Structure
Article

EM-Fold: De Novo Folding of α-Helical Proteins Guided by Intermediate-Resolution Electron Microscopy Density Maps

Steffen Lindert,[1,3] René Staritzbichler,[1,3] Nils Wötzel,[1,3] Mert Karakaş,[1,3] Phoebe L. Stewart,[2,3] and Jens Meiler[1,3,*]
[1]Department of Chemistry, Vanderbilt University, Nashville, TN 37212, USA
[2]Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[3]Center for Structural Biology, Vanderbilt University, Nashville, TN 37212, USA
*Correspondence: jens.meiler@vanderbilt.edu
DOI 10.1016/j.str.2009.06.001

SUMMARY

In medium-resolution (7–10 Å) cryo-electron microscopy (cryo-EM) density maps, α helices can be identified as density rods whereas β-strand or loop regions are not as easily discerned. We are proposing a computational protein structure prediction algorithm "EM-Fold" that resolves the density rod connectivity ambiguity by placing predicted α helices into the density rods and adding missing backbone coordinates in loop regions. In a benchmark of 11 mainly α-helical proteins of known structure a native-like model is identified in eight cases (rmsd 3.9–7.9 Å). The three failures can be attributed to inaccuracies in the secondary structure prediction step that precedes EM-Fold. EM-Fold has been applied to the ~6 Å resolution cryo-EM density map of protein IIIa from human adenovirus. We report the first topological model for the α-helical 400 residue N-terminal region of protein IIIa. EM-Fold also has the potential to interpret medium-resolution density maps in X-ray crystallography.

INTRODUCTION

Since the first subnanometer (<10 Å) resolution cryo-EM single-particle reconstructions, determined for the hepatitis B virus capsid in 1997 (Bottcher et al., 1997; Conway et al., 1997), there have been an increasing number of structures determined by cryo-EM in the 6–10 Å resolution range (Booth et al., 2004; Martin et al., 2007; Min et al., 2006; Saban et al., 2006; Serysheva et al., 2008; Villa et al., 2009; Zhang et al., 2003). For example, Saban et al. determined a 6.9 Å resolution structure of adenovirus, Booth et al. reached 9 Å resolution for cytoplasmic polyhedrosis virus, and Zhang et al. elucidated a 7.6 Å resolution structure of reovirus. Because only a fraction of the viral proteins are amenable to structure elucidation by X-ray crystallography, these experiments yield images of viral proteins of previously unknown structure. Cryo-EM can also elucidate the structures of large macromolecular complexes such as blue copper protein hemocyanin (Martin et al., 10 Å resolution), elongation factor Tu-ribosome complex (Villa et al., 6.7 Å resolution), and tetraspanin uroplakins (Min et al., 6 Å resolution). In these cases the density map revealed previously unknown crucial interfaces between subunits of the macromolecular complex. Cryo-EM has also been used to elucidate subnanometer structures of membrane proteins such as the skeletal muscle Ca^{2+} release channel (Serysheva et al., 9.6 Å resolution). Several near-atomic resolution structures (<5 Å resolution) have been determined recently using cryo-EM (Jiang et al., 2008; Ludtke et al., 2008; Yu et al., 2008; Zhang et al., 2008). Although near-atomic resolution maps show details such as β sheets and large side chains (Zhou, 2008), these features cannot be identified reliably at intermediate resolution. However, α helices are resolved as density rods at intermediate resolution (Lindert et al., 2009).

One of the biggest challenges for the interpretation of medium-resolution density regions remains the building of a correct topological model. It is impossible to "thread" the primary sequence through the density map for regions that are assigned to a protein of unknown structure because the connectivity between the density rods cannot be discerned at intermediate resolution. Thus it is not possible to assign particular density rods to specific α-helical regions of the sequence. Even if this obstacle could be overcome, missing loop regions and side-chain coordinates need to be built to arrive at an accurate atomic model.

Several computational tools are available that help in the analysis of cryo-EM density maps. If a high-resolution structure for the map or parts of the map is available, fitting techniques are frequently employed (Rossmann, 2000; Tama et al., 2004a, b; Topf et al., 2008; Topf and Sali, 2005; Trabuco et al., 2008; Velazquez-Muriel and Carazo, 2007; Velazquez-Muriel et al., 2006; Volkmann and Hanein, 1999; Wriggers et al., 1999). If no high-resolution structures are available for fitting, medium-resolution density maps can be interpreted in terms of the α helices that can be seen in the map. α-Helical regions can be identified either manually as rods within the density map, or automatically by methods using segmentation and feature extraction (Dal Palu et al., 2006; Jiang et al., 2001). The skeletonization algorithm in Baker et al. (2007) identifies secondary structure elements and suggests a possible secondary structure topology by connecting density rods based on increased density in short loop connections. A protocol that iteratively improves comparative models by fitting these models into cryo-EM density maps is reported (Topf et al., 2006). This method requires the presence of a comparative model but is independent of the identification of α-helical regions in the density map. Models built with the de

novo protein structure prediction software Rosetta were ranked with respect to their agreement with the cryo-EM density maps using a two-way distance measure (Baker et al., 2006). This approach eliminates the need for an initial comparative model, but it has the drawback that the Rosetta calculation is not driven by the experimental density map. Therefore, the approach only works if Rosetta is capable of folding the protein correctly de novo, which is possible for proteins with up to 150 amino acids (Bonneau et al., 2002b).

De novo protein structure prediction algorithms have experienced considerable improvements during the last ten years. The software Rosetta has been demonstrated to correctly predict the fold of proteins with up to 150 amino acids (Bonneau et al., 2002b; Moult, 2005; Rohl et al., 2004b; Simons et al., 1997, 1999). Structurally variable loop regions up to 12 residues long can be modeled routinely with Rosetta (Rohl et al., 2004a). More recently, iterative side-chain repacking and backbone reconstruction protocols within Rosetta have been shown to refine initial de novo and comparative models to atomic-detail accuracy (Bradley et al., 2005; Misura and Baker, 2005; Misura et al., 2006; Schueler-Furman et al., 2005). For instance, with a benchmark of 16 small proteins (49–88 residues), Bradley et al. demonstrated that accurate atomic-detail models (<1.5 Å) could be reached from initial de novo models for five proteins.

