Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

EMatch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies

Oranit Dror,[a]* Keren Lasker,[a] Ruth Nussinov[b,c]* and Haim Wolfson[a]

[a]School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [b]Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel, and [c]Basic Research Program, SAIC-Frederick, Center for Cancer Research Nanobiology Program, NCI-Frederick, Building 469, Room 151, Frederick, MD 21702 USA

Correspondence e-mail: oranit@post.tau.ac.il, ruthn@ncifcrf.gov

Received 9 February 2006
Accepted 8 October 2006

Structural analysis of biological machines is essential for inferring their function and mechanism. Nevertheless, owing to their large size and instability, deciphering the atomic structure of macromolecular assemblies is still considered as a challenging task that cannot keep up with the rapid advances in the protein-identification process. In contrast, structural data at lower resolution is becoming more and more available owing to recent advances in cryo-electron microscopy (cryo-EM) techniques. Once a cryo-EM map is acquired, one of the basic questions asked is what are the folds of the components in the assembly and what is their configuration. Here, a novel knowledge-based computational method, named *EMatch*, towards tackling this task for cryo-EM maps at 6–10 Å resolution is presented. The method recognizes and locates possible atomic resolution structural homologues of protein domains in the assembly. The strengths of *EMatch* are demonstrated on a cryo-EM map of native GroEL at 6 Å resolution.

1. Introduction

Key cellular mechanisms are carried out through the formation of large macromolecular assemblies. Understanding the three-dimensional structure of these biological machines is essential for comprehension of their function (Alberts, 1998). Nevertheless, owing to their large size and instability, the structures of only a small number of macromolecular complexes have successfully been determined at atomic resolution, comprising a tiny portion of the PDB (Dutta & Berman, 2005; Krogan *et al.*, 2006).

Cryo-electron microscopy (cryo-EM) is a term referring to several different approaches to freezing a sample and reconstructing its three-dimensional structure from a set of two-dimensional projections. Recently, cryo-EM has emerged as a principal tool for structural analysis of macromolecular assemblies that are too large and flexible to be solved at atomic (high) resolution by NMR or X-ray crystallography (Baumeister & Steven, 2000; Frank, 2002; Chiu *et al.*, 2005). The obtained structural information is a three-dimensional grid, called a cryo-EM map, in which each voxel is associated with a mass-density value. The resolution of the map is in the range 6–30 Å. At low resolution (coarser than 15 Å), only the global shape and boundaries of some components are apparent. At intermediate resolution (6–15 Å), individual components can be discriminated. In particular, at 6–10 Å it is possible to reveal secondary-structure elements (helices or β-sheets).

The desire to bridge the resolution gap has stimulated the development of various *in silico* tools for combining intermediate- to low-resolution cryo-EM maps of multi-molecular complexes with atomic resolution data on molecular subunits.

The goal is to assist in providing quasi-atomic structural models of large assemblies. The tools can be classified into two types: (i) those that assume that the atomic structures of the subunits of the complex are known *a priori* and the goal is to find their locations and orientations (and sometimes their conformations) in the complex and (ii) those that do not assume this, but look for closely related known atomic structures of the subunits of the complex and fit them into the map.

Tools of the first type take as input a cryo-EM map of a complex and an experimentally determined atomic resolution structure of one of its subunits. They try to fit the subunit into the cryo-EM map. Some fitting approaches rely on manual placement with the aid of visualization tools, such as *O* (Jones *et al.*, 1991), *VMD* (Humphrey *et al.*, 1996), *Chimera* (Pettersen *et al.*, 2004) and *Amira* (http://www.amiravis.com). However, since these approaches depend on the decisions of an expert, other automated fitting methods have been developed with the aim of replacing the manual approaches. Most of them perform a search over three translational and three rotational degrees of freedom, locating the subunit in the cryo-EM map using different sampling algorithms and variants of cross-correlation as a similarity measure. These methods include *COAN* (Volkmann & Hanein, 1999, 2003), *DOCKEM* (Roseman, 2000), *EMfit* (Rossmann, 2000; Rossmann *et al.*, 2001), *Foldhunter* (Jiang *et al.*, 2001), *CoLoRes* from *Situs* (Chacon & Wriggers, 2002), 3*SOM* (Ceulemans & Russell, 2004) and *Mod-EM* (Topf *et al.*, 2005). In addition, there are a few methods that allow some degree of conformational flexibility of the subunits in the complex, such as *Situs* flexible fitting (Wriggers *et al.*, 2004) and *NMFF* (Tama *et al.*, 2004).

Frequently, the structures of individual subunits in a complex under inspection are unknown, but a cryo-EM map of their assembly is available. In such cases, only the second type of method, looking for closely related atomic structures of the subunits, is applicable. Given a cryo-EM map of a multi-domain protein, *SPI-EM* (Velazquez-Muriel *et al.*, 2005) is a statistical approach for determining the CATH superfamilies to which the domains of the protein belong. Firstly, a fitting method, such as *CoLoRes*, is applied to dock all CATH superfamily members into the map. The superfamilies that achieve the most statistically significant correlation scores are then returned. *Moulder-EM* (Topf *et al.*, 2006) is a method for modelling a target sequence of a single protein in the context of its cryo-EM map. The method iteratively refines comparative models. The models are generated by applying *MODELLER* (Sali & Blundell, 1993) and are refined based on the cryo-EM map of the target structure by applying *Mod-EM* (Topf *et al.*, 2005).

