig. 3A) the helix extraction was executed on a locally normalized map using an expansion threshold of 80% (see Methods). VolTrac identified all 17 known helices of seven or more residues, placing 16 in the top 17 scoring solutions (Fig. 3B and C) and the remaining one at rank 18. The total run time for this example was 10.4 min (using an Intel I7–2600 processor at 3.4 GHz).

The ranking of the results using the empirical L_{Helix} value (Eq. (3)) performed well in this example. In the case of GroEL, the N top results were divided into 16 true positives (the actual helices)

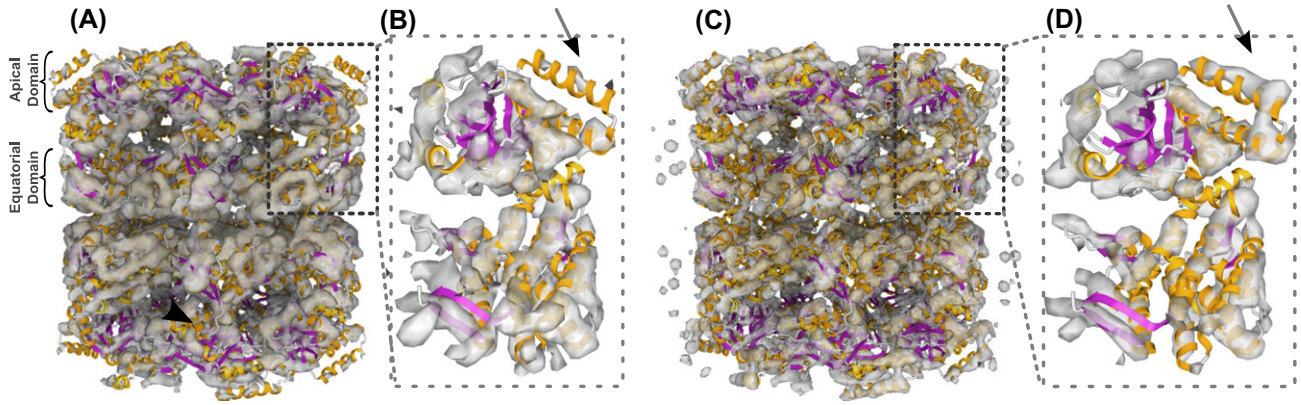

Fig. 2. Gaussian-weighted local normalization applied to a 6 Å resolution experimental map of the chaperonin GroEL (EMDB ID: 1081, Ludtke et al., 2004). (A and B) The map shows higher density values in the equatorial than in the apical domain (isolevel 0.59744). (C and D) After Gaussian-weighted local normalization, the map depicts comparable density value across the map. Arrows indicate area of interest. The crystal structure of GroEL is shown as a reference in ribbon representation, with α-helices depicted in yellow.

