doi:10.1006/jmbi.2001.4633 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) **308**, 1033–1044

JMB

Bridging the Information Gap: Computational Tools for Intermediate Resolution Structure Interpretation

Wen Jiang[1]†, Matthew L. Baker[1]†, Steven J. Ludtke[2] and Wah Chiu[1,2]*

[1] *Program in Structural and Computational Biology and Molecular Biophysics*

[2] *Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston TX 77030, USA*

**Corresponding author*

Due to large sizes and complex nature, few large macromolecular complexes have been solved to atomic resolution. This has lead to an underrepresentation of these structures, which are composed of novel and/or homologous folds, in the library of known structures and folds. While it is often difficult to achieve a high-resolution model for these structures, X-ray crystallography and electron cryomicroscopy are capable of determining structures of large assemblies at low to intermediate resolutions. To aid in the interpretation and analysis of such structures, we have developed two programs: *helixhunter* and *foldhunter*. *Helixhunter* is capable of reliably identifying helix position, orientation and length using a five-dimensional cross-correlation search of a three-dimensional density map followed by feature extraction. *Helixhunter's* results can in turn be used to probe a library of secondary structure elements derived from the structures in the Protein Data Bank (PDB). From this analysis, it is then possible to identify potential homologous folds or suggest novel folds based on the arrangement of alpha helix elements, resulting in a structure-based recognition of folds containing alpha helices. *Foldhunter* uses a six-dimensional cross-correlation search allowing a probe structure to be fitted within a region or component of a target structure. The structural fitting therefore provides a quantitative means to further examine the architecture and organization of large, complex assemblies. These two methods have been successfully tested with simulated structures modeled from the PDB at resolutions between 6 and 12 Å. With the integration of *helixhunter* and *foldhunter* into sequence and structural informatics techniques, we have the potential to deduce or confirm known or novel folds in domains or components within large complexes.

Keywords: macromolecular assemblies; structure; fold recognition; electron cryomicroscopy; bioinformatics

Introduction

Through the efforts of X-ray crystallography, NMR, electron cryomicroscopy and modeling, the current structural database, the Protein Data Bank (PDB),[1] which has greater than 14,000 structures, represents only a portion of all existing folds.[2,3] Efforts in structural genomics have focused on establishing a highly representative library of unique folds, encompassing the vast majority of known atomic resolution models. One of the underrepresented areas in the current structural library is folds derived from large macromolecular assemblies. Thus, folds obtained from these assemblies may provide additional insight to known folds, as well as a source for novel structural arrangement.

Often, the initial structural analysis of a large complex is limited to intermediate resolution.[4–10] In the absence of atomic models, structural interpretation of large complexes at intermediate resolution is a formidable task, although structural information is still present. However, if individual domains, components and structural elements can be identified, pieces of the ''jigsaw puzzle'' may be put together to yield a tentative atomic model of the entire complex. With this in mind, the task at hand becomes the mining of structural and functional information of these complexes.

We have developed two computational methods which will facilitate the quantitative identification

†Contributed equally to this work.

Abbreviations used: PDB, Protein Data Bank.

E-mail address of the corresponding author: wah@bcm.tmc.edu of structural features in terms of known folds in three-dimensional (3-D) density maps at low to intermediate resolutions (Figure 1). This provides a means to identify structural homologs of components in large macromolecular complexes containing alpha helices. The first method, implemented in *helixhunter*, is used to analyze a 3-D map for alpha helix content at intermediate resolutions, 6-10 Å, the resolution at which such secondary structure elements become discernable. Correlation of secondary structure elements using alpha helix location and orientation, with known protein structures from the PDB, can then be used to identify similar folds. The second method, implemented in *foldhunter*, can be used to localize known or predicted folds/domains within larger macromolecular assemblies at lower resolutions (<20 Å). This could in turn lead to high-resolution interpretation of pieces within the macromolecular puzzle.

The development of *helixhunter* and *foldhunter* is just the first step in the multi-level structural interpretation process. Localization and identification of structural features provides the necessary template for the integration of these tools with sequence, structure and computation based analysis. Further development and integration could allow for the generation of structural and/or functional models based on low to intermediate resolution structures of macromolecules. The end result would represent a much more thorough understanding of macromolecular complexes.

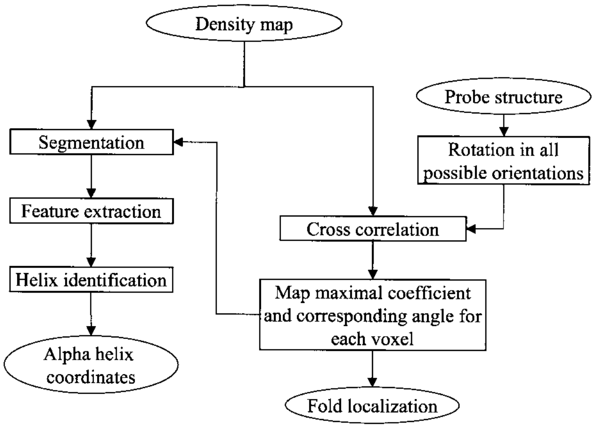

Figure 1. *Helixhunter and foldhunter* algorithm for identifying structural elements in intermediate resolution structure. The general flow of *helixhunter* and *foldhunter* algorithms are shown. The left side represents the basic *helixhunter* search, while the right-hand side represents the cross-correlation search of *foldhunter*. The central arrow indicates that the cross-correlation method can serve as a pre-processing step in *helixhunter*. The probe structure in the right side of the diagram can represent a prototypical helix in the case of *helixhunter* or a given sub-structure in *foldhunter*. Also, in the case of *helixhunter*, a five-dimensional search is done rather than a six-dimensional search in *foldhunter*.

Results

As seen in the PDB, protein structures vary considerably in alpha helical and beta sheet content, as well as adopting many types of folds and aggregation states. There have been many efforts to classify these structures, biochemically and structurally, into related groups. One of the most widely accepted classifications is SCOP, a Structural Classification of Proteins.[11] The majority of known structures can be broken-down into four major SCOP families: all alpha (α), all beta (β), alpha and beta (α/β), and alpha plus beta (α + β). To comprehensively sample these classes, we have examined over two dozen proteins with varying structural configurations. Here, we present four typical proteins from our larger test data set: bacteriorhodopsin (1C3W, α),[12] triose phosphate isomerase (1TIM, α/β),[13] tyrosine kinase domain from the insulin receptor (1IRK, α + β),[14] and Bluetongue Virus outer shell coat protein (1BVP, a β upper domain and an α lower domain).[15,16] These four proteins represent the major SCOP classes and are used to validate the results of *helixhunter* and *foldhunter*.

The *helixhunter* program, which identifies putative alpha helices, incorporates a multi-step process including cross-correlation, density segmentation, segment quantification, helix identification, and explicit description of the identified helices. The final helices are represented as cylinders, each specified by six parameters (three for center, two for orientation, and one for length). This abstract helix representation makes visualization of helices easy, as well as allowing for subsequent spatial fold recognition.