For cryo-EM maps at 6–10 Å resolution, a different strategy has been proposed (Jiang *et al.*, 2001; Chiu *et al.*, 2005). This strategy utilizes the observations that at this resolution it is possible to recognize secondary-structure elements (SSEs) and their spatial arrangement defines the scaffold of the examined structure. Firstly, SSEs are identified by methods such as *Helixhunter* (Jiang *et al.*, 2001) and *Sheetminer* (Kong & Ma, 2003; Kong *et al.*, 2004). Their three-dimensional configuration is then used to probe a library of known atomic resolution protein structures to find potential structural homologues. The strategy has been tested by applying *Helixhunter* to recognize helices, the *DEJAVU* (Kleywegt & Jones, 1997) or *COSEC* (Mizuguchi & Go, 1995) methods to find high-resolution structures with similar helix configuration and *Foldhunter* (Jiang *et al.*, 2001) to fit these structures into the cryo-EM map.

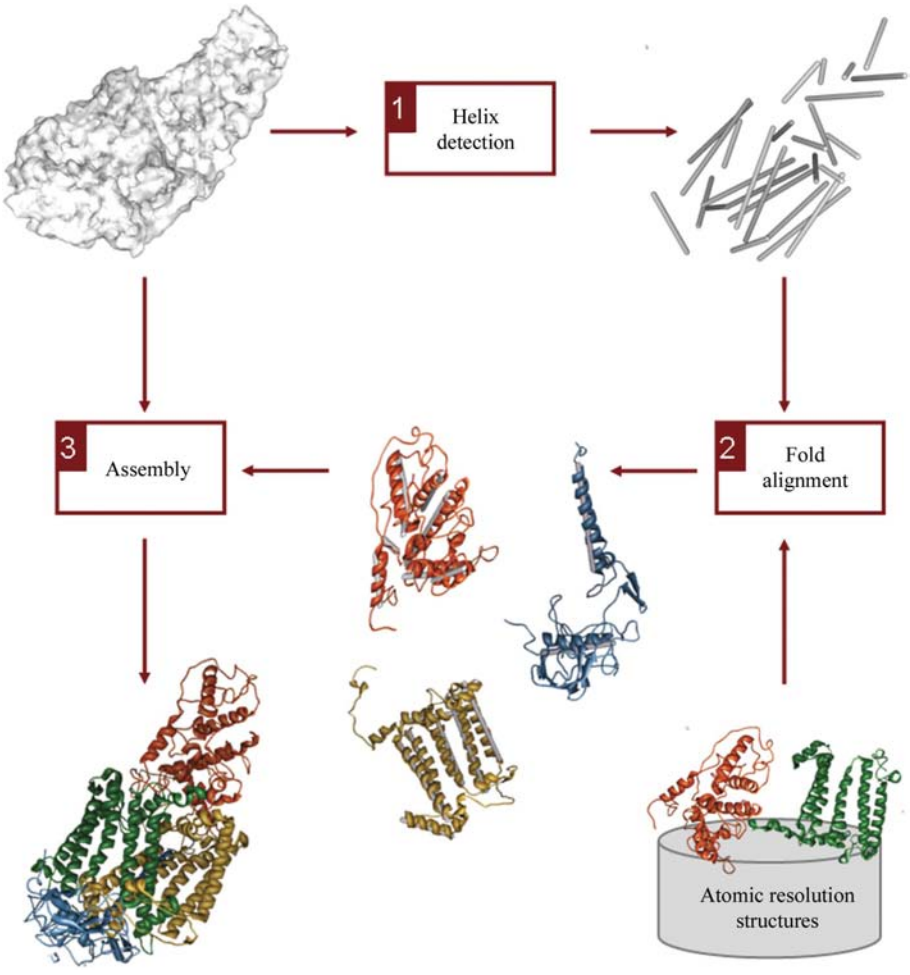

Figure 1
EMatch flow. The strategy of *EMatch* consists of three stages. In the first stage, helices are identified in a given cryo-EM map of a protein complex. Their spatial arrangement is then used to query a data set of atomic resolution folds to find potential structural homologues of domains appearing in the map. In the final stage, which is currently under development, the potential atomic structural homologues of the domains are assembled into a quasi-model of the complex.