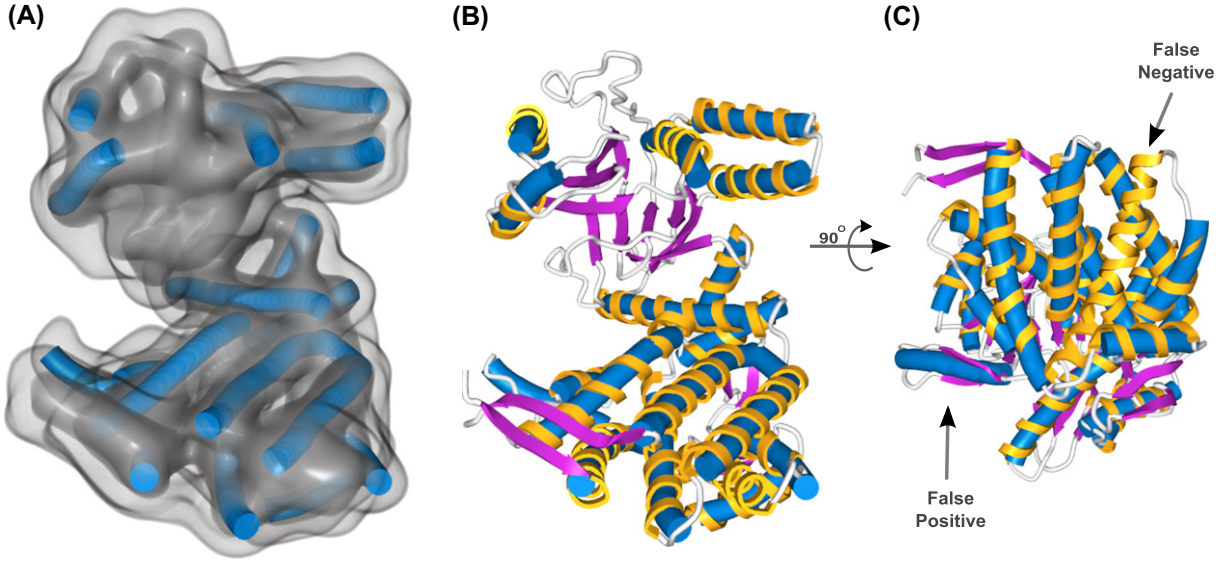

Fig. 3. (A) A simulated map obtained by low-pass filtering a GroEL monomer to 8 Å resolution is presented along with the helices predicted by VolTrac (represented as blue tubes). (B) Side and (C) bottom views of VolTrac results (blue cylinders) overlapping the target crystal structure (α-helices represented as yellow ribbons).

and one false positive. False positives may be found in rod-like regions that do not correspond to alpha helices, for example (anti) parallel β-sheets (Fig. 3C). On the other hand, false negatives represent correct helices that are found lower in the ranking (such as rank 18 here), below the N top predictions. Our tests have shown that these false negative helices either are of smaller size (~7–10 residues) or may lack the characteristic rod-like shape.

To measure the performance of VolTrac, we compute the hSe value, i.e., the percentage of true positive helices predicted by VolTrac, to hSe = 94.1%. The agreement between the predicted and known axis points was characterized by the measures pSe = 90.5%, pPPV = 86.8%, and RMSD = 0.68 Å. As a control, we repeated the calculation in the absence of the local normalization, which was expected to perform less favorably (see Section 2). The resulting performance measures without local normalization were hSe = 76.4%, pSe = 89.6%, pPPV = 65.6%, and RMSD = 1.06 Å.

3.2.1. Performance as a function of resolution

For a systematic benchmark of VolTrac we generated simulated maps from monomer GroEL structures (17 long helices of seven or more residues; 57 kDa total molecular weight; Braig et al., 1995), succinate dehydrogenase (33 long helices; 118 kDa; Yankovskaya et al., 2003), and photosynthetic reaction center (34 long helices; 132 kDa; Baxter et al., 2004). The α-helical secondary structure content ranged from 37% to 51% for these structures. Each structure was low-pass filtered to resolutions ranging from 6 Å to 10 Å (2 Å voxel size). A local normalization and an expansion threshold of 80% were applied (see Methods). Fig. 4A–C shows the helix and point-based performance measures hSe, pSe, pPPV, and RMSD as a function of the resolution. The numeric values were obtained as averages over three statistically independent runs.

Overall, VolTrac detected α-helices reliably up to ~7 Å resolution, beyond which the performance started to degrade toe to the loss of secondary structure detail. In all three cases the hSe was typically above 70% (Fig. 4A) and the geometric accuracy of the prediction was better than 1.4 Å (Fig. 4B). Moreover, the pSe estimated on the axis points was better than 80%, indicating that any missed helices were short, as they accounted for only a small number of points (Fig. 4C). The pPPVs were systematically lower than the corresponding pSe values, indicating a VolTrac bias to