In many instances, the structures of fragments, domains or subunits of a larger complex are known by X-ray crystallography. The most popular method for fitting these parts into the complex structure is still manual fitting using visualization tools,[17] for which accuracy of the fitting is subjective. Here we provide a template based cross-correlation tool called *foldhunter* to fit the known structures objectively employing all six possible degrees of freedoms (three rotations and three translations).

Testing *helixhunter*

Each of the aforementioned proteins was modeled at 8 Å resolution, representing intermediate resolution structure determination by electron cryomicroscopy or X-ray crystallography (Figure 2). In all four cases, *helixhunter* correctly localizes better than 88% of the helices. Typically, helical assignment was within five degrees of the true orientation and within one turn of the correct length. This level of accuracy was largely unaffected by composition or size of the test model, which ranges from 24 to 116 kDa. Additionally, the misidentified helices were kept to a minimum. Out of the 54 total helices in the four proteins, only two non-helical segments were falsely identified as

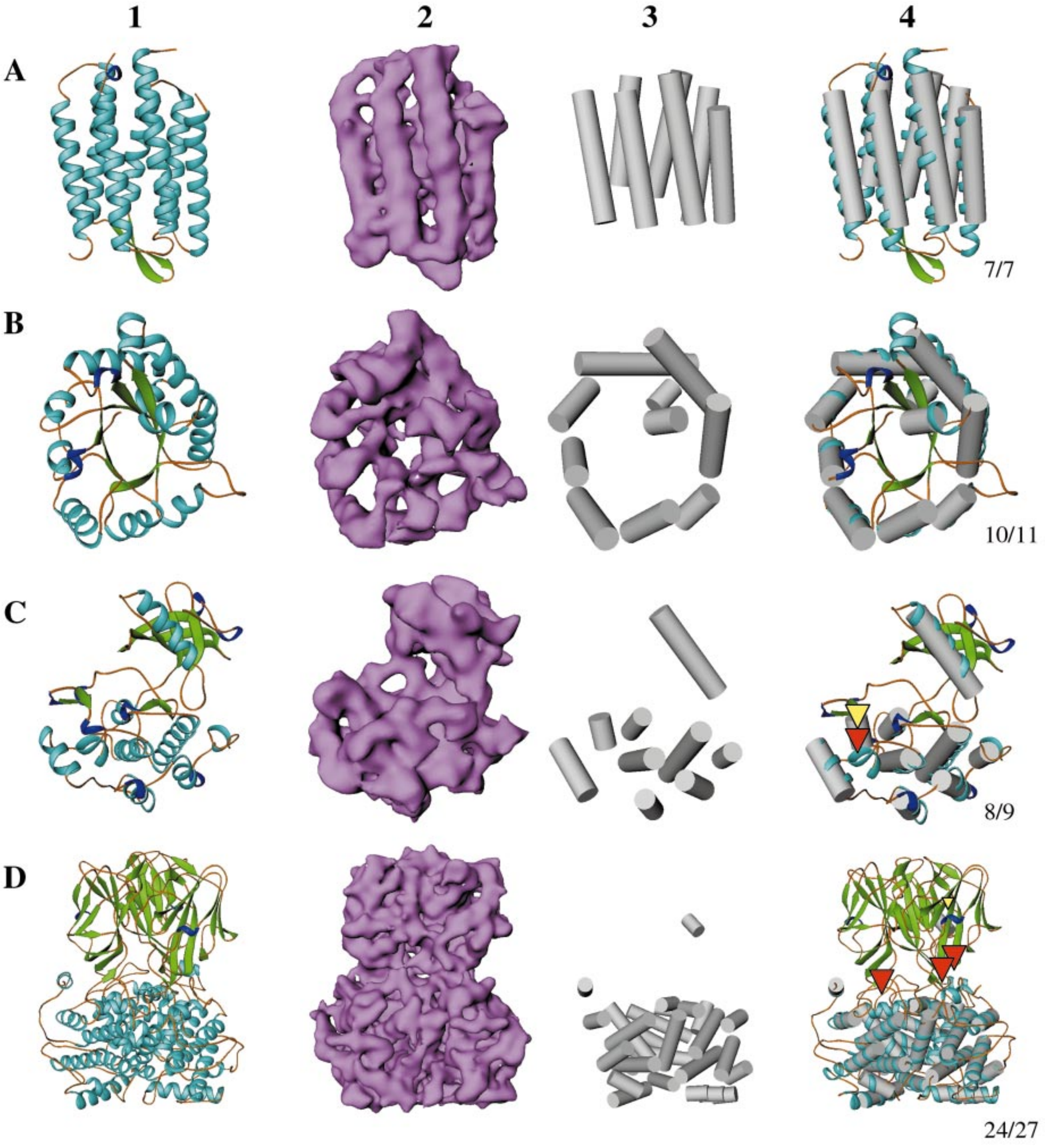

Figure 2. Testing *helixhunter* on model data. The four model structures: 1C3W (a), 1TIM (b), 1IRK (c) and 1BVP (d) are displayed as ribbons for their atomic structure in column 1, and as surface renderings of the simulated 8 Å structures, contoured at approximately the molecular mass, in column 2 (purple). Column 3 shows the found helices from *helixhunter* as cylinders (grey). Column 4 shows the superimposition of the found helices (cylinders) onto the ribbon diagram of the known structures. The red triangles show helices not identified by *helixhunter*, while the yellow triangles represent false positives. The number of correctly identified helices *versus* the total number of helices is listed just after column 4. In the simplest case, bacteriorhodopsin (1C3W), a seven transmembrane helix protein, *helixhunter* was able to identify the location and orientation, within five degrees, of all the helices. Additionally, the length of the *helixhunter* assigned helices was in agreement to within one turn of the known helices. In the triose phosphate isomerase (1TIM) model, an alpha/beta protein, a very high degree of agreement was seen between the real and predicted helices using *helixhunter*. All but one short helix, smaller in length than the prototypical helix in the correlation routine and immediately adjacent to a long helix, was not identified. The insulin receptor tyrosine kinase structure (1IRK), an $\alpha + \beta$ protein, also showed a good correlation with the *helixhunter* helices and actual helices, missing only a short 1.5 turn helix. In the final test model, a trimer of BTV VP7 containing an all-helical lower domain and an all beta sheet upper domain, 24 of the 27 helices were correctly identified, agreeing in location and length. A total of three short helices were missed and a single turn in the beta sheet region was misidentified.

a helix, while four helical segments were not identified. Typically, these missed or incorrectly assigned segments were either short helices or turns.

Structure identification

Due to the fact that *helixhunter* ultimately provides the user with spatial identification of helices, we have an explicit representation of all helical parameters. Thus, the information encoded by the parameters as well as the relative position and orientation of the helices provides an extremely coherent descriptor for protein structural features. Using this information, we used the structural matching programs, *DejaVu*[18] and *COSEC*[19,20] to identify homologous structures based on spatial arrangement of secondary structural elements of a protein using a library based on protein structures in the PDB.

