FOLD-EM: Fully Automated Fold Recognition in Low Resolution Electron Density Maps

Mitul Saha[1,*] and Marc C. Morais[1,*]

[1]Sealy Center for Structural Biology and Molecular Biophysics,

Department of Biochemistry and Molecular Biology,

University of Texas Medical Branch, Galveston, TX 77555.

[*] **Corresponding authors:**

Mitul Saha, Ph.D.

Sealy Center for Structural Biology and Molecular Biophysics,

University of Texas Medical Branch

301 University Boulevard, Galveston, TX 77555-0647

Email: misaha@utmb.edu

Phone: 409-747-1406

Marc C. Morais, Ph.D.

Sealy Center for Structural Biology and Molecular Biophysics,

University of Texas Medical Branch

301 University Boulevard, Galveston, TX 77555-0647

Email: mcmorais@utmb.edu

Phone: 409-747-1401

Condensed Title: FOLD-EM: Fully Automated Fold Recognition

Number of Characters: 34543

ABSTRACT

Scale-invariant feature transform (SIFT) algorithms are used in the field of computer vision to detect and describe local features in images. They implement an approach that has many characteristics in common with neuronal responses in primate vision. Although SIFT-based approaches have been primarily implemented for analysis of two-dimensional images, they also have great potential in structural biology to facilitate the interpretation of noisy, cluttered, three-dimensional electron density maps of cells, organelles, and other large macromolecular assemblies. Here, we have adapted the SIFT algorithm to develop a new computational tool FOLD-EM that identifies folds, motifs, and other features of biological macromolecules in electron density maps and returns a model of the backbone of the constituent polypeptides. We further demonstrate the inherent ability of our algorithm to detect and characterize conformational changes in different biological states of a particular macromolecular complex. FOLD-EM is available to the structural biology community as a free open source software at: http://cs.stanford.edu/~mitul//foldEM/

INTRODUCTION

SUMMARY: Scale-invariant feature transform (SIFT) algorithms are used in the field of computer vision to detect and describe local features in images. Implementing an approach that has many characteristics in common with neuronal responses in primate vision, SIFT-based algorithms transform an image into a large collection of local feature vectors, each of which is invariant to any scaling, rotation or translation of the image. These collections of feature vectors provide a description that can then be used to locate similar objects in an image or collection of images. In addition to being invariant to object scaling and rotation, SIFT-based approaches are also resilient to the effects of clutter, partial occlusion, and "noise" present in images. Although SIFT-based approaches have been primarily implemented for analysis of two-dimensional images, they also have great potential in structural biology to facilitate the interpretation of noisy, cluttered, three-dimensional electron density maps of cells, organelles, and other large macromolecular assemblies. Here, we have adapted the SIFT algorithm to develop a new computational tool FOLD-EM that identifies folds, motifs, and other features of biological macromolecules in electron density maps and returns a model of the backbone of the constituent polypeptides. We further demonstrate the inherent ability of our algorithm to detect and characterize conformational changes in different biological states of a particular macromolecular complex. FOLD-EM is available to the structural biology community as a free open source software at: http://cs.stanford.edu/~mitul//foldEM/

Recent technological advances have resulted in exponential growth of the amount of data available at each level of the sequence-structure-function relationship. Along with this expansion of available data comes the need for systematic and objective methods to analyze and interpret this data. For example, the amount of information that can be extracted from the structure of an isolated macromolecule is limited; to fully understand how a macromolecule functions in a cell requires

knowledge of not only its interaction partners, but also how mutually induced conformational changes that occur upon complex formation give rise to integrated biological function. Toward this end, structural biology continues to tackle larger and larger targets, ranging from X-ray structures of binary protein complexes to cryo-electron microscopy image reconstructions of large macromolecular complexes and cryo-electron tomograms of entire cells. Unfortunately, cryo-EM/ET maps typically have low signal-to-noise ratios, making their analysis and interpretation challenging and somewhat subjective, depending on the skill of specialized investigators. Hence, there is a need for computational methods to systematically and quantitatively analyze maps of macromolecular assemblies, organelles and whole cells. In particular, tools capable of 1) identifying individual proteins within larger complexes and 2) characterizing conformational rearrangements relevant to macromolecular function would provide non-structural specialists access to structural data, thus allowing for enhanced biological perspectives.

We have developed a new software tool, FOLD-EM, to identify macromolecular folds in cryo-EM/ET electron density maps and to characterize conformational changes that accompany different biological states of macromolecules. Although inspired by SIFT (1, 2), FOLD-EM is not simply an extension of SIFT; it is also coupled with advanced computational techniques such as 6DOF clustering and graph analysis to carry out fold recognition and structure analysis in electron density maps. The method works by constructing rotationally invariant, low-dimensional representations of local regions in the input atomic resolution structures and cryo-EM maps. Correspondences are established between the reduced representations by comparing them using a simple metric. These correspondences are then clustered using hash tables and graph theory to identify structurally equivalent domains or motifs. Because correspondences are built from matching smaller structural units, the program will work even if only portions of domains are structurally homologous. Similarly, FOLD-EM will also automatically determine if different transformations are necessary for fitting

different regions of the input search model; as a result, FOLD-EM is inherently able to characterize conformational differences between the structures being compared. Here, we demonstrate the effectiveness of FOLD-EM using synthetic and real data, and discuss some advantages of FOLD-EM compared to existing software that fit atomic resolution structures into medium to low resolution electron density maps.

RESULTS

The FOLD-EM algorithm

FOLD-EM is based on our previous work (3), and solves the structural comparison problem **P,** defined as follows: compare a non-atomic resolution structure (i.e. a cryo-EM map from EMDB) with another structure, typically either another electron density map or atomic coordinates obtained via X-ray crystallography or NMR, and identify conserved structural domains or motifs between the pair of input structures. The full algorithm used by FOLD-EM to solve **P** is outlined in the Methods. Briefly, the input to FOLD-EM is a pair of volumetric electron density maps obtained either experimentally from cryo-EM image reconstructions or calculated from atomic coordinates. The program uses geometric processing, statistical analysis, and graph theory to detect conserved regions between the input pair by executing six steps. In laymen's terms, FOLD-EM first tries to find small regions of similarity between two maps, and then determines how each small region in one map must be rotated and translated to superimpose it on the equivalent region in the other map. Next, FOLD-EM groups regions that require similar rotations and translations, and if the members of a group also form contiguous regions in the maps then this region is considered a structurally equivalent domain/motif. Since FOLD-EM also gives the relative spatial orientation of the common sub-structures extracted, the fitted domain/motif is returned along with its corresponding transformation matrix. If no homologous structural motifs are identified between the input pair, then no fitted structure is returned. Hence, FOLD-EM only outputs fitted coordinates if meaningful structural homology can be detected. In this regard, FOLD-EM differs from other fitting programs which always return a fitted structure regardless of whether or not the determined fit is meaningful.

In essence, FOLD-EM identifies conserved domains/motifs in large macromolecular assemblies. Because domain/motif correspondences are built from matching smaller structural units, no prior knowledge of the extent of homologous domain/motif structures is required and the program will thus work even if only portions of domains are similar. In contrast, other fold recognition/fitting algorithms require that the structures being compared are similar over the entirety of the search structure. Similarly, FOLD-EM is inherently able to compare and fit structures that have undergone conformational changes; the 'bottom-up' approach of assembling structural correspondences in FOLD-EM assures that discrete conserved structural units are automatically identified and fitted separately, thus providing a computationally objective approach for performing flexible fitting. As a result, FOLD-EM automatically identifies domains/motifs in large macromolecular assemblies that remain conserved upon conformational rearrangement. As a by-product, non-conserved regions in structures are also revealed, which can point to potentially important molecular flexibility. Hence, FOLD-EM has the potential to facilitate biomedical research and discovery by accelerating the rate at which structures of large macromolecular assemblies can be determined and analyzed. Below, we present the performance of FOLD-EM for various structural feature recognition problems and discuss its advantages and limitations compared to existing docking softwares.

Evaluating FOLD-EM as a fitting/docking tool

To verify that FOLD-EM is capable of recognizing and fitting conserved structural motifs into cryo-EM maps, we tested the algorithm using simulated and real cryo-EM data. First, electron density maps, comparable to those obtained via cryo-EM, were calculated for a GroEL monomer in the 5 – 20 Å resolution range. To assess the effect of search model size on fitting, we split the GroEL monomer into 3 separate domains: 1) the equatorial domain (249 residues); 2) the apical domain (182 residues); and 3) the intermediate domain (90 residues) (Fig. 1(a)). Table S1 reports the results from

fitting each domain. The reported error values for fitting each domain into maps of different resolutions are low, demonstrating the effectiveness of FOLD-EM in accurately fitting domains/motifs of varying sizes into relatively low resolution maps. We have also tested the ability of FOLD-EM to fit atomic resolution structures into experimentally determined cryo-EM maps. For instance, Figs. 1 (b) , (c), & (d) show, evaluate, and compare (with other popular fitting software) the result of using FOLD-EM to fit known GroEL atomic resolution domains into a 6 Å experimentally determined cryo-EM map of GroEL. Figs. 1 (e) & (f) show the results of another test using real cryo-EM data. Here we used FOLD-EM to fit the HK97 and bacterial immunoglobulin (BIG2) domains of the capsid protein of bacteriophage phi29 (5) into their corresponding densities in the 7.9 Å reconstruction of phi29 particles (5). Fig. 1 (e) shows the fit of both the HK97 and BIG2 domains obtained using FOLD-EM in a fully automated fashion; these results agree well with previously reported results (5) obtained using semi-automated means. Fig. 1 (f) shows the failure of a popular fitting software in fitting the BIG2 domain.

Fitting in the presence of extraneous regions

A strength of the feature recognition algorithm used by FOLD-EM is its ability to carry out partial matching. That is, FOLD-EM can match and align two objects precisely even though they match only partially with each other. Figs. 2 (a) & (b) schematically illustrate the need for this. As seen in the figures, while trying to fit a structural homologue (black wire) into its corresponding region in a cryoEM map (blue region), presence of extraneous regions (red wire; which does not have any corresponding region in the target map) can act as noise and introduce fitting error. We will now show that FOLD-EM is indeed able to ignore any extraneous regions and preserve accuracy, while docking/fitting, both in simulated and real/experimental data. In simulated data, as earlier, we used FOLD-EM to fit the three domains of GroEL into simulated cryo-EM maps of GroEL. However, here, we added extraneous

structural features/residues to each domain of the search model (as seen in Figs. 2 (c) ,(d), & (e)) that are not present in the simulated maps. Tables S2 (a), (b), & (c) show the results obtained by using FOLD-EM to fit these altered high resolution domain structures, each with differing amounts of extraneous structures/residues introduced, into simulated cryoEM maps of GroEL. The low RMSD errors demonstrate the effectiveness of FOLD-EM in fitting structures in the presence of extraneous, non-homologous structural features. We also tested the partial matching capabilities of FOLD-EM by correctly fitting our structurally altered GroEL domains into an experimentally determined 6 Å cryo-EM map of GroEL (6). Figs. 2 (f) & (g) show some successful fittings done by FOLD-EM in the presence of extraneous non-homologous structure. The figures also show how another popular fitting program failed (due to their inability to handle extraneous residues) in these cases. We also wanted to test if FOLD-EM could successfully perform partial fitting using structures obtained from lower resolution structural methods such as low resolution X-ray crystallography, small angle X-ray scattering, and cryo-EM. As a test case, we fitted (Figs. 2 (h)-(o)) the mature conformation of the mature bacteriophage P22 capsid protein, obtained via cryo-EM, into density corresponding to the immature conformation of the capsid, also obtained using cryo-EM (10). As a bonus, here, FOLD-EM was also able to improve the alignment of the two subunits reported earlier in (10), as seen in Figs. 2 (n) & (o) (see Text S1 for additional commentary on this result).