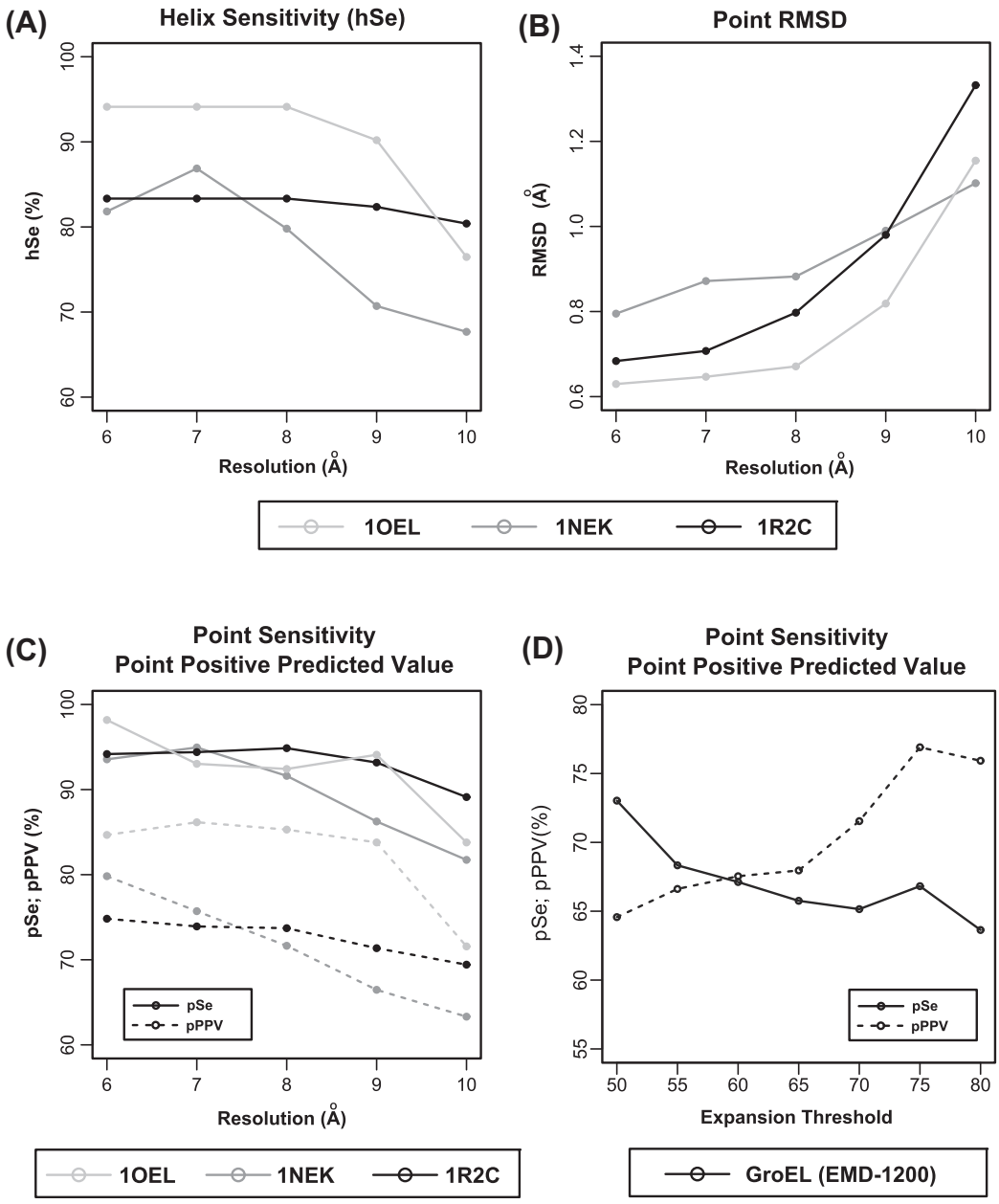

Fig. 4. (A–C) VolTrac performance validation as a function of resolution for the simulated maps of GroEL (PDB ID: 1OEL, 17 helices), succinate dehydrogenase (PDB ID: 1NEK, 33 helices), and photosynthetic reaction center (PDB ID: 1R2C, 34 helices) (A) Helix sensitivity hSe (see text). (B) RMSD of the helical axes. (C) The pSe and pPPV measures (see text) based on the axis points. (D) Variation of pSe and pPPV as a function of the expansion threshold, plotted for the experimental map of GroEL solved at 7.8 Å resolution (EMDB ID: 1200). Fig. 5E, F and the Table 2 show the results for an expansion threshold = 60%; The plotted values were obtained as averages over three statistically independent runs.

predict slightly longer helices than justified by the known structure. A more detailed inspection revealed that some axis points were predicted at the ends of known helices, where occasionally the backbone showed helix-like organization. We assumed that such minor false positive predictions would be preferable to false negatives. If desired, a user could reduce the resulting helix length by increasing the default score threshold in the bidirectional expansion (see Methods).

The helix sensitivity, hSe, of GroEL surpasses that of the other two cases. This result may have been influenced by the relatively small number of helices of GroEL, 17 compared to 33/34 in the other test cases. Short helices ranked lower in the list drop in and out of the top N = 17 solutions, resulting in possibly larger fluctuations for GroEL hSe. In the two other test cases, the sensitivities, hSe and pSe, are highest for relatively lower resolutions 7–8 Å (Fig. 4A and C). At this resolution, the rod-like shape of helices is devoid of any helical structure (or side chains) and are therefore easily detected using the cylindrical template.

3.3. Experimental validation

The above tests were based on idealized maps in the absence of noise. To assess the performance of the algorithm on realistic cryo-EM maps in the presence of noise and 3D reconstruction artifacts, we chose six benchmark maps for which atomic structures were available as a control. Specifically, we used maps of GroEL at resolutions of 4.2 Å (Ludtke et al., 2008), 6.0 Å (Ludtke et al., 2004), and 7.8 Å (Stagg et al., 2006), a map of the 20S proteasome at 6.8 Å

resolution (Rabl et al., 2008), a map of rice dwarf virus at 7.9 Å resolution (Liu et al., 2007), and a map of kinesin at 9 Å resolution (Sindelar and Downing, 2007) (Table 1). Fig. 5 presents an overview of the six benchmark systems.