Here, we describe a new computational knowledge-based method, named *EMatch*, aimed at detecting a quasi-atomic structural model of a protein assembly for which a cryo-EM map at 6–10 Å resolution is available. Similar to the strategy suggested by Jiang *et al.* (2001), *EMatch* is a three-tier algorithm (see Fig. 1). Firstly, helices are identified in the given cryo-EM map. Their spatial arrangement is then used to query a data set of atomic resolution protein folds to find potential structural homologues of domains appearing in the map and their locations in the complex. The aim of the final stage, which is currently under development, is to assemble the potential atomic structural homologues of the domains into a quasi-model of the complex. An important novel contribution of the method is its ability to identify 'partial alignments' between the detected set of helices and the data-set folds. The method is capable of aligning structural homologous folds even if (i) only some of the helices of the folds are matched with helices in the cryo-EM map and/or (ii) the matched helices are not necessarily of exact length and orientation. Thus, the method is tolerant to noise in the cryo-EM map and capable of aligning structures that are not fully homologous to domains in the complex (for example, sequentially remote domains of the same fold). Another important strength of the method is its high efficiency, which makes the method applicable to both interpreting large complexes and querying a massive data set of possible folds.

2. Method

Here, we give an outline of the algorithm. A more detailed technical description can be found in Lasker *et al.* (2005).

2.1. Helix detection

We seek to detect all the helices appearing in a given cryo-EM map at 6–10 Å resolution. To attain this goal, we exploit the observation that at this resolution helices appear as continuous, long, thin and highly dense cylindrical regions (Jiang *et al.*, 2001). Our aim is to find regions of voxels in the cryo-EM map that are most likely to be associated with a helix based on these unique characteristics.

The algorithm consists of four stages. In the first stage, we enhance voxels that are likely to be part of a helix and suppress the others by thresholding and fitting techniques. The objective of the second stage is to calculate an initial satisfactory segmentation of the map into regions such that each region satisfies a cylinder predicate. The predicate is defined in such a way that voxels of the same helix are likely to be clustered into the same region and each of the remaining voxels is considered as a separate region. The quality of a segmentation is usually quantified by two contradicting measurements; namely, (i) homogeneity, which is the similarity between voxels in the same region, and (ii) separability, which is the dissimilarity between voxels in different regions. We find a satisfactory segmentation as defined by Felzenszwalb & Huttenlocher (2004), which tries to balance the two measurements. In the next stage, we link noncontiguous regions that are likely to be part of the same helix based on geometrical considerations. Finally, we select those regions that are most likely to be associated with a helix and represent each one of them as an undirected segment. The direction of the segment is parallel to the eigenvector that corresponds to the largest eigenvalue of the covariance matrix of the locations of the voxels in the region. The end points of the segment are determined by projecting each of the voxels in the region onto its direction and selecting the extreme projected points.

2.2. Fold alignment

The fold-alignment algorithm is partially based on the *MASS* method for aligning multiple three-dimensional structures of proteins using their secondary-structure elements (SSEs; Dror *et al.*, 2003*a*,*b*). However, while *MASS* aligns high-resolution protein structures by also utilizing their atomic information, the fold-alignment algorithm is suitable for aligning structures for which the only available information is a coarse representation of their SSEs. The input for the fold-alignment stage is a cryo-EM map and a set of undirected three-dimensional line segments representing the central axes of SSEs appearing in the map (in the current application, only helices are used). The goal is to fit all atomic resolution protein folds from a predefined data set into the given cryo-EM map based on the spatial configuration of their SSEs.

The rationale behind the method is that a biologically interesting common substructure consists of at least two SSEs. Thus, ordered pairs of nonlinear SSE segments, which we call bases, are used to fit each data-set structure into the cryo-EM map. Given a data-set structure, the method examines whether some of its bases share a similar three-dimensional configuration with bases in the input set of SSE segments. For each such pair of bases with a similar three-dimensional configuration, the method computes two possible transformations for superimposing one base onto another. Each transformation defines an initial alignment between the cryo-EM map and the data-set structure for which at least two SSEs are matched. In the next stage, the initial alignments are clustered and extended by finding additional matched SSE segments in the two structures (two SSE segments are matched if their line distance, midpoint distance and angle are below predefined thresholds). The extended alignments are then clustered and sorted by their core size and the r.m.s.d. (Kaindl & Steipe, 1997) between the midpoints of the corresponding segments. Finally, the top-ranking alignments (ten by default) are re-ranked by their correlation score (defined as the normalized cross-correlation coefficient) with the cryo-EM.

3. Results and discussion

We have successfully validated *EMatch* on a number of simulated cryo-EM maps (Lasker *et al.*, 2005). Here, we evaluate the method on an experimental cryo-EM map of native GroEL. GroEL is a chaperone that assists protein folding in prokaryotes. Its three-dimensional structure is highly symmetric, comprising 14 monomers that are arranged

Table 1
A priori known domain reconstruction.

For each domain, the data appearing in the columns are the domain name, the number of matched helices of the top-ranking alignment between the high-resolution domain and the cryo-EM map, the r.m.s.d. between the axial midpoints of the matched helices, the average angle and average line distance between the matched helices, the *Z* score of the top-ranking alignment, the running time of the fold-alignment stage and the r.m.s.d. between the domain ring in the suggested atomic quasi-structural model and the corresponding domain ring in the X-ray crystal structure of the complex (PDB code 1oel) after superimposing the two structures with minimum r.m.s.d.