It has been demonstrated that guiding the de novo protein structure prediction technique Rosetta with low resolution or sparse experimental data yields structural models with accurate atomic detail. Inclusion of nuclear magnetic resonance data within RosettaNMR has improved the quality of created atomic models (Bowers et al., 2000; Meiler and Baker, 2003b, 2005; Qian et al., 2007; Rohl and Baker, 2002). Similarly, EPR data have been combined with Rosetta for enhanced model building (Alexander et al., 2008; Hanson et al., 2008).

The approach presented in this article combines computational structure prediction methods with experimental cryo-EM density maps to build topological models for large proteins without an atomic resolution structure or an available comparative model. The algorithm first identifies α-helical regions in the density map and in the protein's primary sequence, utilizing a consensus secondary structure prediction protocol. The predicted α helices are placed into specific α-helical density rods of the density map using a novel Monte Carlo assembly algorithm. Then loop regions and side-chain coordinates are added using Rosetta's iterative side-chain repacking and backbone reconstruction protocols to arrive at a model with atomic detail present.

Currently, EM-Fold is tailored toward α-helical proteins because β strands are typically not well resolved in medium-resolution density maps. β strands become visible at 5–7 Å resolution (Lindert et al., 2009). We plan a future development stage of EM-Fold that simultaneously assembles α helices and β strands. This method will be implemented during the next several years as more density maps become available that have both types of secondary structural elements resolved.

Here we present the results of EM-Fold with ten mainly α-helical benchmark proteins and simulated cryo-EM density, as well as with experimental cryo-EM density maps of bovine metarhodopsin and adenovirus protein IIIa. In the case of metarhodopsin, the EM-Fold models are compared with the atomic resolution structure of rhodopsin.

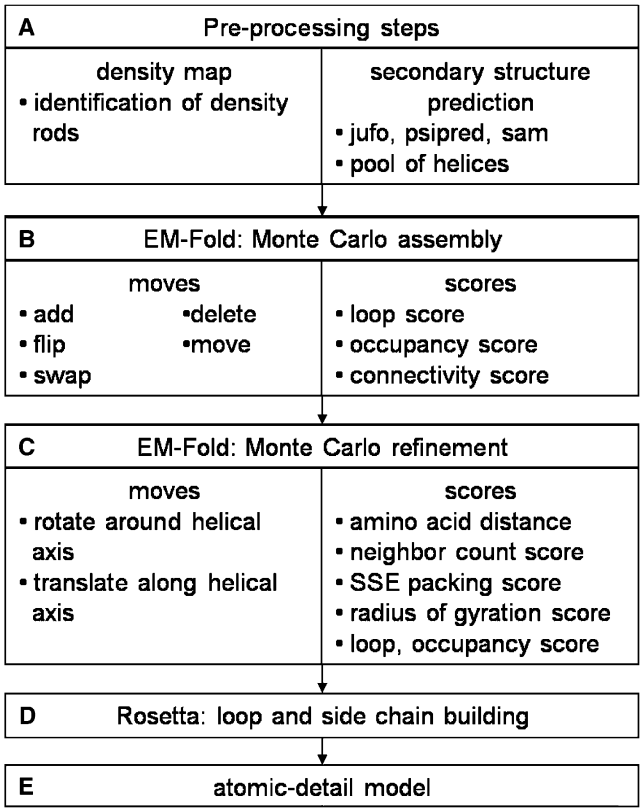

Figure 1. Flowchart of the Entire Protocol

(A) Density rods are identified in a medium resolution density map. A pool of α helices is built using secondary structure prediction algorithms.

(B) The assembly step of EM-Fold places α helices from the pool into density rods.

(C) An EM-Fold refinement step improves the placement of α helices within the density rods.

(D) Loops and side chains are built in Rosetta for the best of the refined EM-Fold models.

(E) One of the final full-atom models is likely to be very close in rmsd to the native structure.

RESULTS AND DISCUSSION

Benchmark Database of Ten α-Helical Proteins with 250 to 350 Residues

To test the reliability as well as to optimize the parameters of the proposed assembly algorithm EM-Fold, it has been benchmarked on ten proteins of known structure following the protocol outlined in Figure 1. The proteins were chosen to be mostly α-helical (60%–68%) and of substantial size (255 to 347 residues) (see Table S1 available online). Except for one protein (1OUV), all the benchmark cases possess contact orders of 40 or higher. Thus these proteins constitute complex folds, making de novo computational structure prediction challenging (Bonneau et al., 2002a). In order to mimic cryo-EM density maps, simulated density maps at 6.9 and 9.0 Å resolution were generated for each of the ten proteins. The positions and lengths of the density rods are virtually indistinguishable at both resolutions. The maps, however, differ by the information they contain in loop regions as well as in delineation of the density rods. The