Flexible Multi-Domain Docking to detect and characterize conformational change

Another unique aspect of FOLD-EM is its ability to carry out simultaneous multi-domain fitting that accounts for domain movements and conformational changes that may have occurred in a cryo-EM/ET map relative to the search model. Similar to the situation described above where only part of a cryo-EM structure is homologous to a search model, conformational differences between structures being compared can lead to inaccurate fitting results; it is not generally possible to simultaneously

align multiple domains if each domain requires a different geometric transformation to fit it into its corresponding electron density. Although there are approaches for flexible fitting (11), they are based on assumptions regarding domain boundaries and molecular motions that may or may not be true. Here, we show that FOLD-EM can automatically determine the extent of discreet structurally homologous domains/regions shared by two structures, and then separately fit each structural unit/domain. As a result, FOLD-EM is inherently capable of performing unbiased fully automated flexible fitting that makes no assumptions regarding domain boundaries or motions.

Briefly, FOLD-EM carries out flexible fitting by iteratively fitting each domain. For example, in the case of the three-domain GroEL structure (Fig. 3 (a)), the entire three domain structure can be input to FOLD-EM along with the input map. FOLD-EM will identify the largest domain via the **P** solver algorithm described in the SI Method and fit this domain into its corresponding sub-volume in the input cryo-EM map (Fig. 3 (b)). The remaining unfitted remnant structure will consist of the input structure minus the largest domain. In the next iteration, FOLD-EM will take the remnant structure and search the remaining map (the original map minus the sub-volume where the first domain was fitted) for the best fit. In this way, FOLD-EM recursively docks each domain to its corresponding sub-volume in the electron density map. The independently fitted domains can then be connected via appropriate linkers, and the geometry of the resulting structure regularized by an energy minimization program.

As a test case, we took the three domains of GroEL, and arbitrarily rearranged them to create three different GroEL conformations that consist of two, three, and four domains, respectively (Figs. 3 (e)-(g)). We then calculated cryo-EM maps in the range of 5-20 Å from the X-ray structure of GroEL. FOLD-EM was then used to carry out flexible fitting of each GroEL confomer into the simulated maps (Tables S3 (a)-(c), Figs. 3 (i)-(k)), resulting in excellent fits. The low RMSD errors listed in Tables S3

(a)-(c) show that FOLD-EM is capable of unbiased, fully automated flexible fitting. To confirm that the flexible fitting routine works with real data, we used FOLD-EM to fit the high resolution structure of one conformation of GroEL into 4 Å and 6 Å cryo-EM maps of GroEL in a different conformation (Figs. 3 (l)-(s)).

The above test cases demonstrated that FOLD-EM can accurately account for conformational changes when fitting noise-free high resolution structures into noisy cryo-EM maps. We also wanted to determine if FOLD-EM could characterize conformational changes when comparing one noisy low resolution structure to another. As a test case, we used FOLD-EM to fit a cryo-EM reconstruction corresponding to one conformation of the 70S ribosome into another reconstruction of an alternate conformer. Figs. S5 (a)-(f) shows the correspondences established by FOLD-EM between the first extracted domain pair and the second extracted domain pair. Visual inspection indicates that the correspondences are reasonably located. Furthermore, both the extracted domain pairs, when aligned, have cross-correlation values of 0.92 & 0.91, respectively. Additionally, the geometric transformations yielded by FOLD-EM that align the extracted domain pairs agrees with previously reported results (12). Hence, even when two relatively low resolution noisy structures are being compared, FOLD-EM is able to deduce the extent of two structurally homologous regions/domains, determine the different geometric transformations necessary to fit each domain, and thus provide a fitted structure that accounts for the conformational differences between the two maps (See the included Video S1 to see the deduced conformation change between the two). As a by-product, non-conserved, conformationally flexible regions of the molecule are also revealed (red regions in Figs. S5 (c)-(d)). As discussed above, multi-domain fitting allowing inter-domain rearrangement typically requires that the user decide on the extent of individual domains and then dock each domain separately. FOLD-EM executes the whole operation automatically. Furthermore, since FOLD-EM is

based on a global search, its flexible fitting capabilities are fully automated, avoid local minima, and make no assumptions regarding molecular movements.

Fully automated fold detection and large scale structural comparisons using FOLD-EM

The fold recognition/fitting scenarios described above assume the user knows the fold they are searching for in an electron density map. The user chooses either the identical molecule or a suitable homolog as a search model for fold recognition/fitting. Although homologous search models can often be identified via sequence comparisons, it is not always possible to identify a suitable homolog based on sequence homology. However, lack of sequence homology does not preclude structural homology, as it is well known that structural similarities often persist over large evolutionary distances where sequence vanishes. Hence, it would be useful to have a tool that systematically compares structural features of an electron density map to a large structural database and returns the best fitting homolog/s. However, rather than fitting entire structures, the goal here is simply to fit individual domains. More complex structures can then be inferred from the relative arrangement of individual domains. There are several advantages to this approach. First, domain databases are designed to include only representative folds, thus avoiding the redundancy present in the PDB. Second, the combinations and relative arrangements of individual domains can vary greatly in multi-domain proteins; by fitting domains separately the search is not necessarily confined to the different domain arrangements present in known structures. Hence we believe our modular approach to locating independent structural units is more akin to the modular design of proteins in nature, and is thus capable of a comprehensive search in spite of including only a limited number of structural units.

In this application, about 4000 representative protein domains from the SCOP database have been chosen as search models. The first member of a domain family is picked as the representative

structure, and additional structures from the same domain family are included if they structurally differ by > 5 Å RMSD from each other. These structures represent all super families of the five SCOP-domain classes: all-alpha, all-beta, alpha+beta, alpha and beta, and small proteins. Next, each domain is scored against the input electron density map using a modified scoring version of the module from FOLD-EM that has been optimized for speed. The domains with the best score are then returned as potential fits for the input electron density map. Below, we describe the use of FOLD-EM to screen the SCOP database and return a CA backbone model in a fully automated fashion, thus removing subjectivity from map analysis and relieving the user of the burden of identifying appropriate homologues as inputs.

As before, we have used the well-known structure of GroEL as a test case. Synthetic cryo-EM maps were calculated in the resolution range 5-20 Å. FOLD-EM was then used to search the SCOP database, identify the constituent domains in each map and return the fitted structures as CA backbone models for each of the simulated maps (Fig. 4 (a)). Tables S4 (a) lists candidate domains, selected by FOLD-EM, along with their associated scores and geometric transformations for the simulated 10 Å GroEL map. The first row of the table reports that the chosen 90 residue intermediate domain was docked into its corresponding region in the map with an RMSD error of 0.49 Å (with respect to the domain used to simulate that map region; Table S4 (b)). Similar results were obtained for simulated maps calculated at 5, 15, and 20 Å resolutions (fitted structures not shown); all reported error values are quite low (Table S4 (b)) demonstrating the ability of FOLD-EM to correctly identify and fit the constituent domain structures of GroEL.

To verify that FOLD-EM is capable of correctly identifying independent structural domains present in actual cryo-EM data with representative noise levels, we selected as test cases several moderate resolution cryo-EM maps where the domain structures of their constituent macromolecules is known.

These structures include: a 6 Å map of GroEL (6); a 7.9 Å map of the bacteriophage ϕ29 capsid protein (5); a 6.8 Å map of the Rice Dwarf Virus capsid protein (13); the 6.8 Å map of the 20S proteasome (14); and a 12.5 Å structure of the 70S ribosomal subunit (12) (also see Text S2). Table 1, Tables S4 (c)-(e) list candidate domains for different regions of each protein along with associated scores that were automatically determined by FOLD-EM. The domains with the highest scores were selected as constituent domains of the output CA models (Figs. 4 (b)-(f)). In every case except for one the highest scoring domains corresponded to the known domain structures for each input map. The one instance where FOLD-EM reported a better score for a SCOP domain different than previously reported was for the bacterial immunoglobulin domain of the capsid protein of bacteriophage ϕ29, where the correct fold had the fourth highest score.

The only previous work that comes close to our first-of-its-kind automated fold recognition work is EMATCH (17,18). However EMATCH is not an independent software, *i.e.*, it requires an input map to be first converted into a collection of helices (which requires manual specification of appropriate map density thresholds (9)) that exists in the input map, ignoring any other non-helical information in the input map. Hence this approach is not suitable for those input cryoEM maps which are predominantly defined by non-helical entities or even hardly-detectable helices (like ones that are short or occur in maps coarser than 10 Å resolution). FOLD-EM on the other hand, does not do any reduction of input maps and works on their full form and is fully automated (*i.e.*, it does not require the user to guess and provide appropriate input parameters). Hence unlike EMATCH, FOLD-EM is applicable to any kind of macromolecular cryoEM map.

In our subsequent publications, we will demonstrate the ability of FOLD-EM to also detect secondary structures (alpha-helices, beta sheets) in sub-nanometer resolution cryo-EM maps.

CONCLUSIONS

We have developed a new software tool, FOLD-EM, to automatically and systematically identify protein folds and fit atomic resolution macromolecular structures into cryo-EM electron density maps without any prior knowledge. FOLD-EM is based on the SIFT algorithm, a recent breakthrough enabling feature detection in computer vision applications including tracking by robots, 3D scene/object modeling and recognition/tracking, human action recognition, and brain analysis in 3D Magnetic Resonance images. We have adapted and extended the SIFT algorithm to automatically identify folds and characterize conformational changes in cryo-electron density maps of large macromolecular assemblies. FOLD-EM works by constructing rotationally invariant, low-dimensional representations of local regions in the input atomic resolution structures and cryo-EM maps. Correspondences are established between the reduced representations by comparing them using a simple metric. These correspondences are then clustered using hash tables and graph theory to identify structurally equivalent domains or motifs. Because correspondences are built from matching smaller structural units, the program will work even if only portions of domains are structurally homologous. Similarly, FOLD-EM will also automatically determine if different transformations are necessary for fitting different regions of the input search model; as a result, FOLD-EM is inherently able to characterize conformational differences between the structures being compared. Using FOLD-EM, we have demonstrated its effectiveness in: 1) partial matching, i.e. successful docking/fitting in the presence of extraneous protein residues; 2) fitting multi-domain structures into cryo-EM maps in a single step while taking into account flexibility due to inter-domain motions; and 3) performing fully automated large scale fold recognition and fitting using a protein domain database. The ability to automatically and objectively carry out these challenging tasks allows non-specialists to perform sophisticated structural analysis and sets FOLD-EM apart from other existing docking packages.

METHODS

This is how the large scale fold recognition in FOLD-EM works. In the first step FOLD-EM chooses about 4000 representative protein domains from SCOP. Usually the first member of a SCOP domain family is picked. More are picked from the same domain family if they are structurally at-least 5 Å RMSD from each other. These represent all superfamilies of the five true classes of SCOP: all-alpha, all-beta, alpha+beta, alpha and beta, and small proteins. Next we score each domain against the input cryoEM structure using a new scoring module, which we call MOTIF-EM. The domains with the best score are returned as potential fits for the input cryoEM maps. We now describe the scoring module MOTIF-EM. A less efficient version of MOTIF-EM has been published earlier as (3) by us. The version presented in this paper is ~10 times faster.