Local normalization and expansion thresholds ranging from 55% to 70% were applied (see Section 2 and Table 2). In some cases, the expansion thresholds were adjusted from the 70% default value to optimize the hSe measure, depending on the quality of a particular map. Similar to the simulated test cases, different performance measures were assessed, which are presented in Table 2. As in the simulated test cases, VolTrac predicted helices that were slightly (~1–2 turns) longer compared to those in the crystal structure.

Historically, GroEL was solved at increasing resolution by cryo-EM, whereas initial reconstructions at 11–13 Å showed only the overall shape of the chaperonin (Ranson et al., 2001; Ludtke et al., 2001), secondary structure elements were detected in intermediate resolution maps (Ludtke et al., 2004; Stagg et al., 2006), and a recent map at 4.2 Å resolution even afforded a trace of the protein backbone (Ludtke et al., 2008). VolTrac was executed here for a single monomer (extraction of the monomer was performed by masking the map using a docked atomic structure). Fig. 5A–F shows the outcome of VolTrac for the three investigated GroEL maps, depicting in blue the helices detected in the top N = 17 solutions and in green the helices found lower in the list. Overall, the observed hSe values varied between 70.6% and 82.4% (Table 2). Similar values were also identified for pSe (66.4–86.3%). The axis RMSD values were below 2.06 Å.

The GroEL tests show that the accuracy of VolTrac depends on the quality of a particular map. As expected, the performance was best for the 4.2 Å map (as judged by TP, hSe, pPPV, RMSD, and ΔTurns measures). The 6.0 Å map, despite its relatively high resolution, required a particularly low expansion threshold of 55% to accommodate visible variations in the rod densities.

The other three systems in the experimental benchmark measured helix-based hSe values ranging from 78.2% to 100%, and axis point-based pSe values ranging from 74.3% to 100.0% (Table 2). VolTrac performance was essentially perfect for the 20S proteasome at 6.8 Å resolution (Fig. 5G,H), where the algorithm predicted all 11 helices with an RMSD of 1.10 Å and ΔTurns of 0.4. In comparison, the slightly lower resolution rice dwarf virus and kinesin cases exhibited RMSD values of 1.12 and 1.33 Å, respectively, and ΔTurns values of 1.4 and 1.17, respectively, when compared to the crystal structure (Fig. 5I-L).

To investigate the effect of the expansion threshold, we tested VolTrac on the experimental map of GroEL solved at 7.8 Å resolution. Fig. 4D shows the computed point sensitivity (pSe) and point positive predictive value (pPPV) for expansion threshold values ranging from 50% to 80%. At low threshold values, the pSe is high, indicating that a large number of points of the real helix axes are properly predicted by VolTrac. The pPPV is rather low as the algorithm tends to over-estimate the length of the helices, due the template expanding beyond the length of the actual helix. On the other hand, when the expansion threshold is high, the pSe decreases, yet the pPPV is considerably enhanced indicating that most of the predicted axis points correspond to real helix axis points. In choosing the values of the expansion threshold, one wishes to maintain a balance between the pPPV and pSe that allows reasonable values for both indicators, i.e., predict as many of the points of the axis of the real helix with reduced false positives. Fig. 5E, F and Table 2 show the results obtained for an expansion threshold equal to 60%.

The execution time of VolTrac is largely independent of the size of the map. The time mainly depends on the number of helices to identify. For example, the map of GroEL at 7.8 Å resolution has the same dimension as kinesin (Table 1), yet the times vary with 12 and 20 min, respectively (Table 2). Overall the execution times ranged from 9.3 to 39 min (on an Inter I7–2300 Processor at 3.4 GHz).

4. Discussion

VolTrac combines a genetic algorithm, a bidirectional expansion, and a tabu search strategy to trace alpha-helical densities in cryo-EM maps. The genetic algorithm performs a global search to identify fragments of helical regions, while the bidirectional expansion determines the curvature and length of the entire region. As the algorithm annotates the helices, they are placed in the tabu list to prevent them from being revisited later in the search.

VolTrac is fully automatic and only required a rough estimate of the number of helices, and a selection of the score threshold in the bidirectional expansion (see below). Similar to earlier approaches by other groups, VolTrac was designed under the assumption that α-helices can be identified as rod-like densities in cryo-EM maps. However, VolTrac also considers the possibility of curvature based on the observed shape of α-helices in high-resolution structures. The bending of a helix is characterized in the bidirectional expansion by using a short cylindrical template that traces the rod-like feature. Such a strategy has the advantage of rendering the helical axes as smooth continuous curves that follow closely the curvature of the true helical axes, as illustrated in Fig. 1C (center) and by the low RMSD values (Fig. 4 center and Table 2). As opposed to approaches such as Helix/SSEhunter (Jiang et al., 2001; Baker et al., 2007) or EMatch (Lasker et al., 2005), VolTrac predicts a single contiguous helix instead of several piecewise straight cylinders. Although the building of complete atomic models was not an aim of this paper, such curved predictions might facilitate the atomic level tracking of the molecular structure. Similar to our approach, Dal Palù and colleagues have implemented curvature (Dal Palù et al., 2006) in a gradient-based technique, yet no validation was provided on experimental maps where the noise may considerably affect the outcome.