The assigned helices from the *helixhunter* results on the 8 Å resolution models were compared to a modified *DejaVu* library of secondary structure elements generated from the structures in the PDB. Since the C^{α} positions of the polypeptide are not determined in an intermediate resolution structure, we modified the operational definition of alpha helices from the original library of secondary structure elements in *DejaVu*, minimizing the potential center and orientation mismatch as described later in Materials and Methods. Table 1 shows the assignment of the top three putative homologous structures from *DejaVu* to each of the four *helixhunter* results for the model structures. In all cases, the structural assignment from *DejaVu* represents the same structure (same PDB identifier) or related isoforms (same protein, different PDB identifier) to the original proteins tested in *helixhunter*. The center-to-center distance between the corresponding alpha helices of the probe and the putative structural homolog is reported as root mean squared (RMS) deviation. The scores of the top *DejaVu* matches for the *helixhunter* results are similar to those of the scores from matching the secondary structure elements directly defined by the PDB file (data not shown). Of further interest, illustrated in 1IRK, the top identified structures found by *DejaVu* corresponded not only to itself but also with a variety of other serine/threonine kinases, which all share a common fold. Similar results were also obtained using the *COSEC* program and *helixhunter* results (data not shown), verifying the accuracy of the helical assignment in *helixhunter*.

Testing *foldhunter*

Each of the four model structures and their related sub-structures were modeled at 8 Å. These sub-structures were centered in a cube of the same size as the complete model data, rotated and translated from the original position, and then used as probes against the intact model structure using *foldhunter*. In each of the four models, the corresponding sub-structure was correctly localized and oriented with respect to the intact structure within two degrees. As with *helixhunter,* composition and size of the target and the probe sub-structure did not affect the accuracy of localization significantly.

1C3W was probed with a segment containing three alpha helices and two small strands. This sub-structure was placed back into the intact density at the correct position and orientation. The sub-structures for 1TIM, two helices and two strands, and 1IRK, the beta sheet portion, were

Table 1. *DejaVu* output using *helixhunter* results

Model	Identified structure	Name	RMSD (score)
1C3W	1BM1	Bacteriorhodopsin in light adapted state	0.87 (0.73)
Bacteriorhodopsin	1AT9	Bacteriorhodopsin	1.02 (0.87)
	1BRX	Bacteriorhodopsin/lipid complex	1.66 (1.35)
1TIM	8TIM	Triosephosphate isomerase	0.91 (0.81)
Triose Phosphate	1TCD	Triosephosphate isomerase-*Trypanosoma cruzi*	1.24 (1.10)
Isomerase	7TIM	Triosephosphate isomerase complex	1.59 (1.43)
1BVP	1BVP	Bluetongue virus vp7	1.61 (1.37)
Bluetongue Virus	1FIY	Phosphoenolpyruvate carboxylase	3.27 (2.95)
VP7 Monomer	1AXG	Alcohol dehydrogenase	3.59 (3.18)
1IRK	1IRK	Insulin receptor tyrosine kinase domain	0.95 (0.86)
Insulin Receptor	2PHK	Phosphorylase kinase	1.41 (1.21)
Tyrosine Kinase	1PTK	Chicken src tyrosine kinase	1.45 (1.27)
Domain	1AQ1	Human cyclin dependant kinase 2	1.63 (1.40)
	1PHK	Phosphorylase kinase	1.64 (1.46)
	1FGI	Tyrosine kinase domain (fibroblast growth factor)	1.71 (1.55)
	1AD5	Src family kinase complex	1.75 (1.56)
	1CMK	Camp-dependent protein kinase catalytic subunit	1.82 (1.58)

Listed are the results of the *DejaVu* program with the modified helix library and the helices defined by *helixhunter* (Figure 2) for the four structures modeled at 8 Å resolution (first column). The second and third columns denote the PDB identifier and name of structure found by the *DejaVu* search, respectively. The results were ordered by RMSD (in Å), which is found in the last column.

also fitted to the correct location within their respective intact structures. The beta sheet region of the BTV VP7 trimer was also used to probe the intact structure. As with the other three structures, the correct location of the sub-structure was identified (Figure 3).

Integrating fold recognition with *foldhunter*

In the case of large macromolecules, often only sequence-derived homologous sub-structures are known. To better represent this type of situation in *foldhunter*, we tested one of the four model data sets, 1BVP, against several related proteins (Figure 4). A number of putatively homologous folds with varying degrees of homology to the β sheet domain of the 1BVP protein were tested to determine whether *foldhunter* was capable of correctly localizing these homologs. The VP7 monomer contains both a β sheet region in the upper domain and a helical region in the lower domain. Using the UCLA-DOE Fold-Recognition Server,[21] several candidate folds were identified as potential structural homologs to the β sheet domain of the 1BVP. The best matches were to itself and the African Horse Sickness Virus capsid protein,[22] while tumor necrosis factor alpha[23] had a borderline score for structure homology.[24] Figure 4(c) illustrates *foldhunter*'s ability to correctly place the capsid protein from African Horse Sickness Virus into the domain of the 8 Å map of the 1BVP. However, the lesser-valued candidate structure, tumor