Table 1. Overview of the Benchmark with Ten α-Helical Proteins

Protein	Rank Assembly[a]	Rmsd Assembly [Å][b]	Rank Refinement[c]	Rmsd Refinement [Å][d]	Rank Loop[e]	Rmsd Loop [Å][f]	α Helices in Final Partial Model[g]
1IE9	1 (1)	3.7 (3.3)	5 (1)	3.7 (2.6)	1 (1)	5.9 (7.8)	4 [4]
1N83	1 (1)	6.2 (3.2)	2 (1)	5.9 (2.4)	1 (7)	7.1 (3.7)	5 [5]
1OUV	6 (10)	3.0 (3.1)	4 (6)	2.9 (2.3)	1 (1)	4.3 (4.8)	9 [9]
1QKM	16 (1)	3.6 (3.1)	2 (1)	2.7 (3.3)	2 (7)	3.9 (4.2)	5 [5]
1TBF	100 (8)	3.1 (3.2)	20 (17)	2.8 (2.7)	1 (3)	4.1 (4.2)	12 [11][h]
1V9M	— (1)	— (3.3)	— (1)	— (2.0)	— (2)	— (6.7)	7 [4]
1XQO	— (2)	— (3.3)	— (7)	— (2.1)	— (1)	— (5.0)	6 [2]
1Z1L	150 (3)	3.1 (3.4)	72 (13)	3.2 (2.5)	1 (1)	5.9 (5.5)	9 [9]
2AX6	1 (1)	4.0 (3.4)	5 (1)	3.2 (3.4)	3 (8)	6.6 (9.2)	5 [5]
2CWC	— (2)	— (2.9)	— (8)	— (2.4)	— (2)	— (7.1)	3 [0]
Rhodopsin	2	3.4	1	3.1	1	7.9	—

Results are shown for both realistic secondary structure prediction, as well as for perfect secondary structure prediction (in parentheses).
[a] Rank of true model after assembly step.
[b] Rmsd of backbone atoms in α helices of true model after assembly step (compared with PDB coordinates).
[c] Rank of true model after refinement step.
[d] Rmsd of backbone atoms in α helices of true model after refinement step.
[e] Rank of true model after loop-building step.
[f] Rmsd of all atoms in true model after loop-building step.
[g] Number of α helices in final partial model based on 50% consensus placement; the number of correctly placed α helices in these partial models is shown in square brackets. These results are also depicted in Figure 4.
[h] The one α helix in the partial model of 1TBF that has not been correctly placed has been placed into the correct density rod, but with antiparallel orientation.

benchmark was performed in two stages depending on the type of secondary structure information used, either the correct secondary structure derived from the atomic resolution structure or a realistic prediction of secondary structure, which can deviate from the true structure.

100% Success Rate for the Perfect Secondary Structure Prediction Benchmark

In a first test, 20,000 models were built for each of the ten benchmark proteins using the correct secondary structure. The Monte Carlo simulation was run until a total of 2000 subsequent steps were rejected with no improvement in the overall score. The agreement with the density, which is simulated for the benchmark proteins, is assessed by an occupancy score (Figure S1), a loop score, and a connectivity score (see Figure S2). A predicted fold is considered correct if all α helices have been placed in the appropriate simulated density rods with the correct orientation of the α-helical axis. A high rank for the correct fold among the 20,000 models generated indicates success of the protocol.

The true model is found among the best ten scoring models for all the benchmark cases (Table 1). In 50% of the cases the true model is ranked first. In the cases where the true model is not ranked first, the better ranking models are similar in topology to the true model and frequently only have a single α helix or a pair of α helices in an incorrect orientation. This demonstrates that the assembly step can clearly distinguish native-like from non-native models if the correct secondary structure is used as input. The root-mean-square deviations (rmsds) of the correct topology models range between 2.9 Å and 3.4 Å over the α-helical residues (Table 1).

For each of the ten proteins, the 50 best scoring models from the assembly step were refined. In this process a wider variety of types of scores (described in Experimental Procedures) is used to evaluate the models. After refinement the rmsds of the best scoring correct topology model range between 2.0 Å and 3.4 Å, again considering only the α-helical residues, and the true model is found among the best 17 scoring models (Table 1). These rankings are within the accuracy limit of the scoring functions.

Rosetta was used to build loops for the 20 best scoring models after the refinement run. The rmsd of the true model after loop building ranges between 3.7 and 9.2 Å (Table 1), which is an excellent level of agreement for de novo models considering the large size of the proteins. After the loop-building step, all of the true models are ranked within the best eight scoring topologies according to the Rosetta score. Thus, EM-Fold is able to identify the true topology within the top ten best scoring models built, given completely correct secondary structure information.

EM-Fold Selects the Best α Helices from a Consensus Pool Generated from State-of-the-Art Secondary Structure Predictions

A combination of three state-of-the art secondary structure prediction programs jufo (Meiler and Baker, 2003a; Meiler et al., 2001), psipred (Jones, 1999), and sam (Chandonia and Karplus, 1999; Karplus et al., 1997) was used to simulate a realistic prediction scenario. The utilization of different programs avoids usage of incorrect secondary structure if one of the methods fails. Wherever an α helix is predicted with a probability of higher than 0.5 for more than nine subsequent residues, this α helix is inserted into the pool of considered secondary structure elements. Smaller α helices are ignored because these

cannot be confidently identified in intermediate resolution density maps. Further, a consensus prediction (average of all three methods) and a consensus prediction where α helices longer than 21 residues are broken into two smaller α helices are included. Within the ten benchmark proteins there are 93 α helices that have at least 12 residues. Each of these α helices is identified by at least one secondary structure prediction technique, although the predicted lengths and confidence levels differ.

Secondary structure predictions tend to yield α helices that are too short, thus three different pools (A, B, and C) of secondary structure elements were tested including lengthened α helices in pools B and C (see Experimental Procedures). The best results for the assembly step are obtained with the most diverse pool of secondary structure elements (pool C), where the average deviation between predicted and correct α-helix length is only 0.4 residues per α helix (Table S2). This finding stresses two points: (1) The more accurate the secondary structure prediction is, the better the results of the assembly algorithm will be—a finding that is also supported by the benchmark test using the correct secondary structure information. (2) A larger pool, which includes many inaccurate secondary structure elements, does not negatively influence the success of the assembly protocol. In other words, the assembly protocol identifies and uses the best possible secondary structure elements available in the pool. Only pool C was used for the realistic secondary structure benchmark because it has been demonstrated to most accurately represent the secondary structure of the proteins.