VolTrac reliably detected α-helices in simulated maps of 6–10 Å resolution and in experimental maps of 4–9 Å resolution, with a true positive accuracy ranging from 70.6% to 100%, as estimated in experimental settings. Although low resolution maps did fare

Table 1
Experimental systems used for validation. Although the systems are multimeric, for the purpose of the validation we only used one monomer. With the exception of the Rice Dwarf Virus test that already shows one monomer, the monomers were extracted by masking with the know high resolution structure of the monomer (indicated by PDB ID). Moreover, the masked maps were zero-padded with a 15 Å-wide boundary.

System units	EMDB ID	Resolution (Å)	Voxel size (Å)	Dimensions (voxels)	PDB ID	Helix count
GroEL	5001	4.2	1.06	86x86x97	3CAU	17
GroEL	1081	6.0	2.08	46x45x50	1OEL	17
GroEL	1200	7.8	2.28	43x42x46	1OEL	17
Proteasome	1740	6.8	1.38	69x58x81	3C92	11
Rice Dwarf Virus	1376	7.9	1.49	75x113x75	1UF2	33
Kinesin	1340	9.0	2.00	44x45x68	1JFF	23

Dimension – size of the map in voxels; PDB ID - ID of the corresponding high-resolution structure; Helix Count - total number of helices with 7 or more residues.