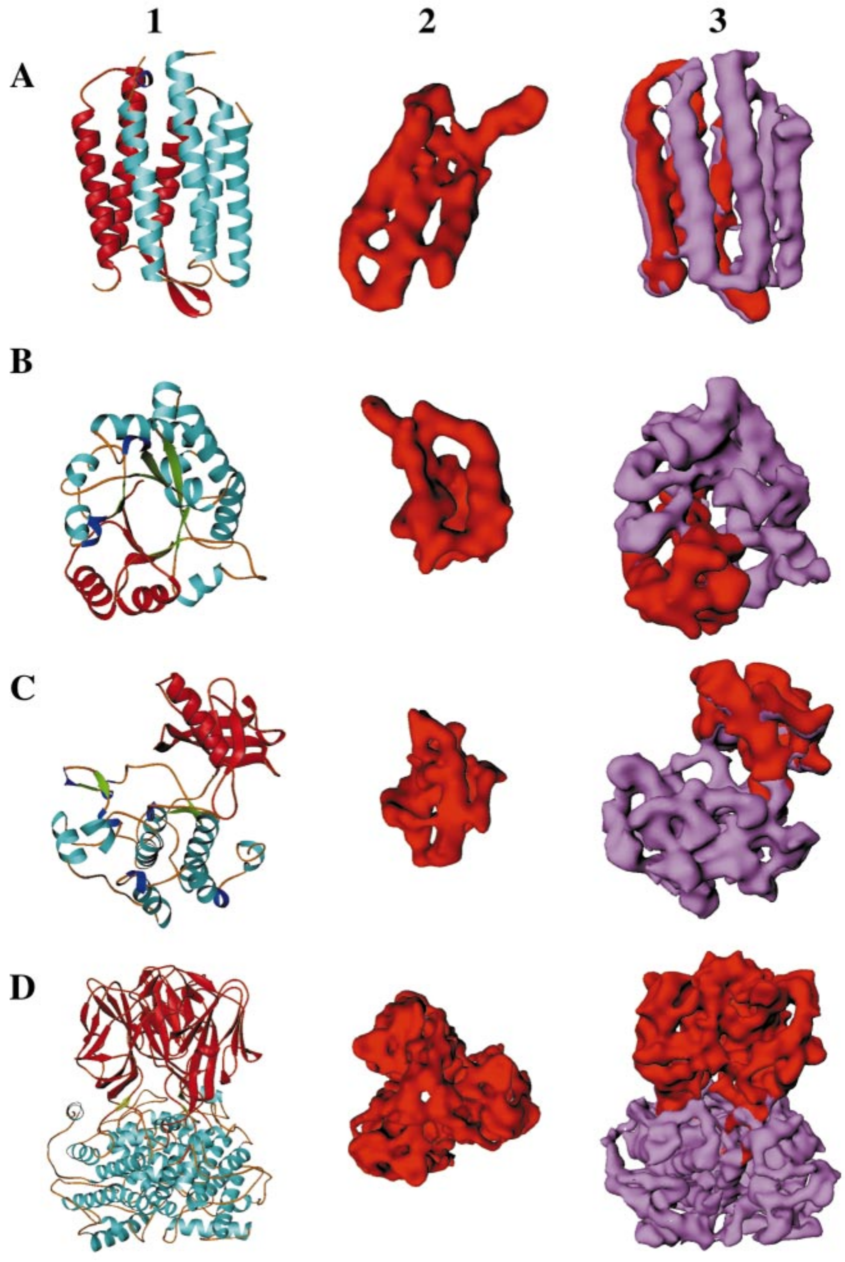

Figure 3. Testing *foldhunter* on model data. The four model structures: 1C3W (a), 1TIM (b), 1IRK (c) and 1BVP (d) are displayed as ribbons in column 1. Portions from these structures, 1C3W (5-99), 1TIM (1-64), 1IRK (981-1078) and 1BVP trimer (117-261), were extracted (column 1, red) and modeled at 8 Å resolution (column 2). Results of *foldhunter* using these extracted portions to map on the entire structure are shown in column 3.

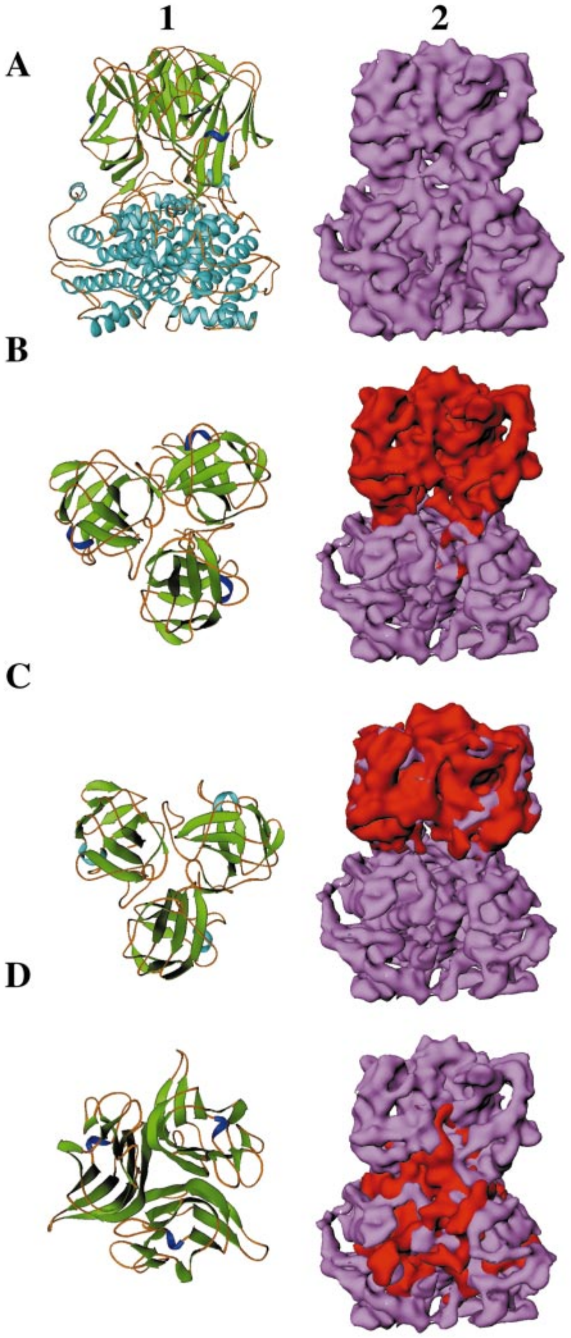

Figure 4. Sensitivity of *foldhunter* with homologous structures. The sensitivity to which *foldhunter* can detect a homologs fold is shown using 1BVP (a) against the beta barrel segment of 1BVP (b), a close homolog (c) (1AHS), and a remote homolog (d) (2TNF) as represented by the ribbon diagrams in column 1, respectively. Results of *foldhunter* are seen in column 2.

necrosis factor alpha (Figure 4(d)), was misplaced in the target structure as seen by visual inspection. Though *foldhunter* can always yield an answer, the match has to be judged by both the correlation coefficient and also visual matching of the two structures. This result suggests that a correctly chosen fold can be placed in the correct region of the target structure by *foldhunter*.

Resolution dependency of *helixhunter* and *foldhunter*

The resolution dependency of *helixhunter* was assayed using 1C3W, modeled at 6, 8, 10, and 12 Å resolution (data shown at http://ncmi.bcm.tmc.edu/~wjiang/hhfh). *Helixhunter* was able to very accurately identify the helices in both the 6 Å and 8 Å models. In the 10 Å model of 1C3W, helix position and orientation retains high fidelity, but helix length is not as accurate. At 12 Å, similar helix positioning to that of the 10 Å model can be seen, however false helices are assigned to the small beta strand region and a turn. Along with the intact structure, the sub-structures for 1C3W were also modeled at the four aforementioned resolutions to test *foldhunter*. Assignment of the correct location and orientation of the sub-structure occurs at all resolutions. Additional testing of *foldhunter* was done using a lower resolution model (20 Å) of bacteriorhodopsin. While the sub-structure was localized correctly, a smaller angular search step was required.

Discussion

The trend in structural biology is to study larger and larger macromolecular assemblies in different functional states. These assemblies generally contain multiple subunits, with a variety of structural elements. The divide and conquer approach has been used to determine the atomic structure of the components or domains, as well as the molecular blueprints for the entire assembly at a lower resolution.[25] While this approach works for some systems, it may not universally apply. In some cases, it may be difficult to obtain an atomic model of the components for various biochemical reasons, as would be the case if a component was in a molten globule state while expressed in the absence of other components.[26] Electron cryomicroscopy and X-ray crystallography have offered opportunities to study large assemblies as a whole at low to intermediate resolutions. As we approach higher resolution, information regarding the structural elements becomes discernable. These elements, from secondary structure to entire domains, must be assembled in the context of the molecular map in order to obtain an atomic model. In this study, we have developed a set of computational tools with which we can mine the data to infer a protein fold or domain from intermediate resolution structures of a large complex. Such analysis could lead to the structural and functional characterization of not only single domains of a large complex, but also the complex itself.