De Novo Folding of α-Helical Benchmark Proteins with Realistic Secondary Structure Predictions

In the initial assembly step (see Figure 1B), 60,000 models were built for each protein using the most diverse secondary structure pool (pool C). Building one model takes approximately 60 s on a single JS20 IBM 2.2GHz PowerPC. The models were ranked by score (Table 1). Our results indicate that despite the inaccuracies of secondary structure prediction, after the assembly step the true model is found among the best 150 scoring models for seven of the ten proteins. In particular, for four of the benchmark proteins the true model is found among the best ten scoring models, and the average rank of the seven correct models is 39. The rmsd of the correct model after the assembly step ranges from 3.0 to 6.2 Å (Table 1). The best 150 models by score enter the refinement protocol without manual analysis.

After refinement (see Figure 1C), the ranking of the correct model improves to at least rank 72, for five of the benchmark cases it even improves to rank 5 or better. Further, the quality of the true model, as assessed by the rmsd, improves for five of seven cases with a range over all seven proteins of 2.7 to 5.9 Å (Table 1). Figure S3 illustrates the improvement of α-helix orientations during the refinement step for three examples. The best 75 models by score enter the loop-building protocol without manual analysis.

Loops are built for the best 75 scoring models after refinement. For each of the 75 refined models 100 loop models are built using Rosetta. After ranking of these 7500 models according to their Rosetta score, the true model is within the best three scoring models for all seven proteins (see Table 1). Even though the average rank of the correct model after the assembly step was 39, the user only needs to consider the top three scoring models after loop building. The accuracy of these models is in the range of 3.9 to 7.1 Å (Table 1). This rmsd range is comparable to those built with correct secondary structure elements and acceptable considering the large size of the proteins. Superimpositions of the final Rosetta model with the native structure are shown for all seven proteins (Figure 2).

Consensus Placement of α Helices Correlates with Correct Positioning and Can Be Used as a Measure of Confidence

In order to develop a measure that is independent of the score and that can evaluate the correctness of a particular model, the consensus placement of α helices into specific density rods was analyzed. Models after the assembly step and after loop construction were evaluated. In both cases, the benchmarks indicate that if a specific α helix is found repeatedly in the same density rod within the set of best scoring models it was placed correctly. Receiver operator characteristic (ROC) curve representations for placement confidence after the assembly and loop-building steps are shown in panels A and B of Figure 3. The total areas under the curve are 0.81 and 0.86, respectively, indicating strong correlations between frequent placement and correct positioning. For example, a placement of a particular α helix into a specific density rod that is found in 70% of the top scoring models after the assembly step has a 71% confidence level of being correct. The results for models after the loop building step are even better, corroborating the ability of the algorithm to enrich for true-topology models. For example, a placement of a particular α helix into a specific density rod that is found in 50% of the top scoring models after the loop building step has an 82% confidence level of being correct.

It would be desirable if the confidence measure allowed distinction between successful and unsuccessful cases in the benchmark. Partial models containing only the α helices placed with a > 50% repetition rate were built for all ten benchmark proteins. A 50% cutoff ensures that no other placement into that density rod can occur more frequently. We evaluated the overall confidence in a model where k α helices have been placed confidently out of a total of n α helices by calculating the number of possibilities to place k α helices into a total of n density rods ($2^k \times n!/(n - k)!$). This equation explicitly takes into account the number of confidently placed α helices (k) and the total number of α helices in the protein (n), and implicitly the fraction of confidently placed α helices. It also accounts for the fact that placing a specific fraction of α helices confidently in a large protein is considerably less likely than placing the same fraction of α helices confidently in a smaller protein. The results of this analysis are plotted in Figure 4. The overall confidence scores for the ten benchmark proteins fall into two regions within this plot. Some proteins have a low number (3–7) and others have a high number (10–14) on this scale (separated by the dashed line in Figure 4). Proteins below the dashed line contain both successful and unsuccessful cases indicating that there is ambiguity for partial models in this range. However, proteins in the upper region (above the dashed line) contain only successful benchmark cases, suggesting that a high value on this overall confidence scale identifies correct topologies. Interestingly, the partial model that we built for adenovirus protein