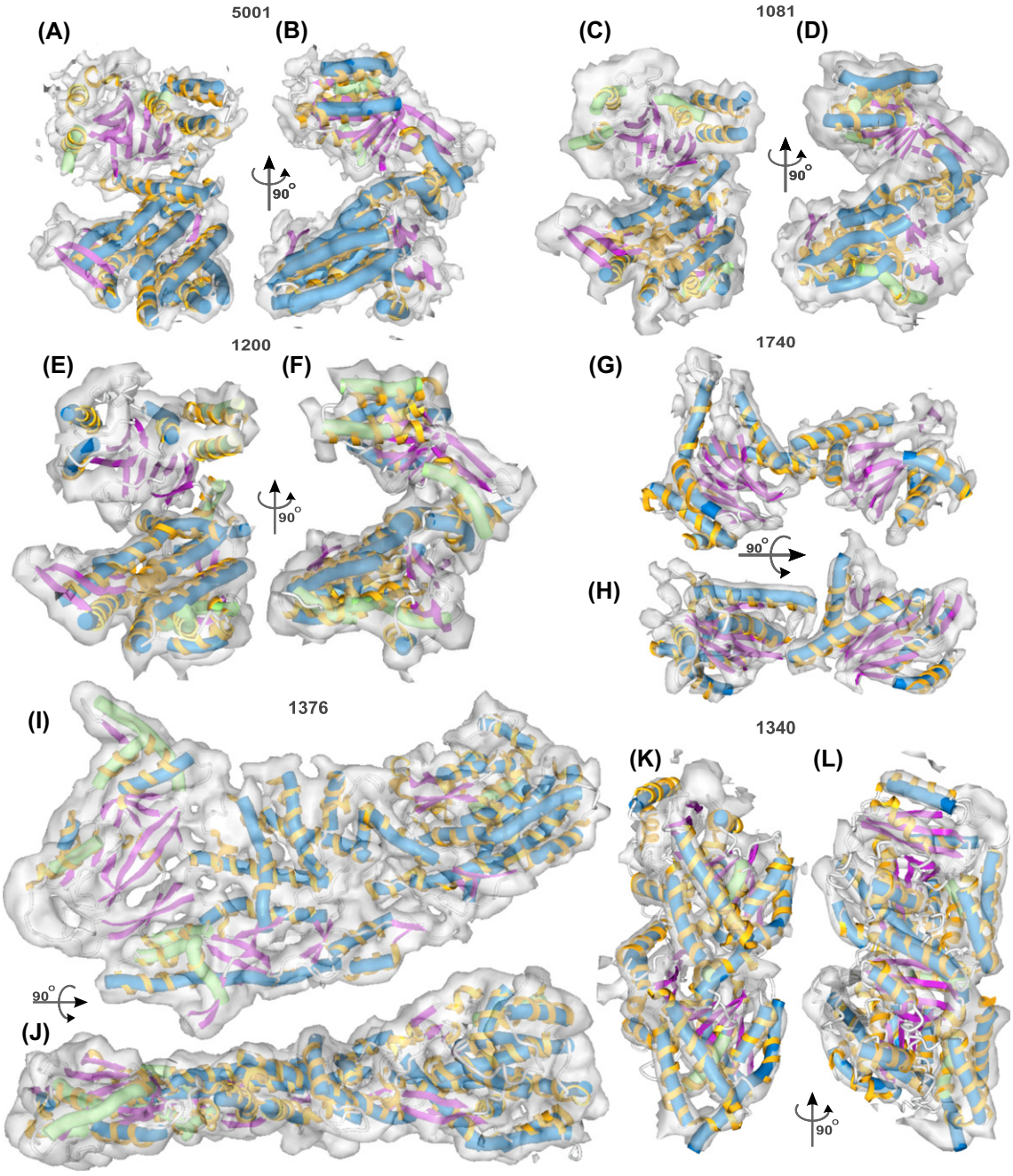

Fig. 5. VolTrac predictions for the experimental cryo-EM maps of GroEL (EMDB ID: 5001, 1081, 1200), 20S proteasome (EMDB ID: 1740), rice dwarf virus (EMDB ID: 1376), and kinesin (EMDB ID: 1340). The predictions are depicted using tube representations, blue for the helices predicted in the top N (column 'Helix count' in Table 2) scoring solutions, and green for false negatives ranked lower in the list of solutions. False positive predictions are not shown. Cryo-EM maps are shown in gray transparent surfaces, and the corresponding crystal structures (column 'PDB ID' in Table 2) use a yellow ribbon representation for the α-helices.

Table 2
VolTrac performance for experimental systems.

System (EMDB ID)	Helix count	TP	Total TP (Rank)	hSe	pSe	pPPV	Δ Turns	RMSD (Å)	Run time (min)	Expansion threshold
GroEL - 5001	17	14	16 (33)	82.4	86.3	81.7	0.80	1.14	14.4	70%
GroEL - 1081	17	12	16 (42)	70.6	70.9	73.6	1.20	2.06	12.3	55%
GroEL - 1200	17	12	17 (40)	70.6	66.4	70.1	1.10	1.76	9.8	60%
Proteasome – 1740	11	11	11 (11)	100	100.0	92.2	0.40	1.10	9.3	65%
Rice Dwarf Virus – 1376	33	26	33 (57)	78.8	86.5	69.2	1.40	1.12	39.0	70%
Kinesin – 1340	23	18	21 (27)	78.2	74.3	66.0	1.17	1.33	20.2	70%

Helix Count – total number of helices with 7 or more residues; TP - True Positive; Total TP - total number of known helices predicted; Rank - rank where the last helix was found (where 0 represents the best scoring helix); hSe - helix sensitivity (see text); pSe - point sensitivity; pPPV - point positive predictive value; Δ Turns - difference in number of turns between predicted and known helices; RMSD - root mean square deviation; Run time - execution time for the algorithm; Expansion threshold - stop criteria for the bidirectional expansion (see Section 2 for definition).

worse, we did not observe a clear correlation between performance and map resolution in experimental maps, as the results in the 4–7 Å resolution range depended on the particular experimental system and on the reconstruction quality.

False negatives often correspond to helices of short length or to those that fail to show the characteristic cylindrical rod shape. Such helices are often still detected but ranked lower in the solutions list. A visual inspection of the results may allow for the selection of such helices according to prior knowledge of the system. To reduce the risk of such false negatives, an user could lower the expansion threshold. The default value of the expansion threshold parameter was set to 70% for our experimental test cases, but it may be lowered to 50–60% for challenging maps that exhibit noisy helical features. Decreasing the value of the parameter entails a reduction in the pPPV and an increase in the length of predicted helices, so there is a tradeoff between tolerance and helical length. In our experimental test cases we were able to optimize the threshold based on the observed hSe values using known atomic structures. If crystal structures are unavailable, docking models may be substituted, or the user may use sequence-based secondary structure prediction (Jones, 1999; Cole et al., 2008; Meiler et al., 2001; Meiler and Baker, 2003; Karplus et al., 1997; Chandonia and Karplus, 1999) as a control.

We recommend to choose a starting expansion threshold value of 70% and decrease this value if the noise in the experimental map causes the helices to break into disconnected segments when compared with secondary structure prediction. In such a situation, the expansion threshold should be decreased progressively until helices are properly traced even when noise is present in the density. Such a decrease in expansion threshold will allow the template to navigate along the helix even when lower densities or gaps are present within the region. However, an extremely low expansion threshold may have detrimental effects on the results as template may 'bleed' into non-helical features.