	Matched helices						
Domain	No.	R.m.s.d. (Å)	Avg. angle (°)	Avg. line distance (Å)	*Z* score	Run time (s)	Evaluation r.m.s.d. (Å)
Equatorial	6	3.40	12.23	1.03	2.96	27	3.72
Apical	3	0.68	14.02	0.30	1.80	12	6.17
Intermediate	4	3.40	20.33	2.97	2.73	9	6.28

in two back-to-back heptameric rings. Each monomer consists of three domains: the equatorial (E), apical (A) and intermediate (I) domains. The input given to *EMatch* is a single ring of *Escherichia coli* GroEL at 6 Å resolution taken from the EMD database (EMDB; map code EMD1081; Ludtke *et al.*, 2004). This complex possesses global cyclic symmetry (Goodsell & Olson, 2000) with seven monomers, which is noted as C_7. In the first stage, *EMatch* identified a set of 168 helices in the input cryo-EM map (hereafter referred to as EM helices) in approximately 50 min. Two types of experiments were then carried out to evaluate the method in cases where (i) the atomic structures of the domains are known in advance and (ii) the atomic structures of the domains are *a priori* unavailable, but closely related structures exist in a predefined data set of high-resolution structures. The experiments were performed on a standard PC (Pentium 4, 2.60 GHz with 2 GB RAM). We present the obtained results below.

3.1. *A priori* known domain reconstruction

The goal of this experiment is to suggest a quasi-atomic structural model of the GroEL complex given its C_7 global symmetry and the atomic structures of the three domains of its monomer [taken from PDB code 1oel (Braig *et al.*, 1995)]. The experiment consists of two stages: fold alignment and assembly. In the first stage, the spatial arrangement of the EM helices is used to locate the three given domains in the complex. In the second stage, the global symmetry of the complex is imposed to assemble the domains and in this way to construct a quasi-atomic model of the whole complex.

3.1.1. Fold alignment. Each of the three input domains was aligned with the set of detected EM helices in less than 30 s (27, 12 and 9 s for the equatorial, apical and intermediate domains, respectively). Figs. 2(*a*), 2(*b*) and 2(*c*) present the matched helices of the top-ranking alignment for the three domains. The top-ranking alignment for the equatorial domain consists of six matched helices with an r.m.s.d. of 3.40 Å between their axial midpoints. For the apical domain, the top-ranking alignment consists of three matched helices with an r.m.s.d. of 0.68 Å between their axial midpoints. Finally, the top-ranking alignment for the intermediate domain has four matched helices with an r.m.s.d. of 3.40 Å between their axial midpoints. Further details of the alignments (including *Z* scores and additional data on the matched helices) are available in Table 1 and at the *EMatch* website (http://bioinfo3d.cs.tau.ac.il/EMatch).

3.1.2. Assembly. Imposing the C_7 global symmetry of the GroEL ring, a quasi-atomic structural model of its overall complex has been constructed. Fig. 2(*d*) shows the suggested model. The model has been successfully evaluated by comparing it with the X-ray crystal structure of one of the GroEL rings at 2.8 Å resolution (PDB code 1oel; Braig *et al.*, 1995). The two structures have been superimposed with a minimum r.m.s.d. of 5.17 Å between their corresponding C^{α} atoms by applying the least-squares fitting technique (Kabsch, 1978). Fig. 3(*a*) shows the superposition. The r.m.s.d. between the corresponding equatorial, apical and intermediate domain rings in this superposition are 3.72, 6.17 and 6.28 Å, respectively (see Figs. 3*b*, 3*c* and 3*d*). The differences between the structures are related to structural flexibility, especially in the intermediate and apical domains, as has been observed by Ludtke *et al.* (2004).

3.1.3. Helix-detection evaluation. We have used the obtained quasi-atomic model of the complex to evaluate the performance of the helix-detection stage. For each detected EM helix, we have searched for the closest helix in the model and calculated the angle, midpoint distance and line distance between them. The thresholds for a correctly detected helix are 40° for the angle, 13 Å (approximately twice the resolution) for the midpoint distance and 4 Å for the line distance. Details for each domain are found in Table 2. In addition, we have quantified the results using the sensitivity and specificity measurements. The true (false) positive rate is defined as the number of detected EM helices that match (do not match) a helix in the quasi-atomic model of the complex. The true (false) negative rate is defined as the number of strands and loops in the quasi-model that do not match (match) an EM helix. Based on these definitions, the sensitivity and specificity of the helix-detection stage are 75% and 65%, respectively. We intend to improve these ratios by applying local refinement to each helix based on the EM density.

3.2. *A priori* unknown domain reconstruction

This experiment is a generalization of the previous one. The goal is to construct a quasi-atomic structural model of the complex in the case where the high-resolution structures of the three domains are unavailable, but structural homologues exist in a predefined data set. To achieve this goal, we need to answer two questions: (i) which are the structural homologues of the domains of the complex and (ii) what are their alignments with the cryo-EM map of the complex? For a structural homologue of a domain appearing in the complex, *EMatch* is able to identify a superimposition into the map that is consistent with the localization of the domain of the complex in

Table 2
Helix-detection evaluation.

For each domain, the data appearing in the columns are the domain name, the number of helices detected by *EMatch* out of the total number of helices in the domain and the average midpoint distance, angle and line distance between the matched helices.