MOTIF-EM solves the structural comparison problem **P** defined as follows: compare a non-atomic resolution structure (i.e. a cryoEM map from EMDB) with another structure (either another cryoEM map or a map blurred from a crystal structure) and identify conserved structural domains or motifs or sub-map (if there is any) between the pair of input structures. The precise algorithm used by MOTIF-EM to solve **P** is outlined in the Figs. S1, S2, & S3. The technique used by MOTIF-EM to detect conserved sub-structures is inspired by a recent breakthrough in 2D object recognition (1). The input to MOTIF-EM is a pair of volumetric electron density maps. The program then uses geometric processing, statistical analysis, and graph theory to detect conserved regions between the input pair by executing the following six steps. In step 1 (Figs. S1 (step 1), S2, & S4 (a)), three-dimensional Cartesian reference frames are assigned to every grid point in each of the input maps. These reference frames are computed by examining the local density variations at each grid point, using singular value decomposition. For example, the primary axis of the reference frame points to the

direction of largest local density variation. In step 2 (Figs. S1 (step 2), S3, S4(b & c)), for each grid point in the input maps, we construct a local region descriptor (LRD) - a rotationally-invariant, low-dimensional representation of electron density variation in the local region around the grid point. The LRD for a grid point p is essentially an orientation histogram of the local density variation vectors around p that were calculated in step 1, i.e., the first axis of the reference frames for the neighboring grid points. In step 3 (Figs. S1 (step 3), S4(d)), for a grid point p in input map 1, we find k potential matches in input map 2, i.e., local regions in map 2 which are “similar” to the local region around p. These matches are essentially those grid points in input map 2, whose LRDs closely match the LRD for point p in map 1. In steps 4 and 5 (Figs. S1 (step 4 & 5), S4(e & f)), we cluster all the matches obtained from step 3, based on the six degrees of freedom geometric transformation that maps a grid point (along with the local reference frame) onto its match in map 2. In step 6 (Figs. S1 (step 6), S4(g)), we choose the most prominent cluster obtained in the previous step. The matches in this prominent cluster form the potential conserved domain between the input maps. False positives occur due to two main reasons: (a) high noise in either/both of the input maps and (b) dimensionally reduced representations (LRDs) used to characterize local regions necessarily result in information loss. However, these false positives are removed using graph theory; a graph is constructed with the matches in the prominent cluster as its node. An edge is added between two nodes in the graph if the inter-point distance (between the two grid points of the same map in the two graph nodes) is preserved across the input map pair. Finally the largest clique in the graph (the sub-graph with an edge between every pair of nodes) is the final predicted domain region that is structurally conserved between the pair of input maps.

Now, inside FOLD-EM we also extend MOTIF-EM to do docking/fitting. This is possible because a high resolution structure can be blurred and fed as input to MOTIF-EM. Since MOTIF-EM also gives

relative spatial orientation of the common sub-structures extracted by MOTIF-EM, it is used by FOLD-EM to eventually dock/fit the structure into the input cryoEM map.

The version of MOTIF-EM presented here is much more efficient (~10 times) than what we presented earlier in (3). This was made possible by clustering only those degrees-of-freedom which occur in dense regions.

FOLD-EM is highly parrallelizable. The fold recognition testcases using the SCOP database took 72-90 hours to execute in a 100 processor computing cluster at University of Texas Medical Branch, Galveston (UTMB). However they will only take few hours in the kind of computing machinery available with www.teragrid.org, where there are computing clusters with thousands of processors. The one-time docking testcases took about two minutes, each, to execute on the 100 processor UTMB cluster.

Simulated cryo-EM maps were generated from atomic resolution structures using EMAN (4).

Inventory of Supplementary Information

Supplementary Figure 1 (Fig. S1)	Outline of the algorithm to solve **P**
Supplementary Figure 2 (Fig. S2)	Outline of algorithm for computing local Cartesian reference frames in Step 1 of Supplementary Figure 1
Supplementary Figure 3 (Fig. S3)	Outline of algorithm for computing local region descriptors (LRDs) in Step 2 of Supplementary Figure 1
Supplementary Figure 4 (Fig. S4)	Cartoon representations of the steps to solve **P**
Supplementary Figure 5 (Fig. S5)	Flexibility in ribosome 70S
Supplementary Table 1 (Table S1)	RMSD error in docking/fitting, using FOLD-EM
Supplementary Table 2 (a)-(c) (Table S2)	RMSD errors in docking/fitting in the presence of extraneous regions.
Supplementary Table 3 (a)-(c) (Table S3)	Flexible fitting
Supplementary Table 4 (a)-(e) (Table S4)	Automated fold recognition
Supplementary Text 1 (Text S1)	Validation of P22 results
Supplementary Text 2 (Text S2)	Fold recognition in ribosome 70S
Supplementary Video 1 (Video S1) Legend	

ACKNOWLEDGEMENT

This research has benefited from discussions with Werner Braun, Kyung Choi, Wah Chiu, Michael Levitt, Gunnar Schroder, Steve Ludtke, Matthew Baker, Yao Cong, Dongua Chen, Xiangan Liu, David Woolford and Junjie Zhang. We also thank the NIH center for Biomedical Computation at Stanford University (http://simbios.stanford.edu) and the National Center of Macromolecular Imaging at Baylor College of Medicine for their initial support. We also thank Dr. Mark White for assistance with using the UTMB computing cluster. We also thank TERA-GIRD (http://www.teragrid.org) for providing their computing resources.

REFERENCES

1. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision,* 60(2), 91-110.

2. http://en.wikipedia.org/wiki/Scale-invariant_feature_transform#Comparison_of_SIFT_features_with_other_local_features

3. Saha M, Chiu W, Levitt M (2004) MOTIF-EM: an automated computational tool for identifying conserved regions in CryoEM structures. M. *Bioinformatics* 26 (12):301-309.

4. Ludtke SJ, Baldwin PR, Chiu W (1999). EMAN: semi-automated software for high-resolution single-particle reconstructions. *J Struct Biol 128:*82-97.

5. Morais MC *et al*. (2005) Conservation of the Capsid Structure in Tailed dsDNA Bacteriophages: the Psuedoatomic Structure of ϕ29. *Molecular Cell* 18:149-159.

6. Ludtke S, Chen D, Song J, Chuang D, Chiu W (2004) Seeing GroEL at 6 Å Resolution by Single Particle Electron Cryomicroscopy, *Structure* 12 (7):1129-1136.

7. Pettersen EF *et al*. (2004) UCSF Chimera--a Visualization System for Exploratory Research and Analysis. *J. Comp. Chem.* 25(13):1605-12.

8. Wriggers W, Birmanns S (2001) Using Situs for Flexible and Rigid-Body Fitting of Multiresolution Single-Molecule Data. *J. Struct. Biol.* 133:193-202.

9. Jiang W, Baker ML, Ludtke SJ, Chiu W (2001) Bridging the Information Gap: Computational Tools for Intermediate Resolution Structure Interpretation. *J. Mol. Biol.* 208:1033-1044 .

10. Jiang W *et al*. (2003) Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nature Struct. Biol.* 10:131-135.

11. Suhre K, Navaza J, Sanejouand YH (2006) NORMA: A tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr. D Biol. Crystallogr.* 62:1098–1100.

12. Valle M *et al*. (2003) Locking and Unlocking of Ribosomal Motions. *Cell* 114:123-134.

13. Zhou ZH *et al*. (2001) Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nat Struct Biol* 8:868-873.

14. Rabl J *et al*. (2008) Mechanism of Gate Opening in the 20S Proteasome by the Proteasomal ATPases. *Molecular Cell* 30:360-368.

15. Nakagawa A *et al*. (2003)The atomic structure of RDV reveals the self-assembly mechanism of component proteins. *Structure* 11(10):1227-1238.

16. Ludtke SJ *et al*. (2008) De novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure* 16 (3):441-8.

17. Lasker K, Dror O, Nussinov R, Wolfson H (2005) Discovery of Protein Substructures in EM Maps. *Algorithms in Bioinformatics* 423-434.

18. Lasker K, Dror O, Shatsky M, Nussinov R, Wolfson HJ (2007) EMatch: Discovery of High Resolution Structural Homologues of Protein Domains in Intermediate Resolution Cryo-EM Maps. *IEEE transactions on Comp. Biol. and Bioinformatics* 4(1):28-39.

FIGURE LEGENDS

Figure 1

(a): We test the sanity of FOLD-EM, as a docker, by docking/fitting domains of three different sizes (equatorial, apical, and intermediate domains of GroEL; ribbon models shown on left) into cryoEM maps (like the gray map shown in middle) simulated from the GroEL monomer (PDB ID: 1OEL), in the resolution range of 5-20 Å. The rightmost image shows a result: the three domains fitted into a map using FOLD-EM.

(b): The fitting of the atomic resolution GroEL domains (green, blue, & yellow wire-frames) into a 6 Å cryoEM map (gray region) of GroEL (from (6)) using FOLD-EM. The fittings are consistent with the results in (6).

(c): Fitting of the atomic-resolution GroEL intermediate domain (blue wire-frame) into the 6 Å GroEL cryoEM map, as determined by FOLD-EM (this is (b), enlarged, with only the map region of the intermediate domain shown for clarity).

(d): In-correct docking of the same intermediate domain (wire-frames) into the equatorial region of the map done by a popular fitting software SITUS (8) (more than one solution is shown; similar failure was encountered by other popular fitting software: FOLDHUNTER (9) and the Chimera fitting tool (7); SITUS, FOLDHUNTER, Chimera were successful in fitting the other two GroEL domains). We believe, these failures occurred because the intermediate domain is too small compared to the target map.

(e): Successful (consistent with (5)) fitting of the HK97 (yellow ribbon) and BIG2 domains (blue ribbon) obtained using FOLD-EM.

(f): In-correct fitting of BIG2 obtained using the popular fitting software SITUS (similar failure was encountered by popular fitting software: FOLUDHUNTER and Chimera fitting tool). We believe, failure occurred because the BIG2 domain is too small compared to the target map.

Figure 2

(a) & (b): Cartoon illustrating problems associated with fitting partial structures into cryoEM maps. The high-resolution structure (wire model in red and black) has extraneous region (red) which does not have corresponding density in the cryoEM map (pale blue region). This extraneous region can act as noise and reduce the accuracy of the fitting and the associated score (as seen in (b)). Since FOLD-EM can separate conserved regions from non-conserved ones, it can potentially detect and eliminate the red extraneous region, yielding an accurate fitting (like the one seen in (a)) and associated score.

(c), (d), & (e): Some domains that were fitted using FOLD-EM. The red region shows the noise/extraneous residues that was incorporated to test the robustness of FOLD-EM against them.

(f): The fitted green ribbon structure shows the correct fitting (consistent with (6)) of the apical domain with some added extraneous residues (like the one shown in (d)), obtained using FOLD-EM. The rest of the ribbon structures in the equatorial domain region show the in-correct fittings obtained using SITUS.

(g): The fitted magenta ribbon structure shows the correct fitting (consistent with (6)) of the equatorial domain with some added extraneous residues (like the one shown in (e)), obtained using FOLD-EM. The yellow ribbon structure is the in-correct fitting (off by at-least 6.2 Å) obtained using SITUS.