Along with the expansion threshold, the dimensions of the template may also influence the outcome. By default a template with a 1 Å radius and a length of 8 Å is used. Although the radius is much smaller than the radius of a helix, the template allowed the proper tracing of the helical regions as opposed to values of 1.5 or 2 Å (data not shown). These results are due to the thinning of the map caused by the local normalization. Longer templates may be employed, but they are more prone to the 'bleeding' into not-helical regions and may encounter difficulties in properly tracing short helices.

Although rod-like densities typically correspond to α-helices, other structural elements may display similar patterns and generate false positives. Examples include the (anti) parallel β-sheet, which exhibits a smaller rod radius than the helices. An inspection of VolTrac results would permit the manual removal of such false positives. Moreover, several α-helices may occasionally be found in sequence, which become blended into a long cylinder without clear ends between distinct sub-helices. Such situations require additional information regarding the α-helix composition of the system, such as sequence-based secondary structure prediction, in order to identify the ends of any sub-helices. Sequence-based prediction methods may also be helpful in situations where the number of helices, N, is not known a priori. To allow for some leeway, we recommend to set the value of N to 10% above the value from the sequence based secondary structure prediction.

In summary, a number of new parameters were introduced for the algorithmic components of VolTrac. These parameters were derived based on geometric considerations and were tested empirically, as is typical for a proof-of-concept paper. We found that the performance of VolTrac was robust for the systems under investigation, but a more systematic refinement of the parameter values could be performed in future research. Based on our experience, these parameters (with the exception of the expansion threshold) should not require any fine-tuning by the user.

VolTrac incorporates parallel computing strategies to take advantage of the multi-core architecture of current workstations. Multiple genetic algorithm runs are launched in parallel. Although independent of each other, the parallel threads use a common tabu list. Each parallel run is finalized by a bidirectional expansion and by a global update of the tabu list before being re-launched until one of the stop criteria is met. The parallel implementation introduces an additional complexity: It is possible that two parallel runs independently identify the same helical region prior to its addition to the global tabu list. To eliminate such redundancies, the predictions are checked for overlap (using a distance criterion based on the known map resolution) and the lower scoring one is discarded when overlap is found. Due to the multi-threaded approach, a typical VolTrac run takes only minutes on a modern workstation. The run time may be decreased by reducing the sampling in the genetic algorithm or in the bidirectional expansion. Our default parameters currently favor sampling over efficiency to ensure a near-continuous exploration of the map. Specifically, we employ an angular step of 1° in contrast to earlier template convolution algorithms such as (Lasker et al., 2005) that employ larger angular steps up to 15°.

During iterative optimization, VolTrac often determines high-ranking helices first. Such characteristics prompted us to integrate VolTrac into the interactive graphics software Sculptor (Birmanns et al., 2011). The user can investigate the results on the fly as they are generated and stop the execution early if desired. VolTrac was included in Sculptor version 2.1, available at URL http://sculptor.biomachina.org. Because Sculptor is primarily intended to function as an interactive graphics program, it can become impractical to wait several minutes for results. Therefore, we have also added a full-featured (and parallelized) version of VolTrac as a command line tool to the Situs package, version 2.7, available at URL http://situs.biomachina.org. All our software is free, open source, and can be used on Linux, Macintosh or Windows computers.

Acknowledgments

We thank Stefan Birmanns, Zbigniew Starosolski, and Manuel Wahle for helpful discussions and for reading the manuscript. This work was supported in part by a grant from National Institutes of Health (R01GM62968).