Alpha helix identification

The most definable structural elements in intermediate resolution structures are alpha helices, as they typically have regular, identifiable features. This has allowed for the interpretation of helices based on simple visual inspection and assignment within the map.[4,5,10] However, experience, bias and interpretation lead to a fairly subjective, non-quantitative and possibly conflicting assignment of helices. We have thus implemented *helixhunter* and tested it on a variety of different structural models, varying in resolution and structural configuration (Figure 2).

From the results of the model data, *helixhunter* has demonstrated the ability to accurately identify helix position, orientation and length at intermediate resolutions. As the resolution of the model data approaches atomic resolution, the reliability of helix assignment increases. At the lower resolutions, short false helical assignments were noticed, while only well defined, long helices were correctly identified in length and location. This resolution dependency is not unexpected. Measurements on the closely packed helices in bacteriorhodopsin reveal minimal helix-helix distance between 10 to 12 Å. This is in agreement with our helical assignments in the resolution dependency test as helix assignment began to deteriorate in accuracy on the 12 Å 1C3W model (http://ncmi.bcm.tmc.edu/~ wjiang/hhfh).

The ability to identify the alpha helices is sensitive to the thresholding of the density map. We apply a segmentation procedure using a neighbor-grouping algorithm, which relies on the ability to separate features into distinct objects. Differences in density between helices and adjacent features in the density map can be subtle. To decrease the threshold sensitivity of the segmentation process, we have adopted a pre-process filtering step using cross-correlation. By correlating the density map with a prototypical helix, we up-weight the helical density and thus suppress the non-helical density. Our segmentation process is thus carried out with the normalized correlation coefficient map rather than the original density map. The improvement results in considerably less threshold dependency, for which successful segmentation of helical densities can be done.

At intermediate resolutions, helices most closely resemble cylinders with a relatively fixed radius of ~2.5 Å and variable length. Thus, by using the basic character of a cylinder, we can evaluate the candidate segments for potential helical assignment. However, non-equivalency of side-chains and bends along the helical axis cause deviation from the idealized cylinder model. Therefore, a cross-section about the helical axis could be described as approximately as an ellipse, with a semi-major and a semi-minor axis (see Materials and Methods). We can thus use the relative values of these two axes as criterion during helix identification, as significant deviation would indicate a non-cylindrical, and therefore non-helical density. As the backbone of a helix has a fairly constant diameter at 5 Å, the two semi-axis values should be close to 2.5 Å. Additionally, helix length can also be used as a criterion in helix identification by specifying minimum and/or maximum values for helix length.

Fold recognition based on alpha helices

In the classification of proteins, families can be defined by the spatial arrangement of secondary structure elements, which in turn determine structurally homologous proteins. Thus, in addition to visual interpretation of structural elements, *helixhunter* allows for structure-based fold recognition when used in conjunction with *DejaVu* and *COSEC*. The spatial relationship of *helixhunter* identified helices can be compared to a large database of known secondary structure elements from published PDB files. This assumes that the identified alpha helices accurately represent nearly complete architectural information capable of distinguishing different folds. As an example, true homologous structures for 1IRK were found only when four or more of the longest helices were used. Four helices, half of the helices in 1IRK, thus represent enough unique information regarding the fold. However, the minimum structural information sufficient for fold prediction is likely to vary amongst different protein folds. Thus, by providing a reasonable set of *helixhunter* generated helices as input, we are able to screen all known protein structures for similar occurrences of these helical elements. This serves two purposes: (1) verification of our approach, and (2) possibility of intermediate resolution structure-based recognition of alpha helical folds. While the first of these two objectives is primarily used in testing the model data, it is the second that is of primary interest when examining real data (M.L.B. *et al*., unpublished). Identification of homologous folds or domains of proteins may then be done in the absence of any sequence-derived information. While in some cases a homologous fold may be identified, the absence of a structural relative may suggest the presence of a novel fold.

In Table 1, which summarizes the *DejaVu* results for each of the four test models, we see that the RMS deviation and score are not zero for itself, which would indicate a less than perfect match. While the segmentation step, as mentioned earlier, has been corrected for some of the thresholding dependency, helix length and orientation may still be affected by the threshold. The ends of the helices are less well defined than other regions of the helices, making them more sensitive to threshold effects.

Based on these test data, it appears that the intermediate resolution structure-based fold recognition is feasible and reliable. However, while dealing with the experimental data, the reliability of this method is limited by the accuracy of the map.

Inaccuracies in helix identification and thus fold recognition are considerably higher in a poorly defined map. Furthermore, while two structures may appear visually similar, *DejaVu* and *COSEC* may not identify the structures as being strictly homologous. The quantitative identification for these two programs relies on a strict matching of corresponding helical elements. Local differences in the alpha helices may thus cause insufficient spatial overlap of the secondary structural elements. Therefore, remote homologs may not be detectable by *DejaVu* and/or *COSEC*.

Fold identification and localization

In the ideal case, results obtained from *helixhunter* and *DejaVu/COSEC* will produce a correctly identified homologous structure that can be mapped directly to the low or intermediate resolution structure. As this fold identification is based solely on alpha helices, this method can work with a variety of folds containing a significant amount of helices, such as TIM barrels (1TIM) and protein kinase-like (1IRK). However, recognition might not always be possible, as would be the case for beta-sheet rich proteins. Thus, using structural information about domains or sub-structures, obtained from either sequence-based fold identification or structural homology searches, the task becomes assigning the correct position and orientation to a more general sub-structure within the larger structure.[24] The visual assignment of sub-structure position and orientation to a larger complex is somewhat subjective. Current methods of aligning structures, such as *Situs*,[27] require that the relative location of the probe structure is known within the target structure. In this regard, the corresponding target volume needs to be excised from the large target structure. Thus, this type of method relies on prior knowledge of the corresponding region between the probe and target structures. *Foldhunter* does not have this limitation, as a relatively small portion or sub-structure will be localized in a much larger complex using the cross-correlation. As demonstrated by the fitting of homologous structures, identified by fold recognition, *foldhunter* has shown the ability to correctly localize relatively similar structures. Thus, *foldhunter* serves not only as a method to localize known folds or domains, but also confirms if a predicted homologous fold is indeed truly homologous to the intermediate resolution structure.

Foldhunter, which performed well at intermediate resolutions, also has the ability to work on much lower resolution structures. Since it docks a sub-structure into a larger structure based on the distribution of density, the major requirements of the target and probe structures are their relative structural features. Probe and target structures with relatively well-defined features at low resolutions provide suitable ''landmarks'' for *foldhunter*, making it possible to localize sub-structures at 20 Å or lower resolution.[28]

Building structural and functional models

In intermediate resolution structures, an explicit backbone trace of the protein is not available, resulting in the inability to correlate the amino acid sequence and key functional residues in the context of the structure. However, with the information obtained from the *DejaVu* and *COSEC* searches in conjunction with additional computational analysis (homology modeling, secondary structure prediction, sequence alignments, etc.), one may conceivably produce directionality for each helix and connections between the helical elements. Through further developments and integration, the resulting information may eventually allow for the assignment of sequence to the structural elements, yielding a rough backbone trace.