(h) & (i): Monomers from pre- and post- capsid maturation states of phage P22, respectively (10).

(j) & (k): The conserved region between the monomers (shown in (h) & (i)) is shown in blue, as determined by FOLD-EM. The rest of the region is shown as red.

(l) & (m): Here only the conserved region (colored as blue in (j) & (k)) is shown, from two different views.

(n) & (o): (enlarged *wrt* (l) & (m)): alignment of the extracted conserved pairs (shown in view #2 of (l) & (m)) using FOLD-EM and data from (10), respectively. Circled regions in (o) highlight areas of poor local alignment, determined by visual inspection.

Figure 3

(a): A high resolution GroEL conformation (PDB ID: 1AON). (b): A lower resolution (4 Å) GroEL in a different conformation. FOLD-EM docks (a) into (b) by deforming it into a conformation (c) which fits well into (b) as shown in (d).

(e)-(g): (the figures on left): We create three fictitious atomic-resolution resolution GroEL conformations. (e, left) has two domains, (f, left) has three domains, and (g, left) has four domains. Then we attempt to dock each (which requires domain rearrangement) into a low resolution version of the structure in a different conformation. For example, the three domain structure (f, left) is docked onto a map (simulated from the three domain GroEL structure with PDB ID: 1OEL) in a different conformation shown in (f, right). Embedded ribbon structures shown in the figures are the ones used to simulate the respective maps.

(i)-(k): A result: conformation #2 (i or (f, left)) transformed into a new structure (j) using FOLD-EM which fits very well (k) into a simulated GroEL 10 Å cryoEM map.

(l)-(o): (l) is an atomic-resolution resolution GroEL conformation (PDB ID: 1AON). (m) is a lower resolution (4 Å) GroEL in a different conformation. FOLD-EM docks (l) into (m) by deforming it into a conformation (n) which fits well (consistent with (16)) into (m) as shown in (o).

(p)-(s): (p) is an atomic-resolution resolution GroEL conformation (PDB ID: 1AON). (q) is a lower resolution (6 Å) GroEL in a different conformation. FOLD-EM docks (p) into (q) by deforming it into a conformation (r) which fits well (consistent with (6)) into (q) as shown in (s).

Figure 4

(a): The construction of a CA backbone model for a simulated GroEL map from the best scored candidate domains (1st row in Table S4 (a)).

(b): the fitting (done by FOLD-EM; consistent with (4)) of the three chosen domains into the 6 Å cryoEM map of GroEL.

(c): the fitting (done by FOLD-EM; consistent with (5)) of the two chosen domains into the 7.9 Å cryoEM map of Phi29 from (5).

(d): the fitting (done by FOLD-EM; consistent with (13, 15)) of the two chosen domains into the 6.8 Å cryoEM map of Rice Dwarf Virus from (13).

(e): the fitting (done by FOLD-EM; consistent with (14)) of the chosen trimer domain into the 6.8 Å cryoEM map of 20S proteasome from (14).

(f): the fitting (done by FOLD-EM) of the two chosen domains (30S and 50S) into the 12.5 Å cryoEM map of ribosome 70S from (12).

TABLES

1KP8 (A:2-136,A:410-526)	0.57	1KID (A)	0.45	1KP8 (A:137-190, A:367-409)	0.45
1KID (A)	0.45	1LS1 (A:1-88)	0.44	2B5E (A:142-239)	0.40
1LS1 (A:1-88)	0.44	2GOY (A:7-138)	0.34	1ABV (A)	0.40
2GOY (A:7-138)	0.34	1H5P (A)	0.33	1YSJ (A:178-292)	0.40
1H5P (A)	0.33	1M9L (A)	0.30	2RLT (A:1-99)	0.40

Table 1: lists candidate domains, with associated scores, automatically picked by FOLD-EM for building the CA backbone of the GroEL map. Three domains were picked: equatorial (column 1&2; column 2 is the associated FOLD-EM generated score), apical (column 3&4), and the intermediate domain (column 5&6). The first row lists the three domains with best scores, which are finally chosen by FOLD-EM to build the CA model of the map.