Domain	Identified helices	Avg. midpoint distance (Å)	Avg. angle (°)	Avg. line distance (Å)
Equatorial	9/10	7.59	15.09	2.06
Apical	4/5	4.53	10.65	1.37
Intermediate	3/4	6.44	18.18	1.90

the map. However, answering the first question is still not fully supported and ongoing work deals with this challenge.

3.2.1. Fold alignment. The data set used in this experiment contains 1538 atomic structures of protein domains representing all superfamilies of the seven true classes in SCOP: all-α, all-β, α/β, $\alpha + \beta$, α and β, small proteins and membrane and cell-surface proteins (Murzin *et al.*, 1995). Each candidate domain in the data set has been aligned into the cryo-EM map by *EMatch*. Notwithstanding the large number of identified EM helices and the size of the data set, the whole process took less than 5 h. The result is the top-ranking alignment with the cryo-EM map for each domain in the data set.

One of the criteria for the success of this stage is that given a SCOP representative that is structurally homologous to some domain of the complex in the cryo-EM map, its superimposition into the map, as defined by the top-ranking alignment, should be consistent with the localization of the domain of the complex in the map. We have therefore evaluated the top-ranking alignment obtained for each of the three SCOP superfamily representatives of the GroEL domains. These are SCOP:19490 (PDB code 1a6e, chain *A*, residues 17–145 and 404–519) for the equatorial domain, SCOP:109289 (PDB code 1we3, chain *A*, residues 190–373) for the apical domain and SCOP:66226 (PDB code 1iok, chain *A*, residues 137–190 and 367–409) for the intermediate domain. The quasi-atomic structural model revealed in the previous experiment has been used for the evaluation process. The reason that we have decided to use this model instead of the X-ray crystal structure is the difference between the X-ray structure and the cryo-EM

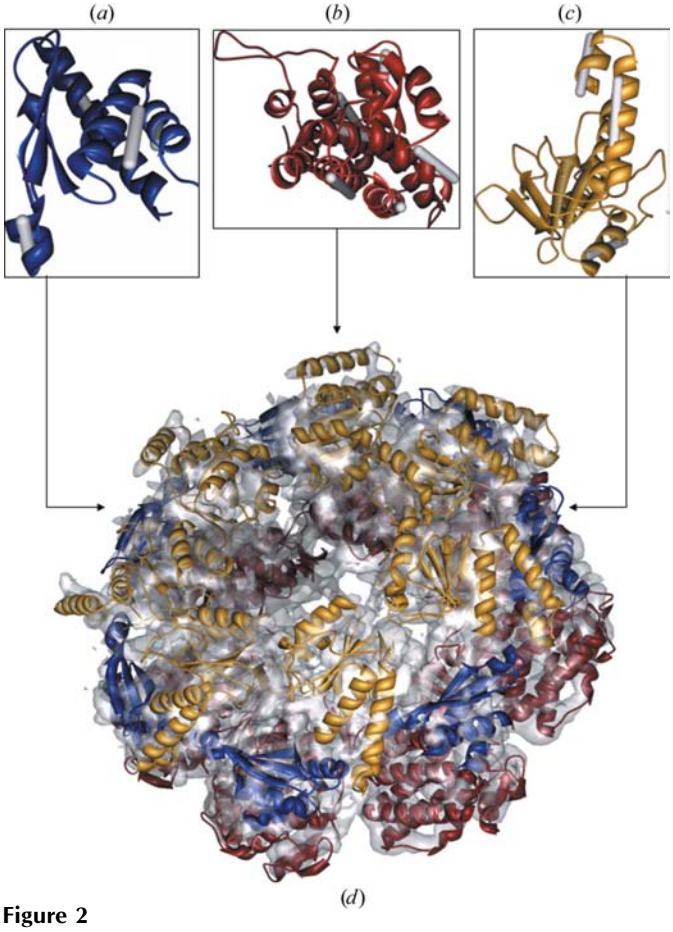

Figure 2
A priori known domain reconstruction. (*a*–*c*) The matched helices of the top-ranking alignment for the intermediate (blue), equatorial (red) and apical (yellow) domains, respectively. (*d*) A quasi-atomic structural model of a GroEL ring as revealed from the cryo-EM map (depicted in grey). This figure and subsequent figures were prepared using *Chimera* (Pettersen *et al.*, 2004).