With a backbone model and the correlation of biochemically and/or genetically relevant residues to the structural features, it may be possible to suggest a putative function of the molecule, providing additional guidance for future biochemical experiments. As an example, the *helixhunter* results for 1IRK, the tyrosine kinase domain from the insulin receptor, indicate the presence of eight helices. When submitted to *DejaVu*, all of the top structures identified belonged to tyrosine kinases or serine/threonine kinases (Table 1). If this structure had been unknown or was an unrelated sequence through convergent evolution, these identified structures would have accurately provided both functional and structural information. The related homologous structures, when aligned to the ''unknown'' structure's structural elements would have also provided a template for further structural modeling and biochemical analysis. Extending this approach beyond a single domain protein, it may be possible to identify structural and functional domains of large complexes. These domains, often composed of non-contiguous sequence segments, might otherwise be impossible to differentiate based on a purely sequence-based analysis, as seen in proteins from a large virus (M.L.B. *et al.*, unpublished results).

In large multi-protein complexes, the atomic resolution models of the individual proteins may have been determined. While this provides us with interesting structural and functional information for the individual proteins, less information is given about the localization, association, and function within the larger complex. By using *foldhunter* to accurately localize and orient a protein (or sub-structure) within the complex, we can obtain previously unavailable information regarding the positioning of the protein within the complex. The potential information from the localization is two-fold: (1) positioning of individual proteins within a complex and (2) relative contacts and associations of proteins within the complex. Thus, the integration of structure and function with positional information provides a much clearer picture of the overall mechanism of macromolecule complex function.

Here we demonstrate the validity of two programs primarily targeted at the interpretation of low to intermediate resolution structures, obtained by electron cryomicroscopy or by X-ray crystallography at an initial stage of analysis.[4–10] It has been demonstrated here that both *helixhunter* and *foldhunter* provide a basic means of bridging and integrating multiple sources of structural and functional information from both known or potentially novel folds. Instead of being viewed solely as methods for interpretation of structure, these programs may be viewed as two nodes in the structure/function prediction network. Not only are they capable of correlating existing information, but they also provide insight into the function, composition and interaction of sub-structures within a larger complex, seen in the 2.4 MDa calcium release channel.[28] Furthermore, such structure analysis may prove valuable in structural genomics, by quickly screening large assemblies for potentially unique or homologous structures.

Materials and Methods

Density segmentation

At intermediate resolutions, the densities for helical regions appear to be higher than other regions of the density map, with the highest density located along the central axis of the helix. By masking out densities below an appropriate threshold value, most of the non-helical density in the map can be eliminated. The extracted densities can then be segmented into separate objects, consisting only of connected voxels. To enhance the helical regions, while suppressing the other non-helical regions, we have implemented a pre-processing step to re-scale the density map. This pre-processing step involves the commonly used template-based cross-correlation technique, by which a centered prototypical helix, with a Gaussian-like radial density distribution and width of 5 Å oriented at different angles, is correlated to the density map. The highest correlation coefficient among all the prototypical helix orientations is recorded for each voxel. The helical regions would therefore have larger correlation coefficients than other regions, resulting in a coefficient map with greater differences in values between helical regions and non-helical regions than the original density map. Two variants of the correlation function,[29] the cross-correlation function (CCF) and mutual correlation function (MCF), were implemented as user selectable options in the pre-processing step. From this ''enhanced'' map, density segmentation, as described above, may result in cleaner identification of helices.

Segment quantification

Critical to segment characterization is the definition of segment shape and related helical axis, assuming the criteria for helix shape have been met. After segmentation of the density map, the attributes (center, mass, volume, second moments tensor, principal axes, length, and width) for each segment are quantified (equations (1)-(3)):

Mass:

$$m = \int f(\vec{\mathbf{x}})d\vec{\mathbf{x}} \qquad (1)$$

Center:

$$\vec{\bar{\mathrm{X}}} = \frac{1}{m}\int f(\vec{\mathbf{x}})\vec{\mathbf{x}}d\vec{\mathbf{x}} \qquad (2)$$

Second moment tensor:

$$M_{ij} = \frac{1}{m}\int f(\vec{\mathbf{x}})(x_i - \bar{x}_i)(x_j - \bar{x}_j)d\vec{\mathbf{x}}, \ (i, j = 0, 1, 2) \qquad (3)$$

Here, $\vec{\mathbf{x}}$ is the voxel, $f(\vec{\mathbf{x}})$ is the density value and i,j are dimensional indices.

Among all the attributes, the 3 × 3 second moments tensor (equation (3)) is the most valuable attribute, as it describes the nine essential shape parameters of the segment density. This attribute is used in computing the nine parameters (two angles representing each of the three principal axis directions and three defining segment dimensions) by eigen-analysis.[30,31] The tensor matrix is symmetric and can be diagonalized using the Jacobi transformation.[31] The resulting three eigen-vectors then become direction vectors of the three principal axes. The three eigen-values represent segment dimensions, with segment length represented by twice the largest of the three eigen-values. The remaining two eigen-values represent segment radii about the segment axis.

Helix identification

The relationship among the three eigen-values describes the shape of the object, which for helices at intermediate resolutions is cylindrical. This therefore requires the density segment to satisfy conditions based on cylindrical parameters. The largest eigen-value must be significantly larger than the two smaller eigen-values, while the two smaller eigen-values must be similar in value. Additionally, the two smaller eigen-values should be less than 3 Å, representing the alpha helix radius. Eigen-values outside these conditions indicate that the segment does not correspond to an alpha helix. In addition to the eigen-values, minimal segment length can also be included as a criterion for identifying a helix. By specifying a ''short-helix'' length, small non-cylindrical densities that still meet the initial eigen-value criteria are excluded, decreasing the likelihood of false peaks. The default short-helix value was set to 9 Å, representing just under two complete turns.

Representing helices

As stated previously, at intermediate resolution an alpha helix can be represented as a cylindrical object. This can now be defined by seven parameters: three for the center position, two angles for the axis direction, one for the helix length, and a constant diameter of 5 Å. Once these parameters are determined, the assigned helices can then be visualized and subjected to further processing. Currently, four output formats are supported: Open Inventor, VRML, SSE (*DejaVu*) and *COSEC* format. The Open Inventor format output can be displayed using 3-D visualization tools, such as *Iris Explorer* (NAG, Downers Grove, Illinois), where the helices are shown as stylized cylinders. This format was the primary method used for visualization, as the cylinders (helices) and the density map can be viewed simultaneously. Apart from helping to visualize helices, the explicit description of the starting and ending position of each helix can be represented in a secondary structure element file. The topological structural information encoded by the relative positions and orientations among the helices make it possible to search for structural homologs, even in the absence of an atomic model. This is accomplished by using the ''bones'' search in the *DejaVu* package, which ignores directionality and sequential order of structural elements. Additionally, representing helices in a vectorized format, we can do a similar search for structural homologs using *COSEC*.[19,20]

Rebuilding helix specific *DejaVu* library

The operational definition of alpha helices differs between *DejaVu* and *helixhunter*. In *helixhunter*, the helix is represented by starting and ending positions along the central helical axis. In contrast, the original *DejaVu* library uses only the first and last C^{α} as the starting and ending coordinates in the spatial definition of the helix. Therefore, when both termini occur on a single side of the helix, the positioning of helix could be off by approximately the radial distance of a helix, 2.5 Å. However, when the termini of the helices reside on opposite sides, the orientation of the helices could be subject to a few degrees of error. To minimize this effect, the secondary structure library was rebuilt using the projections of the starting and ending C^{α} onto a centrally fitted helical axis. Additionally, this new library excludes non-helical structures making it more suitable for analysis of *helixhunter* results.