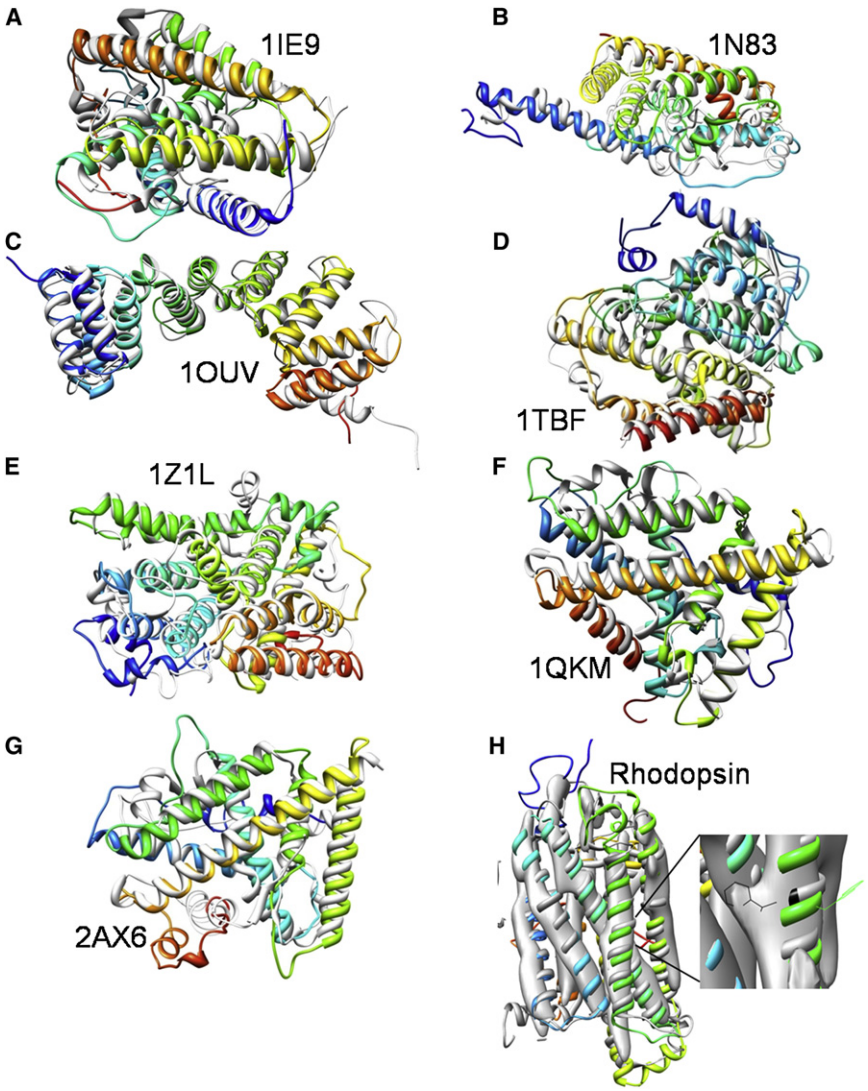

Figure 2. Comparison of the Computational Models with the Crystal Structures

Superimposition of the Final Models (Colored in Rainbow) of 1IE9 (A), 1N83 (B), 1OUV (C), 1TBF (D), 1Z1L (E), 1QKM (F) and 2AX6 (G) with the Original PDB Structures (gray). These proteins range in size from 255 to 345 residues. The displayed models have rmsds ranging from 3.9 Å to 7.1 Å compared with the PDB structure. Regions that are only seen in the models (such as the N terminus of 1TBF) correspond to parts of the protein that are missing in the PDB file. Panel (H) shows the model of rhodopsin after the loop building step (rainbow) in the experimental density model. The crystal structure of rhodopsin is shown in gray for comparison. The model and crystal structure have an rmsd of 7.9 Å. A blow-up of one Trp side chain and its corresponding density bump is shown. The Trp side chain of the crystal structure is shown in black for comparison. It is apparent that the Trp in the model was placed in the correct height of the density rod. The rotation of the α helix in the model is off by about 150°. This is not unexpected and could be corrected by a subsequent refinement protocol.

IIIa (discussed below) clusters with the benchmark proteins in the high range of this scale. This gives credence to the protein IIIa model in the absence of an atomic resolution structure.

Poor Secondary Structure Prediction Leads to Poor Assembly Results

The three proteins that were not successfully assembled have the poorest secondary structure prediction with an average deviation of 0.8 residues per α helix in pool C, compared with an average deviation of 0.3 residues per α helix for the remaining seven proteins (Table S2). This underscores the fact that failure to find the true solution is not a shortcoming of the assembly algorithm but rather a result of suboptimal secondary structure prediction. The correct solution of 2AX6 is found despite its poor secondary structure prediction (average deviation of 0.8 residues per α helix in pool C) because this protein is small with only six α helices. In this case, the assembly algorithm has to probe a considerably smaller search space and thus can overcome the limitation of poor secondary structure information.

Rosetta Iterative High-Resolution Refinement Achieves Accurate Atomic-Detail in Parts of the Protein Models

One of the main challenges of computational protein structure prediction is recovering accurate atomic detail of interfaces within proteins. The top ten scoring loop models of all the seven proteins where the correct topology was identified after loop building were subjected to an iterative Rosetta refinement protocol (see Experimental Procedures). The objective of this protocol was to test the ability of the method to build accurate atomic-detail structural models at least in part of these proteins. Further, it was investigated whether it is possible to uniquely identify the correct topology by the Rosetta energy score.

Figure 5 shows close-up views of three α-helix-helix interfaces in the best scoring correct topology model for 1QKM after iterative high-resolution refinement. The protocol was able to recover native side-chain packing in some of the α-helical interfaces (Figures 5A and 5B). However, even in the best scoring model there are still interfaces that are not recovered (Figure 5C). Figure S4 shows the total full-atom Rosetta energy plotted versus the rmsd of the model for all of the proteins. Although low RMSD models cannot be identified solely by energy, in six of seven cases the correct topology can be identified by its enrichment in the 10% model with lowest energy (7.6 for 1Z1L, 4.0 for 1IE9, 3.8 for 1OUV, 2.6 for 1QKM, 1.6 for 1TBF, and 1.2 for 1N83). We hypothesize that these enrichments are due to lower energy (higher quality) of the fraction of α-helical interfaces that were built accurately at atomic detail. At the same time, non-native α-helix interfaces introduce a background noise that make the energy of models with correct topology often comparable to those of incorrect topology.