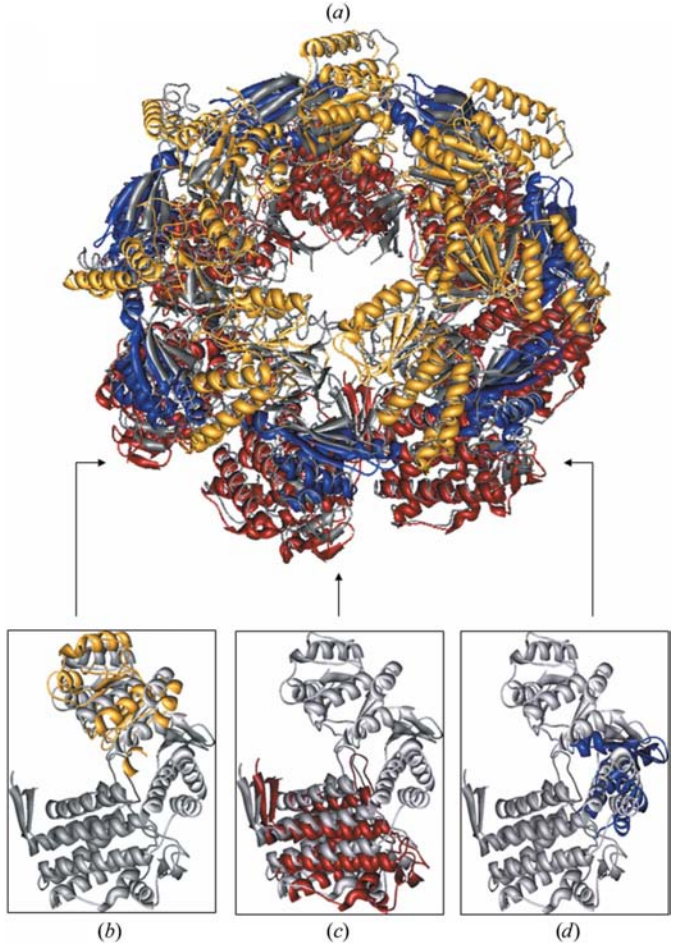

Figure 3
Evaluation of *a priori* known domain reconstruction. (*a*) A quasi-atomic structural model of a GroEL ring (coloured as in Fig. 2) superimposed on its X-ray crystal structure (PDB code 1oel; grey) with a minimum r.m.s.d. of 5.17 Å. (*b*–*d*) Enlargement of the superimposed apical, equatorial and intermediate domains, respectively.

map as observed by Ludtke *et al.* (2004). The following procedure was applied for each of the three GroEL SCOP representatives.

(i) Let R be a GroEL SCOP representative and let T be the transformation to align R onto the GroEL cryo-EM map as defined by the top-ranking alignment of *EMatch*.

(ii) Transform R onto the cryo-EM map by applying T and then impose the C_7 symmetry of the GroEL ring on $T(R)$. Let $C_7\ T(R)$ be the result.

(iii) Use *MASS* (Dror *et al.*, 2003*a*,*b*) to align R onto the quasi-atomic structural model revealed in the previous experiment and then impose the C_7 symmetry of the GroEL ring on the result. Denote the obtained structure $C_7\ T'(R)$.

(iv) Compute the r.m.s.d. between the C^{α} atoms of $C_7\ T(R)$ and $C_7\ T'(R)$. In the following, we refer to this r.m.s.d. as the evaluation r.m.s.d.

The top-ranking alignment for the SCOP superfamily representative of the equatorial domain contains six matched helices with an r.m.s.d. of 3.50 Å between their axial midpoints. For the superfamily representative of the apical domain, the top-ranking alignment consists of three matched helices with an r.m.s.d. of 1.20 Å between their axial midpoints. Finally, the top-ranking alignment for the superfamily representative of the intermediate domain contains three matched helices with an r.m.s.d. of 3.20 Å between their axial midpoints. For all the three superfamily representatives, the top-ranking alignment has achieved an evaluation r.m.s.d. lower than 7 Å, namely 1.07, 6.33 and 6.65 Å for the equatorial, apical and intermediate domains, respectively. The evaluation r.m.s.d. for the whole constructed complex is 4.76 Å. Further details of the alignments (including Z scores and additional data on the matched helices) are available in Table 3, Fig. 4 and the website. The results clearly demonstrate the potential of *EMatch* to detect alignments that are almost as accurate as atomic based ones. This success in bridging the resolution gap is achieved by the capability of *EMatch* to extract sufficient secondary-structure information from the cryo-EM maps and to find partial alignments between the SSEs and the high-resolution structures.