Definitions of helical segments were extracted directly from the PDB headers. No additional parameterization of helical character was done, contrary to the *DejaVu* method, which uses its own criteria for helix definition. Fitting of the central helical axis was done using the rotational least-squares (*rotfit*) algorithm.[32] The C^{α} endpoints were then mapped to the central axis, yielding an accurate representation of the helix. The library of SSE files were then constructed using this technique with a relatively recent version of the PDB (July 1999, release no. 89). Shown in Figure 5, is bacteriorhodopsin, 1C3W, represented using the original *DejaVu* method and the central axis method. Helical positioning in comparison with the atomic model suggests an increase in general fidelity using the central axis description of helices.

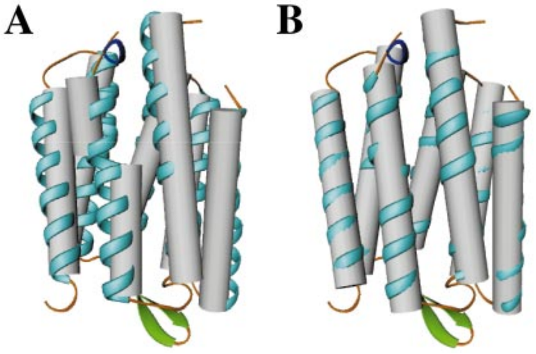

Figure 5. Comparison of helix representations. Shown in (a) is a ribbon model for 1C3W, bacteriorhodopsin, and its corresponding helices using the original *DejaVu* representation method. In (b), the new central axis representation of helices is shown, in addition to the same 1C3W ribbon model. There are noticeable differences in the helical axis orientation. Artifacts in the original *DejaVu* representation method include tilting and lateral translation of the helix center. No such artifacts can be seen in the central axis representation.

Structure fitting

The technique used by *foldhunter* is also a correlation-based method. The fragment or sub-structure being localized within the reconstructed model is rotated to all possible orientations (three degrees of freedom). A 3-D cross-correlation map is then calculated and the maximal correlation coefficient and the corresponding orientations are recorded for each voxel. A sorted list of best solutions (rotation and translation) is then produced. The top solutions, rotational and translational parameters, represent the likely fit of the probe structure to the target structure. Thus, the *foldhunter* results reflect only the most probable solutions relative to the accuracy of the probe, target map, and angular search sampling. Therefore, it is necessary to visually inspect the resulting fitted structures.

For a 3-D structure alignment, the exhaustive six-dimensional search is computationally demanding, however, with current computational resources, this process falls within acceptable time constraints for most searches. For instance, just under three hours was required to probe a structure of 116 kDa on a single Origin 2000 R10000 processor (195 MhZ). *Foldhunter* and *helixhunter* have been parallelized for use on shared-memory supercomputers. Thus, even for large-scale problems, computational time requirements should not be prohibitive.

Generating model data

Density maps used for testing were generated using *pdb2mrc*, a program included in the EMAN suite.[33] For each atom in the PDB model, a Gaussian, $n_e e^{-\left(\frac{r}{k}\right)^2}$ is generated in the corresponding position in the density map, where n_e is the atomic number and r is the distance from the center of the atom. The Gaussian width, k, is constant for all atoms to approximately simulate a 3-D structure at moderate resolution. For an 8 Å resolution 3-D structure, k is equal to $8/\sqrt{2}$. This corresponds to starting with an atomic resolution model and applying a Gaussian Fourier filter with a half-width of $1/8$ Å^{-1}. The resolution is sufficiently low that the atomic form factor is negligible, and thus excluded. While this is not a perfect test of how the algorithm will operate on real data, it is a reasonable simulation for evaluation of the *helixhunter*/*foldhunter* algorithms.

Four representative proteins, comprising four distinct super-families in the SCOP classification, were downloaded from the PDB. These structures were then converted into cubic electron density maps at 8 Å resolution using *pdb2mrc*. All structures were generated using a sampling of 1 Å^3 per voxel and centered in a 64^3 map, except for 1BVP, which was centered in a 128^3 map. Additionally, individual sub-structures were extracted from the four test PDB sequences and generated under the previously mentioned conditions. Thus, a standard test set of four test models and four sub-structures were generated for use in *helixhunter* and *foldhunter*.

Two additional data sets were generated, one for testing the resolution dependence of *helixhunter* and *foldhunter*, the other for testing the robustness of *foldhunter*. The bacteriorhodopsin structure (1C3W) and its corresponding sub-structure were generated at 6, 8, 10, and 12 Å resolution, representing current intermediate resolution range of electron cryo-microscopy. The third data set consisted of the Bluetongue Virus VP7 timer (1BVP), the beta-sheet trimer region from BTV, a trimer from the African Horse Sickness Virus major capsid protein (1AHS), and tumor necrosis factor alpha (2TNF). These structures were generated at 8 Å resolution in a 128^3 map as mentioned above.

Testing with model data

For each of the four intact structures from the first data set and the four intact different resolution structures in the second data set, two correlation maps, CCF and MCF, were calculated using a prototypical helix 15 Å in length and an angular search stepsize of five degrees in *helixhunter*. Both the correlation maps and the original density were then further analyzed for helical content with *helixhunter*'s feature extraction routine, with a minimum helical length of 7 Å. The ''percen'' option was used to define the appropriate threshold for helix identification. Additionally, the ''sse'' option was enabled, generating a list of helices in the *DejaVu* SSE file format.[18]

The generated SSE files from the four MCF filtered *helixhunter* results were then used as input into *DejaVu*, in conjunction with the new helix library in order to identify potential structural homology. *DejaVu* was run with the ''bones'' search option using the default values for the most part, however since no directionality of the helices is assigned in *helixhunter*, no directionality was specified in the *DejaVu* options.

Each of the four intact structures was correlated with the corresponding sub-structures as probes in *foldhunter*. An initial search was done using a stepsize of 30 degrees, which was then systematically refined using the ''smart'' option. The ''smart'' option allows for a systematic refinement of fold orientation using increasingly finer angular searches. For the second data set, the various resolution 1C3W structures were probed against the corresponding sub-structure using the aforementioned parameters. The final data set, under the same parameters, used the beta sheet trimer of VP7, 1AHS, and 2TNF to probe the intact BTV VP7 trimer in *foldhunter* for the highest correlation fitting. Additional data sets (not shown, available at http://ncmi.bcm.tmc.edu/ ~ wjiang/hhfh) were analyzed, including structures across SCOP folds, superfamilies, families, proteins and species.