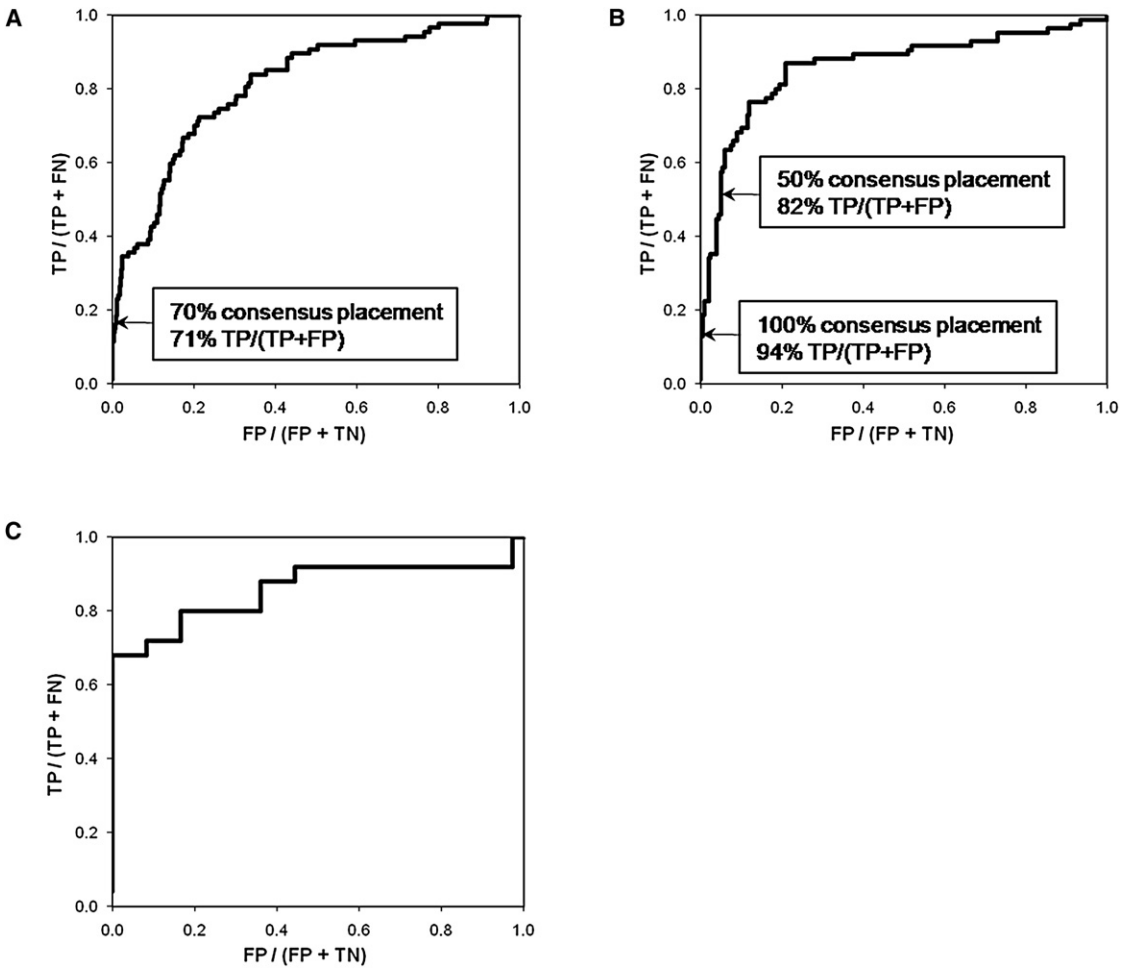

Figure 3. ROC Curves for the Confidence in Repeated Placements and the Performance of the Connectivity Score

(A) ROC curve of the confidence in placements of single α helices into density rods based on repeated placements after the assembly step. The fraction of correct placements (true positives/[true positives + false negatives]) over the fraction of wrong placements (false positives/[false positives + true negatives]) is plotted. The connection between repetition rate and placement confidence has been added to the ROC curve. For example, a placement of a particular α helix into a specific density rod that is found in 70% of the top scoring models after the assembly step has a 71% confidence of being correct. The area under the curve is 0.81 where 0.5 represents a random measure.

(B) ROC curve of the confidence in placements of single α helices into density rods based on repeated placements after the loop-building step. The fraction of correct placements (true positives/[true positives + false negatives]) over the fraction of wrong placements (false positives/[false positives + true negatives]) is plotted. The connection between repetition rate and placement confidence has been added to the ROC curve. For example, a placement of a particular α helix into a specific density rod that is found in 50% of the top scoring models after the loop building step has a 82% confidence of being correct. The area under the curve is 0.86, where 0.5 represents a random measure.

(C) ROC curve of the connectivity score. The fraction of correct connections (true positives/[true positives + false negatives]) over the fraction of wrong connections (false positives/[false positives + true negatives]) is plotted. The steep increase at the beginning demonstrates that the strongest correct connections score all better than any of the wrong connections. The area under the curve is 0.86, where 0.5 represents a random measure.

For all seven proteins, the native structure obtained from the Protein Data Bank (PDB) was minimized in the refinement protocol as well (Figure S4). Its energy is clearly lower than the energy of any of the models built. Thus the absence of models that have accurate atomic detail throughout the entire protein chain is a sampling rather than a scoring problem. This is expected for de novo protein models of 250 and more residues. The size of these systems far exceeds the 90 residue practical limit for de novo high-resolution structure prediction (Bradley et al., 2005). However, our finding of native-like α-helix interfaces in portions of these models is an encouraging result that suggests that all-atom accurate atomic-detail models can be achieved as cryo-EM reaches higher resolution, and as computational techniques improve.

Comparison of EM-Fold with a Computational Prediction Method for α-Helical Membrane Proteins

In 2007, Kovacs et al. introduced a protocol for predicting atomic-resolution details for α-helical membrane proteins guided by EM density maps (Kovacs et al., 2007). This method uses scripts within the internal coordinate mechanics (ICM) software environment. The ICM-based approach was demonstrated with simulated EM density maps at intermediate resolution for three