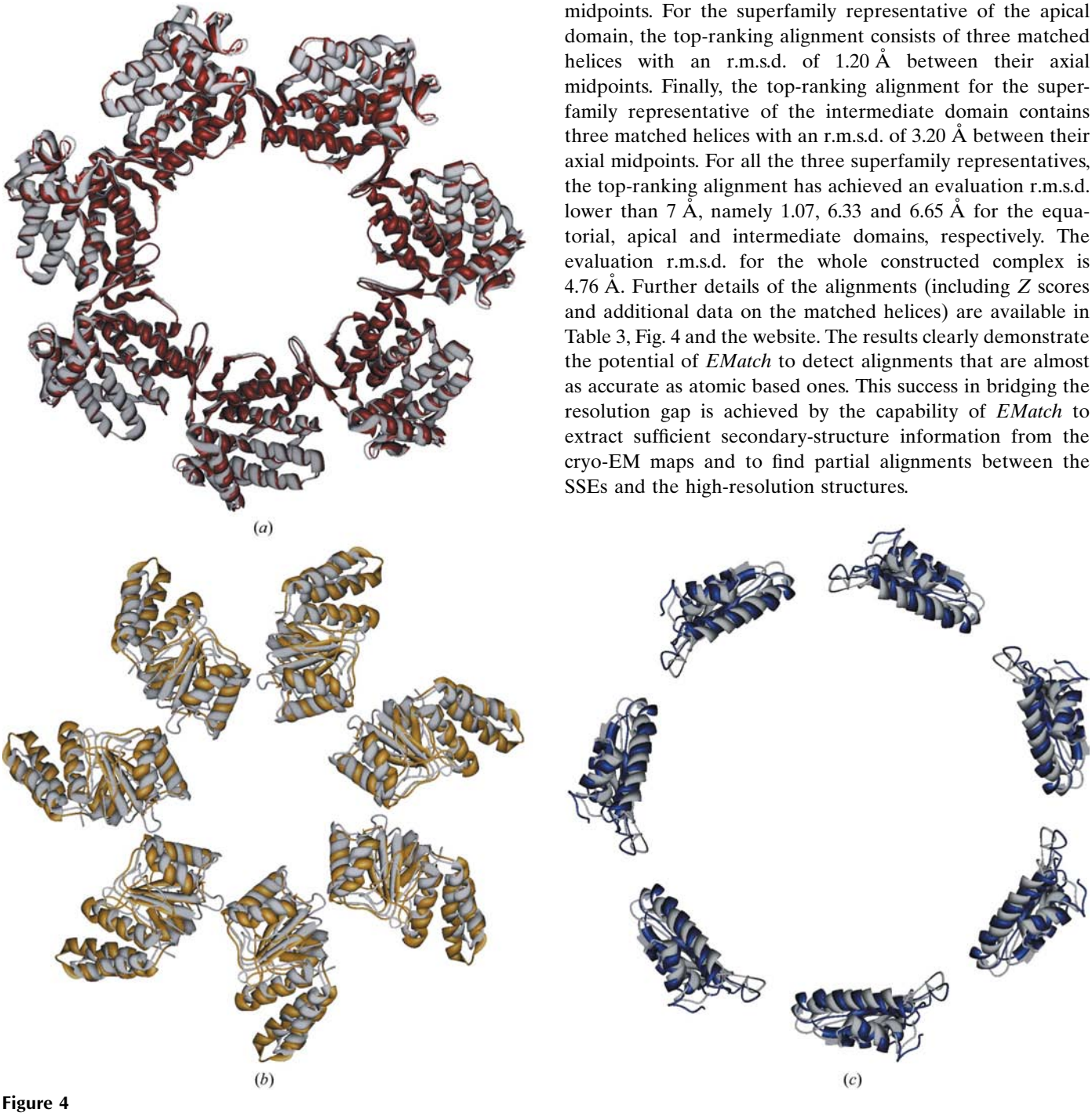

Figure 4
Evaluation of *a priori* unknown domain reconstruction. (*a*) The SCOP superfamily representative for the equatorial domain (red) superimposed by *EMatch* on the cryo-EM map (not shown) and the same structure (grey) superimposed by *MASS* on the atomic quasi-structural model constructed in the first experiment. (*b*) and (*c*) Similar figures for the apical (yellow) and intermediate (blue) domains, respectively.

Table 3
A priori unknown domain reconstruction.

For each GroEL domain, the top-ranking alignment between its SCOP superfamily representative and the cryo-EM map is presented. The data appearing in the columns are the domain name, the SCOP code of the superfamily representative, the number of matched helices, the r.m.s.d. between the axial midpoints of the matched helices, the average angle and average line distance between the matched helices, the *Z* score of the alignment, the running time of the fold-alignment stage and the evaluation r.m.s.d. as defined in the text.

		Matched helices						
Domain	SCOP code	No.	R.m.s.d. (Å)	Avg. angle (°)	Avg. line distance (Å)	*Z* score	Run time (s)	Evaluation r.m.s.d. (Å)
Equatorial	19490	6	3.50	10.64	1.03	2.54	23	1.07
Apical	109289	3	1.20	13.17	1.41	1.81	15	6.33
Intermediate	66226	3	3.20	22.63	0.90	3.43	7	6.65

3.2.2. Assembly (future work). The challenge that we face is to find the SCOP superfamily representatives for which structural homologues appear in the cryo-EM map. To date, this task is only partially addressed by *EMatch*. Particularly for GroEL, when we ranked all the SCOP representatives by their correlation scores, the SCOP representatives of the apical and intermediate domains were ranked in fourth and sixth places with respect to all SCOP representatives. The SCOP representative of the equatorial domain received a lower rank. The reason for this is that smaller domains have a higher chance of receiving a high correlation score. Ongoing work aims to provide a full solution to the assembly task by using additional constraints derived from the cryo-EM map, protein sequences and available high-resolution structures.