Visualization

While default output of the helices is done in the Open Inventor format, which can be viewed in the IRIS Explorer package, we have provided an additional output in the form of the popular 3-D virtual reality markup language (VRML). To visualize an atomic structure model, the PDB files were first rendered in *Ribbons*[34] and subsequently exported as Open Inventor files. The original density, correlation maps, the ribbon diagram and the *foldhunter/helixhunter* results were then visualized in IRIS Explorer on SGI Onyx dual-processor system.

Software availability

We have integrated both *helixhunter* and *foldhunter* into the newly published single particle image processing package, EMAN, which contains additional file handling features. The EMAN package[33] is freely available at http://ncmi.bcm.tmc.edu/~stevel/EMAN/doc.

Acknowledgments

We thank Jon A. Christopher for providing the source code for the helical axis fitting and Michael F. Schmid for his comments. This research has been supported by grants from National Institutes of Health (P41RR02250, R01AI43656, R01AI38469), Robert Welch Foundation and the National Partnership for Advanced Computational Infrastructure (NPACI), through the support of the National Science Foundation. M.L.B. was supported in part by BRASS and the W.M. Keck Center for Computational Biology through a training grant from the National Library of Medicine (2T15LM07093).

References

1. Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallog. sect. D,* **54**, 1078-1084.
2. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595-603.
3. Burley, S. K. (2000). An overview of structural genomics. *Nature Struct. Biol.* **7**, 932-934.
4. Böttcher, B., Wynne, S. A. & Crowther, R. A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature,* **386**, 88-91.
5. Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J. & Steven, A. C. (1997). Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature,* **386**, 91-94.
6. Ban, N., Freeborn, B., Nissen, P., Penczek, P., Grassucci, R. A., Sweet, R., Frank, J., Moore, P. B. & Steitz, T. A. (1998). A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. *Cell,* **93**, 1105-1115.
7. Matadeen, R., Patwardhan, A., Gowen, B., Orlova, B. V., Pape, T., Cuff, M., Mueller, F., Brimacombe, R. & van Heel, M. (1999). The *Escherichia coli* large ribosomal subunit at 7.5 Å resolution. *Structure Fold. Des.* **7**, 1575-1583.
8. Unger, V. M., Kumar, N. M., Gilula, N. B. & Yeager, M. (1999). Three-dimensional structure of a recombinant gap junction membrane channel. *Science,* **283**, 1176-1180.
9. Mancini, E. J., Clarke, M., Gowen, B. E., Rutten, T. & Fuller, S. D. (2000). Cryo-electron microscopy reveals the functional organization of an enveloped virus, Semliki Forest virus. *Mol. Cell,* **5**, 255-266.
10. Zhou, Z. H., Dougherty, M., Jakana, J., He, J., Rixon, F. J. & Chiu, W. (2000). Seeing the herpesvirus capsid at 8.5 Å. *Science,* **288**, 877-880.
11. Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257-259.

12. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. (1999). Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291**, 899-911.
13. Banner, D. W., Bloomer, A., Petsko, G. A., Phillips, D. C. & Wilson, I. A. (1976). Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochem. Biophys. Res. Commun.* **72**, 146-155.
14. Hubbard, S. R., Wei, L., Ellis, L. & Hendrickson, W. A. (1994). Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature,* **372**, 746-754.
15. Grimes, J. M., Basak, A. K., Roy, P. & Stuart, D. I. (1995). The crystal structure of bluetongue virus VP7. *Nature,* **373**, 167-170.
16. Grimes, J. M., Burroughs, J. N., Gouet, P., Diprose, J. M., Malby, R., Zientara, S., Mertens, P. P. C. & Stuart, D. O. (1998). The atomic structure of the bluetongue virus core. *Nature,* **395**, 470-477.
17. Gabashvili, I. S., Agrawal, R. K., Spahn, C. M. T., Grassucci, R. A., Svergun, D. I., Frank, J. & Penczek, P. (2000). Solution structure of the *E. coli* 70 S ribosome at 11.5 Å resolution. *Cell,* **100**, 537-549.
18. Kleywegt, G. J. & Jones, T. A. (1997) Detecting folding motifs and similarities in protein structures. In *Methods in Enzymol.* (Carter, C. W., Jr & Sweet, R. M., eds), vol. 277, pp. 525-545, Academic Press, London, England.
19. Mizuguchi, K. & Go, N. (1995). Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.* **8**, 353-362.
20. Kinoshita, K., Kidera, A. & Go, N. (1999). Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements. *Protein Sci.* **8**, 1210-1217.
21. Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955.
22. Basak, A. K., Gouet, P., Grimes, J., Roy, P. & Stuart, D. (1996). Crystal structure of the top domain of African horse sickness virus VP7: comparisons with bluetongue virus VP7. *J. Virol.* **70**, 3797-3806.
23. Baeyens, K. J., De Bondt, H. L., Raeymaekers, A., Fiers, W. & De Ranter, C. J. (1999). The structure of mouse tumour-necrosis factor at 1.4 Å resolution: towards modulation of its selectivity and trimerization. *Acta Crystallog. sect. D,* **55**, 772-778.
24. Lu, G., Zhou, Z. H., Baker, M. L., Jakana, J., Cai, D., Wei, X., Chen, S., Gu, X. & Chiu, W. (1998). Structure of double-shelled rice dwarf virus. *J. Virol.* **72**, 8541-8549.
25. DeRosier, D. J. & Harrison, S. C. (1997). Macromolecular assemblages. Sizing things up. *Curr. Opin. Struct. Biol.* **7**, 237-238.
26. Kirkitadze, M. D., Barlow, P. N., Price, N. C., Kelly, S. M., Boutell, C. J., Rixon, F. J. & McClelland, D. A. (1998). The herpes simplex virus triplex protein, VP23, exists as a molten globule. *J. Virol.* **72**, 10066-10072.
27. Wriggers, W., Milligan, R. A. & McCammon, J. A. (1999). Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125**, 185-195.
28. Baker, M. L., Seryskeva, I. I., Wu, Y., Sencer, S., Tung, W., Pate, P., Zhang, J. Z., Ludtke, S., Jiang, W., Chiu, W. & Hamilton, S. (2001). Identification of an N-terminal redox sensor in RYR1. *Biophysical J.* **80**, 330a-331a.
29. van Heel, M., Schatz, M. & Orlova, E. V. (1992). Correlation functions revisited. *Ultramicroscopy,* **46**, 307-316.
30. Gonzalez, R. C. & Woods, R. C. (1992). *Digital Image Processing*, Addison-Wesley, Reading, Mass.
31. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1997). *Numerical Recipes in C*, Cambridge University Press, New York, NY.
32. Christopher, J. A., Swanson, R. & Baldwin, T. O. (1996). Algorithms for finding the axis of a helix: fast rotational and parametric least-squares methods. *Comput. Chem.* **20**, 339-345.
33. Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). EMAN: semi-automated software for high resolution single particle reconstructions. *J. Struct. Biol.* **128**, 82-97.
34. Carson, M. (1997). Ribbons. *Methods Enzymol.* 493-505.

Edited by W. Baumeister

(Received 5 December 2000; received in revised form 26 February 2001; accepted 14 March 2001)