FIGURE 1

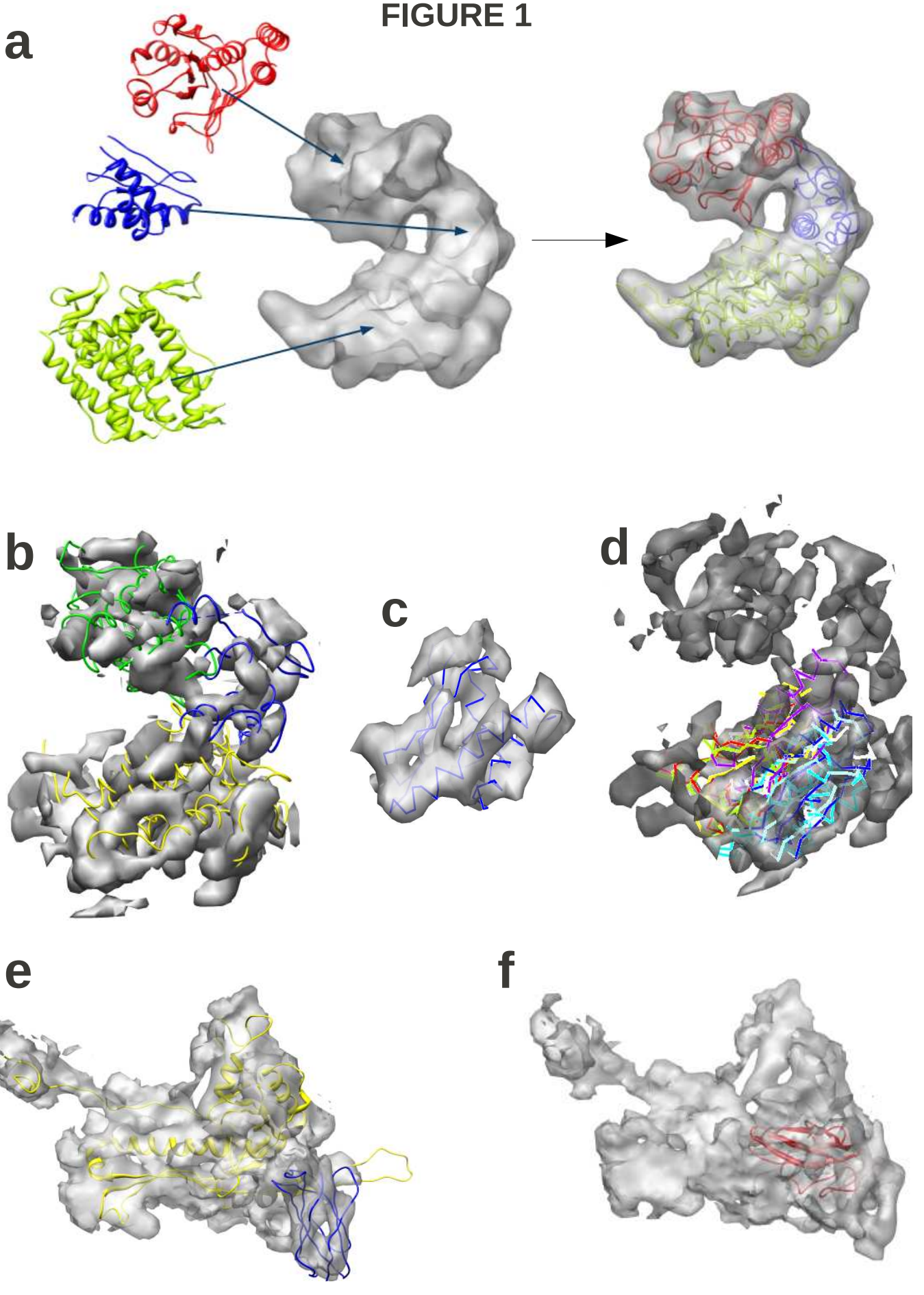


FIGURE 2

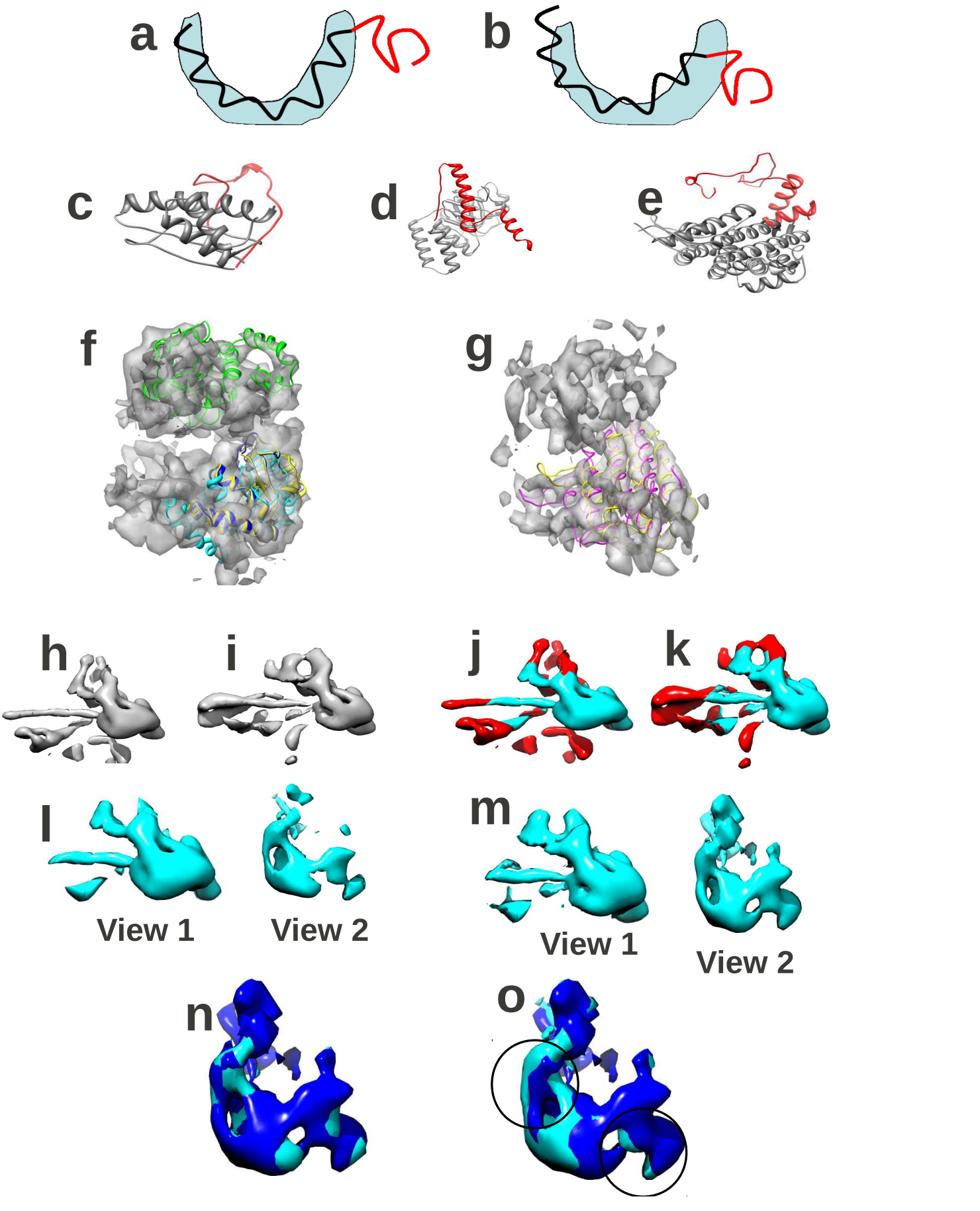


FIGURE 3

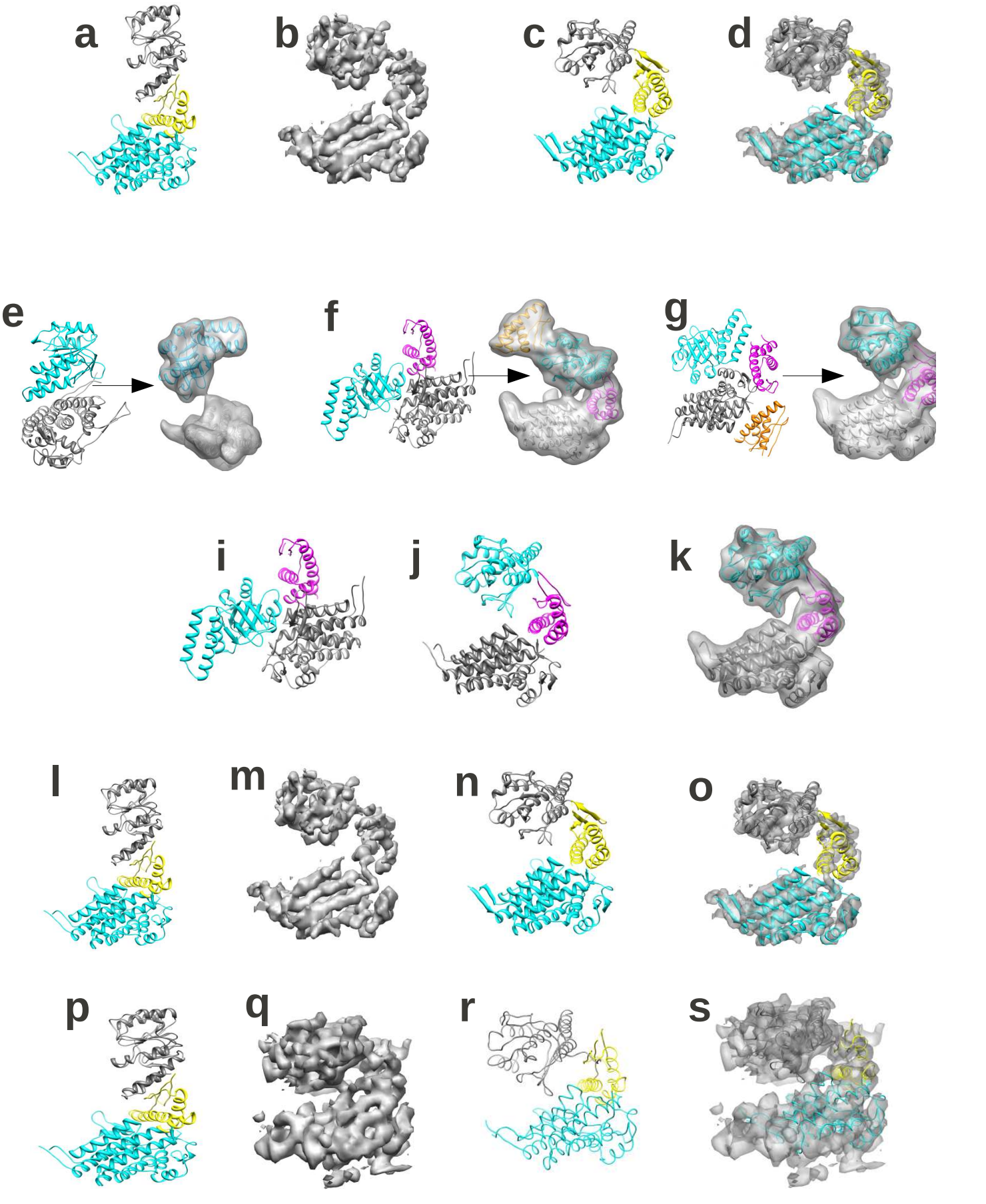
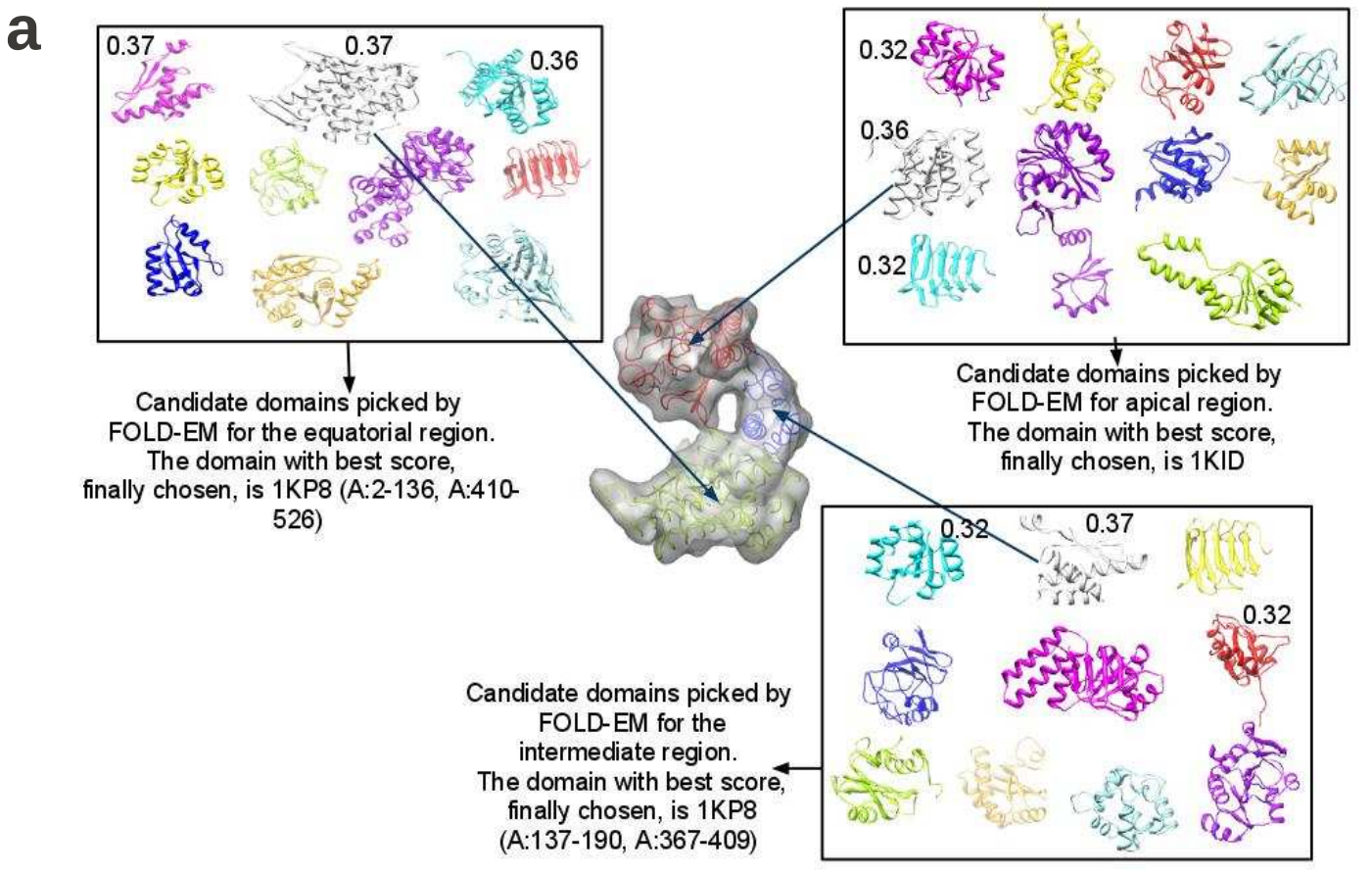
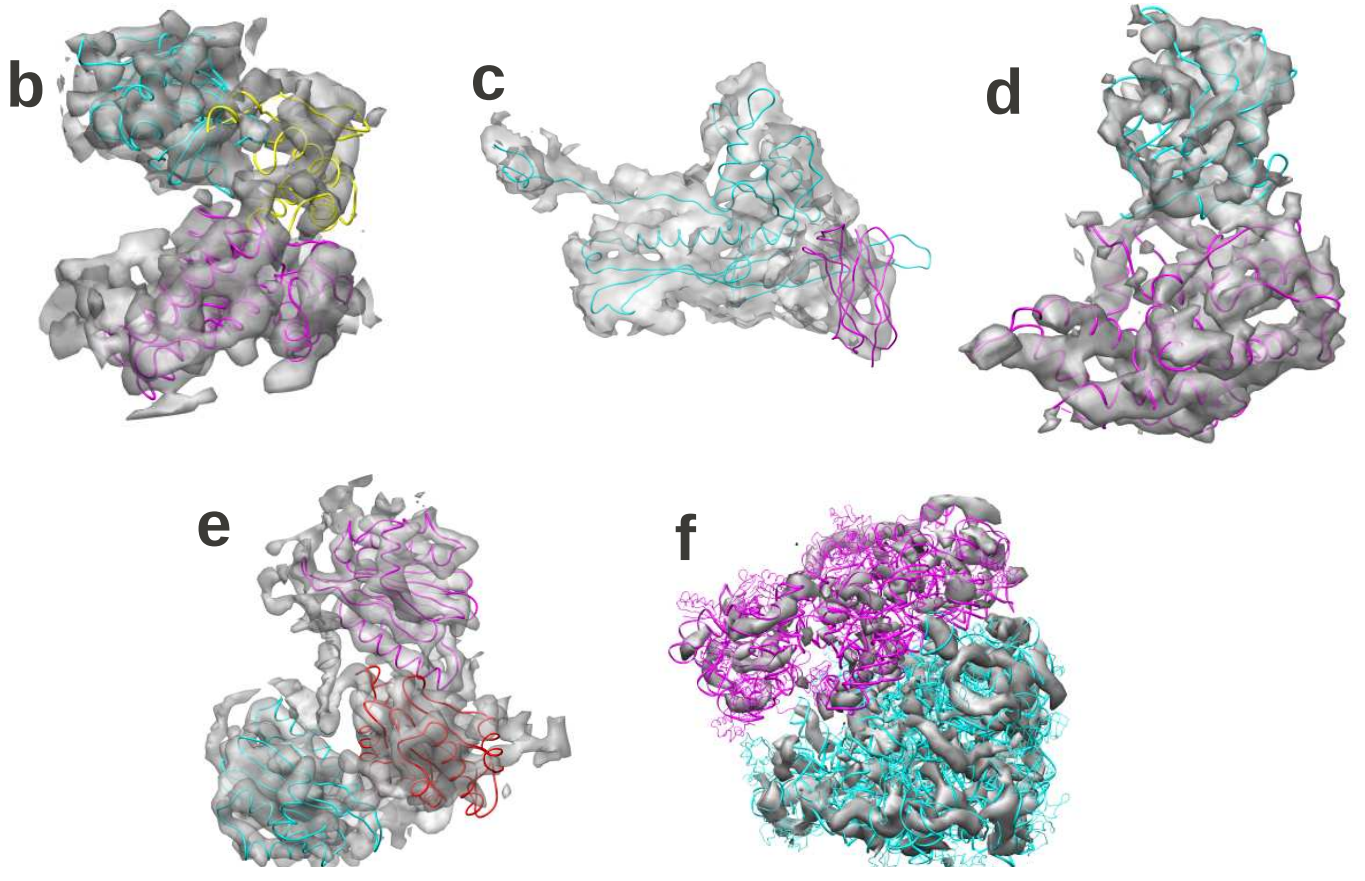


FIGURE 4


a
0.37
0.37
0.36
0.32
0.36
0.32
0.32
0.37
0.32
Candidate domains picked by FOLD-EM for the equatorial region. The domain with best score, finally chosen, is 1KP8 (A:2-136, A:410-526)
Candidate domains picked by FOLD-EM for apical region. The domain with best score, finally chosen, is 1KID
Candidate domains picked by FOLD-EM for the intermediate region. The domain with best score, finally chosen, is 1KP8 (A:137-190, A:367-409)


b
c
d
e
f

SUPPLEMENTARY INFORMATION (SI)

Fully Automated Fold Recognition in Low Resolution Electron Density Maps

Mitul Saha and Marc C. Morais

The Algorithm to solve P:

Notations:

M_i: map i, i=1,2

p^i_j: grid point p_j in map i.