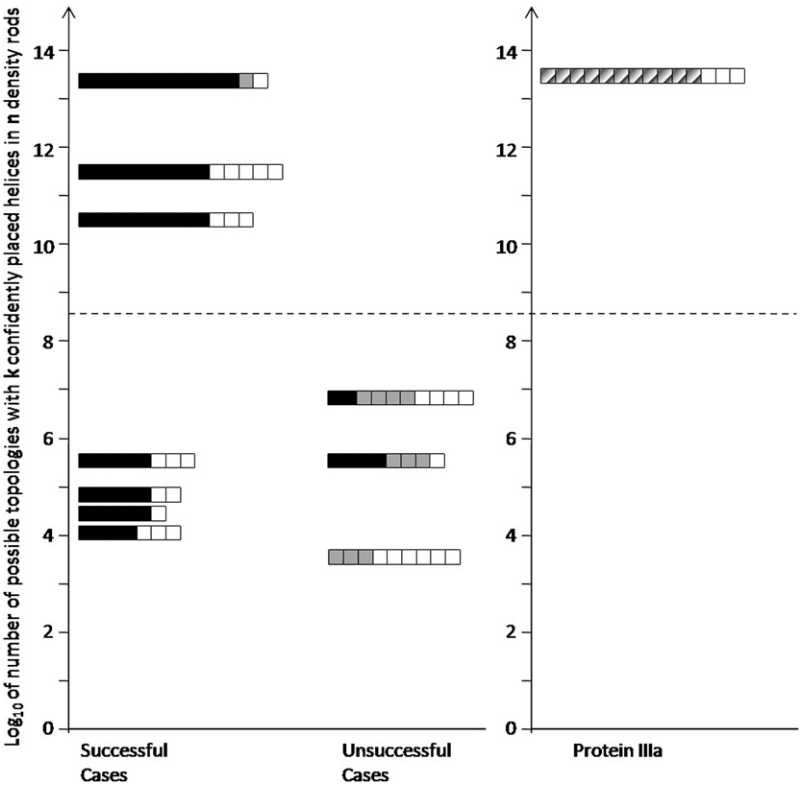

Figure 4. EM-Fold Results for the Ten Benchmark Proteins and Adenovirus Protein IIIa Evaluated on the Basis of the Number of Confidently Placed α Helices and the Total Number of α Helices in the Protein

The *y* axis represents the log base 10 of the number of possible topologies with **k** confidently placed α helices in **n** density rods using the following equation: $(2^k \times n!/(n - k)!)$. The length of each bar in the plot corresponds to the total number of α helices in a protein (**n**). The sum of the black and gray squares within a bar represents the number of α helices that were confidently placed by EM-Fold (i.e., with > 50% repetition rate) (**k**). Within the subset of confidently placed α helices, the correctly placed α helices are in black. The ten benchmark proteins split into two groups as indicated by the dashed line: those with a low number (3–7) and those with a high number (10–14) on this scale. A high number indicates a low probability of confidently placing these α helices by chance. Although there are both successful and unsuccessful benchmark cases below the dashed line, only successful cases are found above the line. For adenovirus protein IIIa, 11 of 14 α helices are confidently placed by EM-Fold (diagonal pattern, **k**) and the *y* axis number is well above the dashed line.

membrane proteins (GpA, KcsA, MscL). ICM-based flexible fitting of α helices, optimization of side-chain conformations, and refinement of atomic models resulted in impressive final rmsds between 0.9 and 1.9 Å for the three test membrane proteins.

Although the general idea of guiding protein structure prediction by α-helical density rods observed in intermediate-resolution EM density maps is the same for the ICM-based method (Kovacs et al., 2007) and EM-Fold, there are substantial differences between the methods. In the demonstration of the ICM approach, perfect secondary structure prediction was assumed. We have tested EM-Fold with both perfect and realistic secondary structure prediction information including variations in α-helix lengths. Second, the test proteins used in the ICM demonstration are sufficiently small (with one or two α helices per monomer), and have α helices of differing lengths (in the case of two α helices per monomer), so that the assignment of α helices into specific density rods is trivial. The centerpiece of the EM-Fold protocol is the assembly step (Figure 1B), which is designed to identify the topology of a protein from its α-helical secondary structure prediction and the positions of density rods in the density map. Subsequent steps (Figure 1C and D) refine the model. The ICM-based algorithm does not have an assembly step, whereas the refinement steps in both protocols follow similar principles. In their current setups these algorithms are complementary, and it is conceivable that models derived from EM-Fold could be input into ICM for further refinement.

Benchmark of EM-Fold on Experimental Bovine Metarhodopsin Density Map

To demonstrate EM-Fold's ability to work reliably in conjunction with experimental data, we built a model for bovine metarhodopsin based on the 5.5 Å resolution cryo-EM density map obtained from the Electron Microscopy Data Bank (EMDB) database (Ruprecht et al., 2004). The crystal structure of bovine rhodopsin (PDB ID 1GZM [Li et al., 2004]) was used to evaluate the results. The crystal structure is in a different conformational state than the cryo-EM structure. The overall fold of the protein is the same, however, because the authors note that the meta I formation involves no large movements or rotations of α helices from their ground state (Ruprecht et al., 2004). So although there might be structural differences in the loop regions, the α-helical regions that are modeled in the protocol are well described by the crystal structure. Interestingly, the authors report density bumps for several Trp side chains in the 5.5 Å resolution cryo-EM density map. Bovine rhodopsin is mostly α helical (63%) and slightly larger than the largest of the ten benchmark proteins (349 residues, Table S1).

The same protocol that was used for the ten benchmark proteins was applied to bovine metarhodopsin. The results are summarized in Table 1. The correct topology is ranked second after the assembly step and is ranked first after the refinement step. After the loop building step the correct topology is the best scoring model. This model has an rmsd of 7.9 Å to the crystal structure. If the crystal structure was not available, we could evaluate the EM-Fold results on the basis of the overlap between Trp side chains and Trp density bumps on rods. Only a single good scoring model has all of the Trp containing α helices in density rods with Trp density bumps. This model corresponds to the correct topology. These results demonstrate the ability of EM-Fold to work accurately in combination with experimental density maps. The rather large rmsd value is in part caused by the conformational change between crystal and cryo-EM