4. Conclusion

We have presented a novel highly efficient computational method, named *EMatch*, for aligning atomic resolution subunits into cryo-EM maps of large macromolecular assemblies at 6–10 Å resolution. The method identifies helices in an input cryo-EM map. It then uses the spatial arrangement of the helices to query a data set of high-resolution folds and finds structures that can be aligned into the cryo-EM map. *EMatch* has been successfully tested on simulated data (Lasker *et al.*, 2005). Here, we have described an example in which *EMatch* has been applied to experimental cryo-EM data of native GroEL at 6 Å resolution. The results show the ability of *EMatch* to identify helices with reasonably high specificity and sensitivity ratios, as well as its capability to align the correct folds into the input cryo-EM map even when the helical information is partial. The running times are immensely satisfying and demonstrate the high efficiency of the method; a typical analysis of a cryo-EM map with several monomers, such as GroEL, takes less than 50 min, and a successive search against a high-resolution structural data set of 1538 domains takes about 5 h on a standard desktop PC. Future work includes developing assembly algorithms that will include additional constraints, such as sequence homology, β-sheet positions, symmetry and other geometric constraints.

5. Availability

Supplementary information is available at the website http://bioinfo3d.cs.tau.ac.il/EMatch.

We thank Maxim Shatsky for stimulating discussions. This research was supported by the Binational Israel–USA Science Foundation (BSF). This research was also supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The research of OD was supported by the Eshkol Fellowship funded by the Israeli Ministry of Science. The research of HJW was supported in part by the Israel Science Foundation (grant No. 281/05) and Hermann Minkowski Minerva Center for Geometry at TAU. The research of RN was funded by Federal funds from the NCI, NIH under contract No. NO1-CO-12400. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organization imply endorsement by the US Government.

References

Alberts, B. (1998). *Cell*, **92**, 291–294.
Baumeister, W. & Steven, A. C. (2000). *Trends. Biochem. Sci.* **25**, 624–631.
Braig, K., Adams, P. D. & Brünger, A. T. (1995). *Nature Struct. Biol.* **2**, 1083–1094.
Ceulemans, H. & Russell, R. B. (2004). *J. Mol. Biol.* **338**, 783–793.
Chacon, P. & Wriggers, W. (2002). *J. Mol. Biol.* **317**, 375–384.
Chiu, W., Baker, M. L., Jiang, W., Dougherty, M. & Schmid, M. F. (2005). *Structure*, **13**, 363–372.
Dror, O., Benyamini, H., Nussinov, R. & Wolfson, H. (2003*a*). *Bioinformatics*, **19**, *Suppl.* 1, i95–i104.
Dror, O., Benyamini, H., Nussinov, R. & Wolfson, H. (2003*b*). *Protein Sci.* **12**, 2492–2507.
Dutta, S. & Berman, H. M. (2005). *Structure*, **13**, 381–388.
Felzenszwalb, P. F. & Huttenlocher, D. P. (2004). *Int. J. Comput. Vis.* **59**, 167–181.
Frank, J. (2002). *Annu. Rev. Biophys. Biomol. Struct.* **31**, 303–319.
Goodsell, D. S. & Olson, A. J. (2000). *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153.
Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.
Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). *J. Mol. Biol.* **308**, 1033–1044.
Jones, T., Zou, J., Cowan, S. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.
Kaindl, K. & Steipe, B. (1997). *Acta Cryst.* A**53**, 809.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 525–545.
Kong, Y. & Ma, J. (2003). *J. Mol. Biol.* **332**, 399–413.
Kong, Y., Zhang, X., Baker, T. S. & Ma, J. (2004). *J. Mol. Biol.* **339**, 117–130.
Krogan, N. J. *et al.* (2006). *Nature (London)*, **440**, 637–643.

Lasker, K., Dror, O., Nussinov, R. & Wolfson, H. J. (2005). *Algorithms in Bioinformatics, 5th International Workshop, WABI 2005*, edited by R. Casadio & G. Myers, pp. 423–434. Berlin: Springer.
Ludtke, S. J., Chen, D.-H., Song, J.-L., Chuang, D. T. & Chiu, W. (2004). *Structure*, **12**, 1129–1136.
Mizuguchi, K. & Go, N. (1995). *Protein Eng.* **8**, 353–362.
Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.
Roseman, A. M. (2000). *Acta Cryst.* D**56**, 1332–1340.
Rossmann, M. G. (2000). *Acta Cryst.* D**56**, 1341–1349.
Rossmann, M. G., Bernal, R. & Pletnev, S. V. (2001). *J. Struct. Biol.* **136**, 190–200.
Sali, A. & Blundell, T. L. (1993). *J. Mol. Biol.* **234**, 779–815.
Tama, F., Miyashita, O. & Brooks, C. L. III (2004). *J. Struct. Biol.* **147**, 315–326.
Topf, M., Baker, M. L., John, B., Chiu, W. & Sali, A. (2005). *J. Struct. Biol.* **149**, 191–203.
Topf, M., Baker, M. L., Renom, M. A. M., Chiu, W. & Sali, A. (2006). *J. Mol. Biol.* **357**, 1655–1668.
Velazquez-Muriel, J. A., Sorzano, C. O., Scheres, S. H. W. & Carazo, J.-M. (2005). *J. Mol. Biol.* **345**, 759–771.
Volkmann, N. & Hanein, D. (1999). *J. Struct. Biol.* **125**, 176–184.
Volkmann, N. & Hanein, D. (2003). *Methods Enzymol.* **374**, 204–225.
Wriggers, W., Chacon, P., Kovacsa, J. A., Tama, F. & Birmanns, S. (2004). *Neurocomputing*, **56**, 365–379.