Λ^i_j: LRD at grid point p_j in map i.

$m(p^i_j, p^k_l)$: A match pair of grid points p^i_j (from map i) and p^k_l (from map k), i~=k

$O(p^i_j)$ or O^i_j: O-XYZ Cartesian reference frame at grid point p^i_j

If S is a set, $S(i)$ is the i-th element of S

X': transpose of X

The Algorithm to solve **P**
Inputs: CryoEM maps M_1, M_2

1. Compute Cartesian frame sets for M_1 and M_2:

$\boldsymbol{O}(M_1)$:$\{O^1_1, O^1_2, \ldots\}$=**compute_frame_set**(M_1)

$\boldsymbol{O}(M_2)$:$\{O^2_1, O^2_2, \ldots\}$=**compute_frame_set**(M_2)

2. Compute LRD sets for M_1 and M_2:

$\boldsymbol{\Lambda}(M_1)$=$\{\Lambda^1_1, \Lambda^1_2, \ldots\}$=**compute_LRD_set**($M_1$, $\boldsymbol{O}(M_1)$)

$\boldsymbol{\Lambda}(M_2)$=$\{\Lambda^2_1, \Lambda^2_2, \ldots\}$=**compute_LRD_set**($M_2$, $\boldsymbol{O}(M_2)$)

3. For a given LRD Λ^1_i in $\boldsymbol{\Lambda}(M_1)$, find k closest LRDs from $\boldsymbol{\Lambda}(M_2)$:

$\boldsymbol{\Lambda}^1_{i\_closest}$=$\{\Lambda^2_{i1}, \Lambda^2_{i2}, \ldots, \Lambda^2_{ik}\}$, Λ^2_j $\boldsymbol{\Lambda}(M_2)$;

Let $m(p^1_i, p^2_{ij})$: $\{\Lambda^1_i, \Lambda^2_{ij}\}$ define a match pair,

Λ^2_{ij} is the j-th element in $\boldsymbol{\Lambda}^1_{i\_closest}$

4. For every match pair $m(p^1_a, p^2_b)$, obtained in step 3, find the corresponding $6DOF(p^1_a, p^2_b)$=**find_dof**($[O^1_a\ p^1_a]$, $[O^2_b\ p^2_b]$)

5. Cluster the 6DOFs obtained in step 4.

6. For each large cluster C_i, from step 5, construct an un-weighted graph G_i. A node in G_i is a match pair from C_i. An edge exists between two nodes in G_i if inter-point distances, corresponding to the match pairs in the two nodes, are preserved. Find the largest clique $S(G_i)$ in G_i and return the match pairs in $S(G_i)$ as the rigidly conserved domain pair.

Fig. S1. Outline of the algorithm to solve P

(Notations: See Fig. S1)

Algorithm **compute_frame_set**

Input: map M

1. At a grid location p_i of M,

   Cartesian reference frame O_i = **compute_frame**(M, p_i)

2. Return {O_1, O_2, …}

Algorithm **compute_frame**

Inputs: map M, grid location p_o in M

S1. Sample k points {p_1, p_2, ..., p_k} uniformly in the

   neighborhood (within r_o radius) of p_o

S2. Let v_i be the density value at p_i in M

   Define matrix P_{kx3} as [w_1*v_1*(p_1-p_o); w_2*v_2*(p_2-p_o); ....]

   - i-th row of P_{kx3} is w_i*v_i*(p_i-p_o)

   - w_i is a Gaussian wt: w_{o1}*exp(-(w_{o2}*|p_o-p_i|^2))

S3. [U_{3x3} D_{3xk} V_{3xk}]=SVD(P_{kx3})

S4. Return the Cartesian reference frame at p_o,

   O-XYZ(p_o): [O_x O_y O_z]=U_{3x3}

Fig. S2. Outline of algorithm for computing local Cartesian reference frames in Step 1 of Fig. S1

(Notations: See Fig. S1)

Algorithm **compute_LRD_set**

Input: Map M, Cartesian frame set: $\{O_1, O_2, \ldots\}$ (O_i is the frame at p_i in M)

1. At a grid location p_i of M, LRD Λ_i = **compute_LRD**(M, p_i, O_i)

2. Return $\{\Lambda_1, \Lambda_2, \ldots\}$

Algorithm **compute_LRD**

Inputs: Map M, grid location p_o in M, Cartesian frame $O(p_o)$:$[O_x\ O_y\ O_z]$ at p_o

S1. Let H be a gradient histogram with m bins: $\{b_1, b_2, ..b_{8*26}\}$

S1.1 divide the region around p_o into 8 equal quadrants: $\{q_1, q_2, \ldots, q_8\}$, in the local frame $O(p_o)$. Let each quadrant have 26 representative directions: **D**:$\{d_1, d_2, \ldots, d_{26}\}$={{[-1/0/1, -1/0/1, -1/0/1]}-[0, 0, 0]}. d_i is finally normalized.

S1.2 bin b_i corresponds to $\{q(\text{ceil}(i/26)), d(1+i\%26)\}$

S1.3 initialize b_i=0

S2. Sample k points $\{p_1, p_2, ...p_k\}$ uniformly in the neighborhood (withen r radius) of p_o

S2.1 let V_i=$O(p_i)_x$

S2.2 let V_{i2}=$(O(p_o).V_i)'$

S2.3 let p_{i2}=$(O(p_o).(p_i-p_o))'$

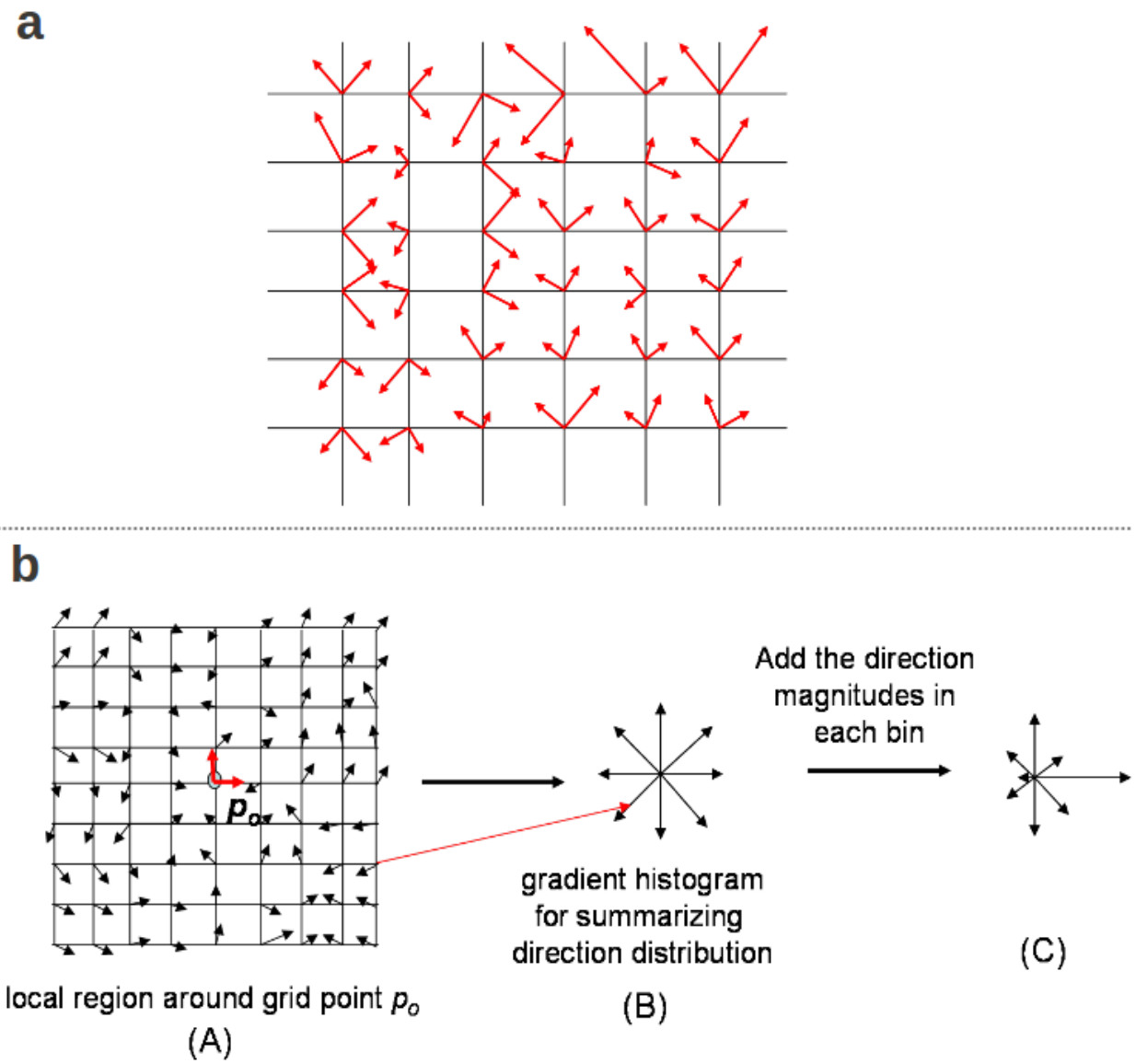
S2.4 find a bin b_i=$\{q_a, d_b\}$, such that p_{i2} is in q_a and d_b is the direction from **D** closest to V_{i2}.