References

Baker, M.L., Ju, T., Chiu, W., 2007. Identification of secondary structure elements in intermediate-resolution density maps. Structure 15, 7–19.
Baumeister, W., Steven, A.C., 2000. Macromolecular electron microscopy in the era of structural genomics. Trends. Biochem. Sci. 25, 624–631.
Baxter, R.H.G., Ponomarenko, N., Šrajer, V., Pahl, R., Moffat, K., Norris, J.R., 2004. Time-resolved crystallographic studies of light-induced structural changes in the photosynthetic reaction center. P. Natl. Acad. Sci. USA. 101, 5982–5987.
Birmanns, S., Rusu, M., Wriggers, W., 2011. Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes. J. Struct. Biol. 173, 428–435.
Braig, K., Adams, P.D., Brünger, A.T., 1995. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. Nat. Struct. Biol. 2, 1083–1094.
Chacón, P., Wriggers, W., 2002. Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol. 317, 375–384.
Chandonia, J.-M., Karplus, M., 1999. New methods for accurate prediction of protein secondary structure. Proteins: Struct. Funct. Bioinf. 35, 293–306.
Cole, C., Barber, J.D., Barton, G.J., 2008. The Jpred 3 secondary structure prediction server. Nucleic. Acids Res. 36 (suppl 2), W197–W201.
Cong, Y., Ludtke, S.J., 2010. Single particle analysis at high resolution. Method. Enzymol. 482, 211–235.
Dal Palù, A., He, J., Pontelli, E., Lu, Y., 2006. Identification of alpha-helices from low resolution protein density maps, In: Computational Systems Bioinformatics Conference, pp. 89–98.
Frank, J., 2002. Single-particle imaging of macromolecules by cryo-electron microscopy. Annu. Rev. Bioph. Biom. 31, 303–319.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA.
Holland, J.H., 1975. Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, MI.
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.
Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.
Karplus, K., Sjölander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., Sander, C., 1997. Predicting protein structure using hidden Markov models. Proteins: Struct. Funct. Bioinf. 29, 134–139.
Kovacs, J.A., Yeager, M., Abagyan, R., 2007. Computational prediction of atomic structures of helical membrane proteins aided by EM maps. Biophys. J. 93, 1950–1959.
Lasker, K., Dror, O., Nussinov, R., Wolfson, H., 2005. Discovery of protein substructures in EM maps. Algorithms in Bioinformatics, 423–434.
Lindert, S., Staritzbichler, R., Wötzel, N., Karakaş, M., Stewart, P.L., Meiler, J., 2009. EM-Fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure 17, 990–1003.
Liu, X., Jiang, W., Jakana, J., Chiu, W., 2007. Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a Multi-Path Simulated Annealing optimization algorithm. J. Struct. Biol. 160, 11–27.
Ludtke, S.J., Baker, M.L., Chen, D.-H., Song, J.-L., Chuang, D.T., Chiu, W., 2008. De novo backbone trace of GroEL from single particle electron cryomicroscopy. Structure 16, 441–448.
Ludtke, S.J., Baldwin, P.R., Chiu, W., 1999. EMAN: Semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. 128, 82–97.
Ludtke, S.J., Chen, D.H., Song, J.L., Chuang, D.T., Chiu, W., 2004. Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. Structure 12, 1129–1136.
Ludtke, S.J., Jakana, J., Song, J.-L., Chuang, D.T., Chiu, W., 2001. A 11.5 Å single particle reconstruction of GroEL using EMAN. J. Mol. Biol. 314, 253–262.
Meiler, J., Baker, D., 2003. Coupled prediction of protein secondary and tertiary structure. P. Natl. Acad. Sci. USA. 100, 12105–12110.
Meiler, J., Müller, M., Zeidler, A., Schmäschke, F., 2001. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J. Mol. Model. 7, 360–369.
Rabl, J., Smith, D.M., Yu, Y., Chang, S.-C., Goldberg, A.L., Cheng, Y., 2008. Mechanism of gate opening in the 20S proteasome by the proteasomal ATPases. Mol. Cell. 30, 360–368.
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.
Roseman, A.M., 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta. Crystallogr. D. 56, 1332–1340.
Rusu, M., Birmanns, S., 2010. Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions. J. Struct. Biol. 170, 164–171.
Sindelar, C.V., Downing, K.H., 2007. The beginning of kinesin's force-generating cycle visualized at 9 Å resolution. J. Cell. Biol. 177, 377–385.
Stagg, S.M., Lander, G.C., Pulokas, J., Fellmann, D., Cheng, A., Quispe, J.D., Mallick, S.P., Avila, R.M., Carragher, B., Potter, C.S., 2006. Automated cryo-EM data acquisition and analysis of 284742 particles of GroEL. J. Struct. Biol. 155, 470–481.
Tagari, M., Newman, R., Chagoyen, M., Carazo, J.M., Henrick, K., 2002. New electron microscopy database and deposition system. Trends. Biochem. Sci. 27, 589.
Wriggers, W., 2012. Conventions and work flows for using Situs. Acta. Crystallogr. D, in press. Invited proceedings article, 2011 CCP4 Study Weekend, University of Warwick, UK.
Wriggers, W., Agrawal, R.K., Drew, D.L., McCammon, J.A., Frank, J., 2000. Domain motions of EF-G bound to the 70S ribosome: Insights from a hand-shaking between multi-resolution structures. Biophys. J. 79, 1670–1678.
Wu, Y., Chen, M., Lu, M., Wang, Q., Ma, J., 2005. Determining protein topology from skeletons of secondary structures. J. Mol. Biol. 350, 571–586.
Yankovskaya, V., Horsefield, R., Törnroth, S., Luna-Chavez, C., Miyoshi, H., Léger, C., Byrne, B., Cecchini, G., Iwata, S., 2003. Architecture of succinate dehydrogenase and reactive oxygen species generation. Science 299, 700–704.