S2.5 let b_i+=$v_i.w_i$

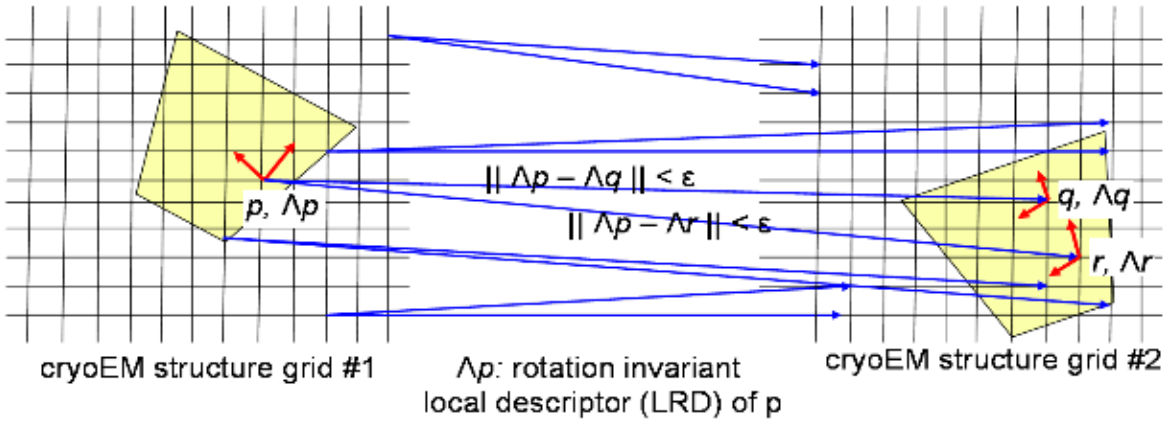
-v_i: magnitude of V_i or $D_{3xk}(1)$ obtained from step S3 in Fig. S2

-w_i: Gaussian wt: $w_{o1}.\exp(-w_{o2}.|p_o-p_i|^2)$

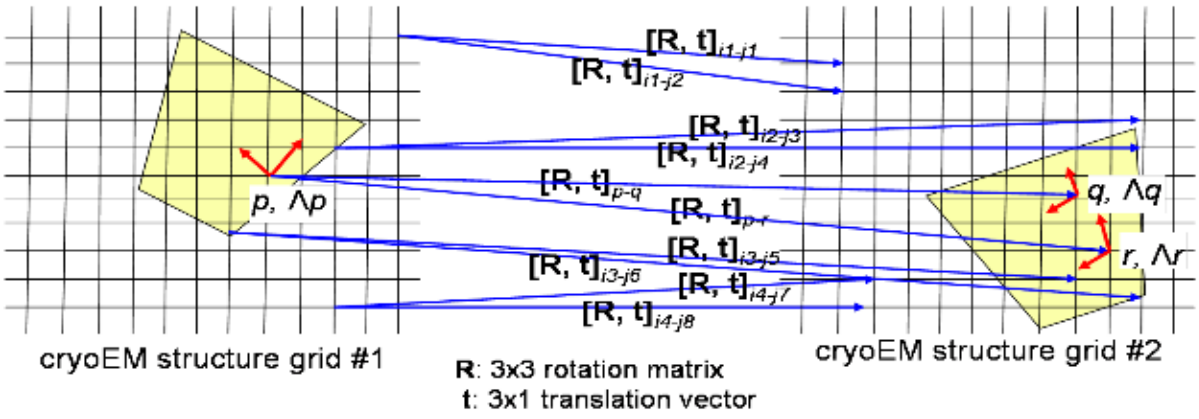
Fig. S3. Outline of algorithm for computing local region descriptors (LRDs) in Step 2 of Fig. S1


a
b
p_o
Add the direction magnitudes in each bin
gradient histogram for summarizing direction distribution
(C)
local region around grid point p_o
(A)
(B)
c
p_o
local region around grid point p_o
A more complex LRD

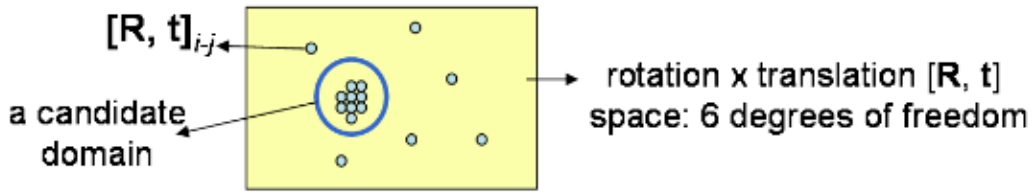
d


p, Λp
|| Λp – Λq || < ε
|| Λp – Λr || < ε
q, Λq
r, Λr
cryoEM structure grid #1
Λp: rotation invariant
local descriptor (LRD) of p
cryoEM structure grid #2

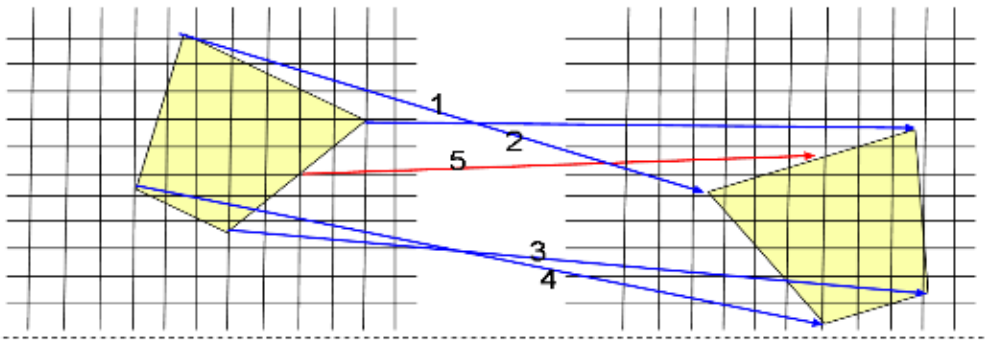
e


[R, t]i1-j1
[R, t]i1-j2
[R, t]i2-j3
[R, t]i2-j4
[R, t]p-q
[R, t]p-r
[R, t]i3-j5
[R, t]i3-j6
[R, t]i4-j7
[R, t]i4-j8
p, Λp
q, Λq
r, Λr
cryoEM structure grid #1
R: 3x3 rotation matrix
t: 3x1 translation vector
cryoEM structure grid #2

f


[R, t]i-j
a candidate
domain
rotation x translation [R, t]
space: 6 degrees of freedom

g


1
2
5
3
4

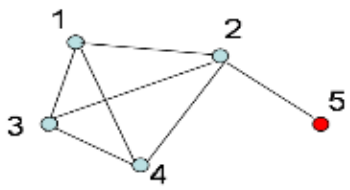

1
2
3
4
5

Fig. S4. Cartoon representation of the steps (of algorithm in Fig. S1) to solve P.

(a): Step 1 of Fig. S1, mimicked in 2D. A Cartesian reference frame is placed at each of the grid points. The length of a frame axis reflects the extent of local density variation along the axis.

(b): Step 2 of Fig. S1, LRD or gradient histogram construction, mimicked in 2D. The principal direction (X axis) of the reference frame of a grid point around p_o is first re-expressed in p_o's reference frame and then stored in the bin (of the gradient histogram) representing the direction closest to the re-expressed one. The magnitudes of the stored gradients in a bin are summed up to obtain a numerical value for each bin (reflected in the length of the directions in (C)).

(c): The local region around p_o can be divided into quadrants. LRDs, one from each quadrant, can be stacked together as a single vector to construct a more complex LRD.

(d): Step 3 of Fig. S1, mimicked in 2D. For a given grid point p in input cryoEM grid 1, locally similar grid points are found in the input cryoEM grid 2 by comparing the LRD at p with LRDs in grid 2.

(e): Step 4 of Fig. S1, mimicked in 2D. For a given match, there exists a spatial rotation **R** and a spatial translation **t**, that transforms the match pair onto each other.

(f): Step 5 of Fig. S1, mimicked in 2D. The match pairs obtained in Step 4 of Fig. S1 are clustered in the [rotation x translation] space.

(g): Step 6 of Fig. S1, mimicked in 2D. A graph is constructed such that match pairs are nodes. An edge between two nodes indicates that the distance between the corresponding two grid points is preserved between the two maps. A clique in the graph (formed by blue nodes) is a collection of grid points whose inter-point distances are preserved between the maps.

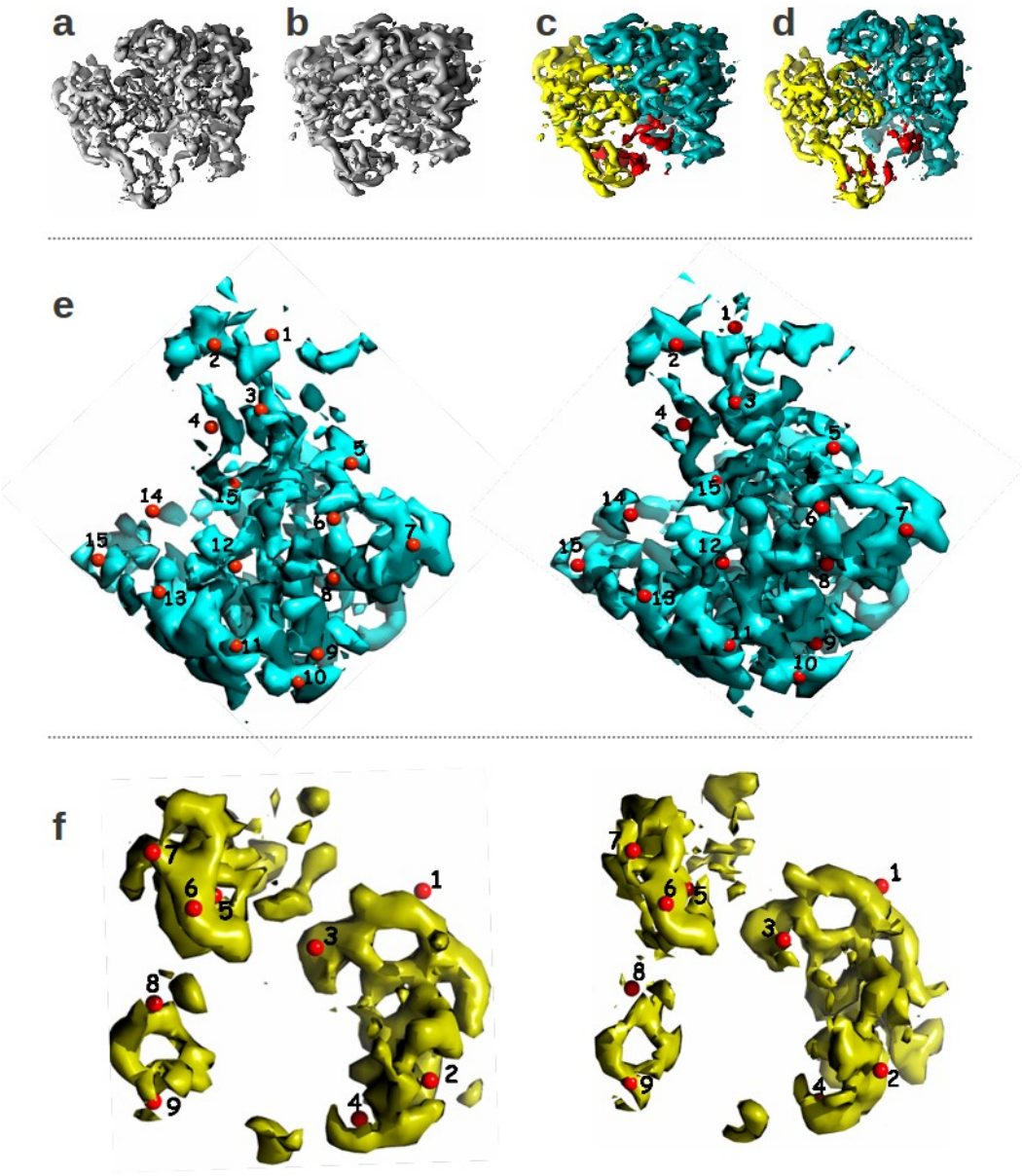


Fig. S5. Flexibility in ribosome 70S.

We also used FOLD-EM to dock the two low resolution (10 Å+) conformations of ribosome 70S (a & b) onto each other. FOLD-EM decomposes the two conformations into two rigid domains (the two regions colored as yellow and cyan) as shown in (c) & (d). The red region in (c) & (d) is the remnant non-conserved region in the input maps (a) & (b).

(e) & (f) (enlarged compared to a & b) show the correspondences (numbered red balls), established by FOLD-EM, between the first extracted domain pair (e) and the second extracted domain pair (f), respectively. The first extracted pair is predominantly the 30S subunit of the 70S ribosome, as per (12). The second extracted pair is predominantly the 50S subunit of the 70S ribosome, as per (12).

Map resolution (Å)	Transformation Error (RMSD Å)	Transformation Error (RMSD Å)	Transformation Error (RMSD Å)
5	0.29	0.23	0.23
10	0.29	0.26	0.31
15	0.47	0.17	0.33
20	0.53	0.36	0.43

Table S1. RMSD error in docking/fitting, using FOLD-EM

Column 2: Fitting errors for the intermediate domain of GroEL (size: 90 residues),

Column 3: Fitting errors for the apical domain of GroEL (size: 182 residues),

Column 4: Fitting errors for the equatorial domain of GroEL (size: 249 residues).

The fittings are done onto simulated GroEL cryo-EM maps with resolution ranging from 5-20 Å (column 1).

(A RMSD error for a fitting is computed between the fitted atomic-resolution domain and the atomic-resolution domain used to simulate the map region where the fitting occurs).

Tables S2 (a)-(c). RMSD errors in docking/fitting in the presence of extraneous regions.

Map resolution (Å)	Transformation Error (RMSD Å) (10% extra noise residues)	Transformation Error (RMSD Å) (20% extra noise residues)	Transformation Error (RMSD Å) (30% extra noise residues)
5	0.11	0.20	0.26
10	0.16	0.25	0.24
15	0.36	0.37	0.47
20	0.68	0.68	0.59

Table S2 (a): RMSD error in fitting, using FOLD-EM, an atomic resolution domain (intermediate domain of GroEL; size: 90 residues) in a simulated GroEL monomer with resolution ranging from 5-20 Å. The rightmost three columns signify cases when 10%, 20%, and 30%, respectively, extra residues were added as noise to the domain to be docked.

Map resolution (Å)	Transformation Error (RMSD Å) (10% extra noise residues)	Transformation Error (RMSD Å) (20% extra noise residues)	Transformation Error (RMSD Å) (30% extra noise residues)
5	0.11	0.15	0.12
10	0.11	0.18	0.16
15	0.18	0.23	0.20
20	0.24	0.24	0.32

Table S2 (b): RMSD error in fitting, using FOLD-EM, an atomic resolution domain (apical domain of GroEL; size: 182 residues) in a simulated GroEL monomer with resolution ranging from 5-20 Å. The rightmost three columns signify cases when 10%, 20%, and 30%, respectively, extra residues were added as noise to the domain to be docked.

Map resolution (Å)	Transformation Error (RMSD Å) (10% extra noise residues)	Transformation Error (RMSD Å) (20% extra noise residues)	Transformation Error (RMSD Å) (30% extra noise residues)
5	0.25	0.20	0.22
10	0.19	0.28	0.26
15	0.35	0.36	0.28
20	0.49	0.43	0.42

Table S2(c): RMSD error in fitting, using FOLD-EM, an atomic resolution domain (equatorial domain of GroEL; size: 249 residues) in a simulated GroEL monomer with resolution ranging from 5-20 Å. The rightmost three columns signify cases when 10%, 20%, and 30%, respectively, extra residues were added as noise to the domain to be docked.

Tables S3 (a)-(c). Flexible fitting.

Map resolution (Å)	Transformation Error (RMSD Å) (equatorial)	Transformation Error (RMSD Å) (apical)
6	0.12	0.07
12	0.24	0.1
18	0.30	0.15

Table S3 (a): Errors in the transformation predicted by FOLD-EM, for each of the two domain: Column #1 for equatorial and Column #2 for apical. This is for conformation #1.

Map resolution (Å)	Transformation Error (RMSD Å) (equatorial)	Transformation Error (RMSD Å) (apical)	Transformation Error (RMSD Å) (intermediate)
6	0.25	0.07	0.07
12	0.28	0.07	0.11
18	0.35	0.17	0.18

Table S3 (b): Errors in the transformation predicted by FOLD-EM, for each of the three domain: Column #1 for equatorial, Column #2 for apical, and Column #3 for intermediate. This is for conformation #2.

Map resolution (Å)	Transformation Error (RMSD Å) (equatorial)	Transformation Error (RMSD Å) (apical)	Transformation Error (RMSD Å) (intermediate)	Transformation Error (RMSD Å) (intermediate-II)
6	0.09	0.08	0.07	0.06
12	0.14	0.15	0.14	0.09
18	0.20	0.16	0.36	0.49

Table S3 (c): Errors in the transformation predicted by FOLD-EM, for each of the four domain: Column #1 for equatorial, Column #2 for apical, Column #3 for intermediate, and Column #3 for the second intermediate domain. This is for conformation #3.

Tables S4 (a)-(e). Automated fold recognition

1KP8 (A:2-136,A:410-526)	0.37	1KP8 (A:137-190, A:367-409)	0.37	1KID (A)	0.36
1KP8 (A:137, A:367-409)	0.37	1KID (A)	0.36	1HF2 (A:100-206)	0.32
1KID (A)	0.36	1HF2 (A:100-206)	0.32	2IOJ (A:206-325)	0.32
2IOJ (A:206-325)	0.32	2IOJ (A:206-325)	0.32	1M1H (A:5-50, A:132-186)	0.32
1HF2 (A:100-206)	0.32	1M1H (A:5-50,A:132-186)	0.32	2HI6 (A:1-132)	0.31
1M1H (A:5-50,A:132-186)	0.31	2HI6 (A:1-132)	0.31	2DST (A:2-123)	0.31
2HI6 (A:1-132)	0.31	2DST (A:2-123)	0.31	1ASS (A)	0.30

Table S4 (a): This lists candidate domains, with associated scores, automatically picked by FOLD-EM for the simulated GroEL 10 Å map. Three domains were picked: equatorial (column 1&2; column 2 is the associated FOLD-EM generated score), apical (column 3&4), and the intermediate domain (column 5&6). The first row lists the three domains with best scores, which are finally chosen to build the CA model of the simulated map.

Map resolution (Å)	Transformation Error (RMSD Å)	Transformation Error (RMSD Å)	Transformation Error (RMSD Å)
5	0.48	1.4	0.63
10	0.49	1.41	0.63
15	0.54	1.4	0.66
20	0.73	1.44	0.68

Table S4 (b): RMSD error in docking, using FOLD-EM

Column 2: Fitting errors for the intermediate domain of GroEL (size: 90 residues),

Column 3: Fitting errors for the apical domain of GroEL (size: 182 residues),

Column 4: Fitting errors for the equatorial domain of GroEL (size: 249 residues).

The fittings are done onto simulated GroEL cryo-EM maps with resolution ranging from 5-20 Å (column 1).

1A7A (A:190-352)	5.35
1QY9 (A:130-297)	5.21
2FS2 (A:1-131)	5.14
1F00 (I:658-752)	4.84
2DI4 (A:406-607)	4.51

Table S4 (c): This lists candidate domains for the first domain, with associated scores, automatically picked by FOLD-EM for building the CA backbone of the Phi29 map. The correct domain 1F00 is ranked #4. After this domain is picked, the final domain (2FT1) is picked as the domain with best score among those which occupied the whole input cryoEM map together with the first picked domain 1F00.

1UF2 (C:1-147, C:301-421)	0.18	1UF2 (C:148-300)	0.15
1UF2 (C:148-300)	0.15	1WN0 (B:11-138)	0.12
1WN0 (B:11-138)	0.12	1RCU (A)	0.08
1RCU (A)	0.08	1SUM (B)	0.06
1SUM (B)	0.06	4AIG (A)	0.06

Table S4 (d): This lists candidate domains, with associated scores, automatically picked by FOLD-EM for building the CA backbone of this RDV map. Two domains were picked: P8 (column 1&2; column 2 is the associated FOLD-EM generated score), P8 top domain (column 3&4). The first row lists the three domains with best scores, which are finally chosen by FOLD-EM to build the CA model of the simulated map.

1YAR (H:1-203)	3.29	1YAR (H:1-203)	2.56	1YAR (H:1-203)	2.48
1HQY (A)	2.59	1HQY (A)	2.50	1HQY (A)	2.46
1YAR (H:1-203)	2.56	1YAR (H:1-203)	2.48	1HQY (A)	2.41
1HQY (A)	2.50	1HQY (A)	2.46	1IAZ (A)	2.00
1YAR (H:1-203)	2.48	1HQY (A)	2.41	1RVV (A)	1.99

Table S4 (e): This lists candidate domains, with associated scores, automatically picked by FOLD-EM for building the CA backbone of this 20S map. Three domains were picked: 1YAR (column 1&2; column 2 is the associated FOLD-EM generated score), 1YAR (column 3&4), and 1YAR (column 5&6). The first row lists the three domains with best scores, which are finally chosen by FOLD-EM to build the CA model of the simulated map.

Text S1 | Validation of P22 results (shown in Figs. 2 (h-o))

We claim that alignment obtained by FOLD-EM (Fig. 2 (n)) is better than in (10) (Fig. 2(o); obtained using FOLDHUNTER (9)) using these two means:

(a) visual inspection: in Fig. 2 (o), we circled the regions where local alignment can be clearly seen (by eyes) as worse than in corresponding regions in Fig. 2 (n).

(b) automated scoring: FOLD-EM had a better alignment score (obtained using Chimera (7)) of 0.91 compared to 0.87 obtained by FOLDHUNTER.

FOLD-EM improves the alignment obtained using FOLDHUNTER by RMSD of 2.8 Å.

FOLD-EM was able to improve the alignment of the P22 subunits done in (10) (using FOLDHUNTER), because it is able to automatically separate the conserved base domain from the rest in the two subunit maps. FOLDHUNTER has no means to the separate conserved (base) and non-conserved regions, and hence looses it's accuracy due to inclusion of non-conserved regions while trying align the subunits by their conserved base. Also, very importantly, FOLDHUNTER needed an initial approximate alignment guess, whereas FOLD-EM didn’t.

Text S2 | Fold recognition in ribosome 70S

The two domains 50S and 30S of the ribosome 70S do not exist in the SCOP database. So, the way we came up with the fold shown in Fig. 4 (f) is as follows. We extracted the two low resolution domains (50S and 30S) from the 70S conformation #1 from (12), by comparing it with 70S conformation #2 from (12) using the **P** solver which is described in the SI Method. We included the extracted domains along with other domains in the SCOP to build the fold for the conformation #2 using FOLD-EM. As expected, conformation #2 scored the best against the the two extracted domains. The authors of (12) have also released the high resolution models for these domains, which we used to construct the final fold shown in Fig. 4 (f). The point of this testcase of building a fold for 70S was to show that FOLD-EM is applicable to real cryo-EM maps with resolutions as low as 13 Å. In the future, we would like to also test FOLD-EM on real cryo-EM maps with resolutions worse that that.

Video S1 LEGEND

Conformation change between the two ribosome conformations, as deduced by FOLD-EM (Fig. S5).

<ps: The video is an ANIMATED GIF AND it could not be uploaded. So it is available at: http://cs.stanford.edu/~mitul/foldEM/SI_Video_1.gif
>