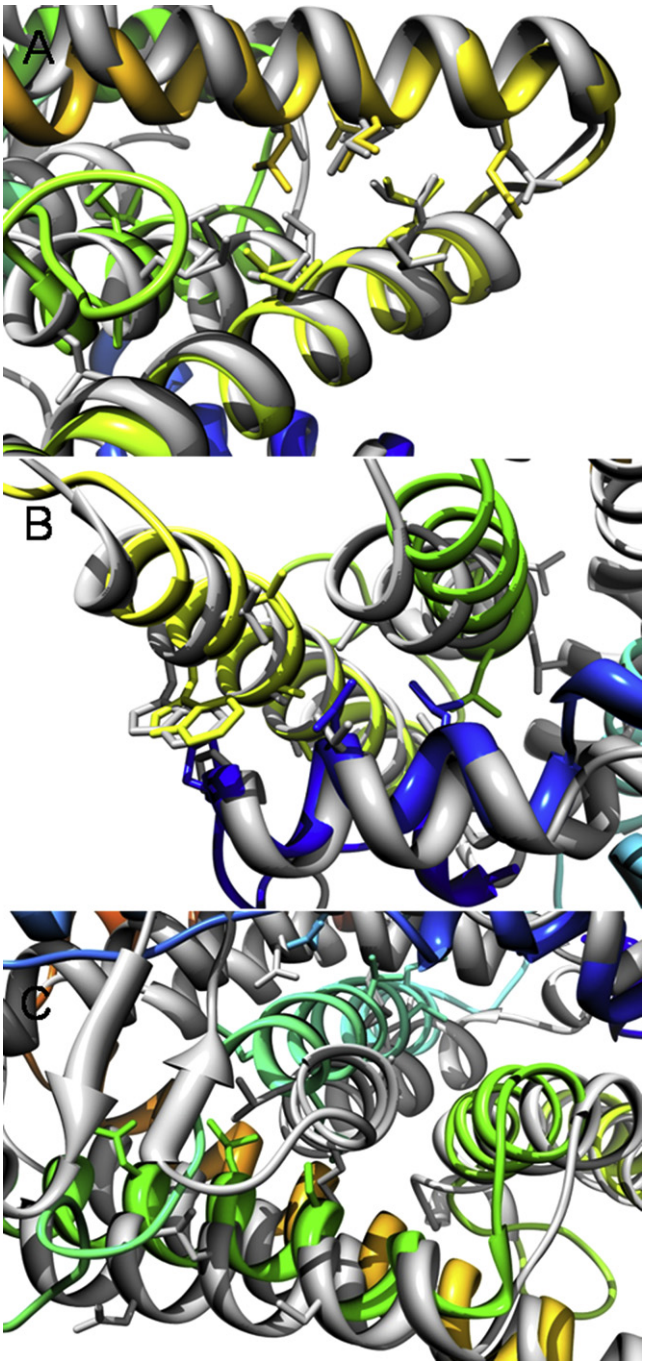

Figure 5. α-Helix-Helix Interfaces within the Best-Scoring, Correct-Topology, Full-Atom Model of Protein 1QKM after Rosetta Iterative High-Resolution Refinement

The full-atom model is shown in rainbow colors, whereas the native PDB is depicted in gray.

(A and B) Examples of near-native interfaces in the final model. The α-helix orientations and positions have been correctly identified and the side-chain conformations are generally close to the native PDB.

(C) An example of a α-helix-helix interface that could not be recovered.

structure, particularly in the loop regions. The rmsd over α-helical residues is only 3.1 Å, making this an excellent model for a protein of this size.

Evaluation of Adenovirus Protein IIIa Folds by EM-Fold

We have also applied EM-Fold to the medium-resolution cryo-EM density assigned to protein IIIa in the adenovirus capsid (Saban et al., 2006). In this case we do not have an atomic resolution structure for protein IIIa. This is a challenging case for EM-Fold because the α-helical region of protein IIIa is larger than any of the benchmark proteins and it has a two-lobe topology (Figure 6). This two-lobe density region contains 14 manually identified density rods and is assigned to the N-terminal 400 residues of protein IIIa, which are predicted to be highly α-helical. Because none of the ten benchmark proteins or rhodopsin have a two-lobe topology, this complication has not been tested in EM-Fold. Therefore, we used experimental information to assign the two lobes and to filter the models produced by EM-Fold.

In order to extend the resolution of the Ad35F cryo-EM structure, we increased the data-set size to a total of 7133 particle images and performed several additional rounds of Frealign refinement. The final Ad35F structure is based on 3040 particle images and has a resolution of 6.8 Å at the FSC 0.5 threshold (and 5.8 Å at the FSC 0.3, and 5.2 Å at the FSC 0.143 thresholds). A plot of the FSC for the refined map can be seen in Figure S5. The crystal structure of the Ad5 hexon reveals that there are two α helices of 10 or more residues that have a Trp (Rux et al., 2003). We observe prominent bumps for the Trp side chains on each of these two α helices in the 6.8Å cryo-EM density map (see Figure S6).

Using the criteria developed to identify Trp in hexon, three possible positions (in rods E, K, L) were identified in the protein IIIa that might correspond to a Trp side chain. The side-chain bump in rod E is at the end of the rod, whereas the bumps in rods K and L are both in the middle of the rods and in fact form a connection between these two rods. Analysis of the protein IIIa sequence indicates that there is only one Trp in a predicted α helix (residue 27) and that it corresponds to the first or second residue in the predicted α helix. This excludes rods K and L, as corresponding to the α helix with a Trp, because the observed bumps are in the middle of these rods. We hypothesize that the observed bumps in rods K and L belong to two aromatic side chains that are in contact. After analyzing the cryo-EM density, we conclude that the rod most likely to contain the predicted α helix with a Trp (amino acids 27–39) is rod E.

This lobe assignment for protein IIIa is in agreement with the N-terminal tagging experiment recently published (San Martin et al., 2008). The protein IIIa peptide tag study localizes the N terminus of protein IIIa to the inner capsid surface close to the interface between penton base and the peripentonal hexons. Specifically, the difference density attributed to an N-terminal FLAG tag on protein IIIa is observed in the vicinity of what we refer to as rod E in lobe 1 of protein IIIa (Figure 6). Therefore, both the analysis of the side-chain density and the protein IIIa N-terminal tagging information indicate that lobe 1 should be assigned to the most N-terminal portion of protein IIIa.

After applying the same EM-Fold protocol used for the ten benchmark proteins and rhodopsin, we analyzed the top 100

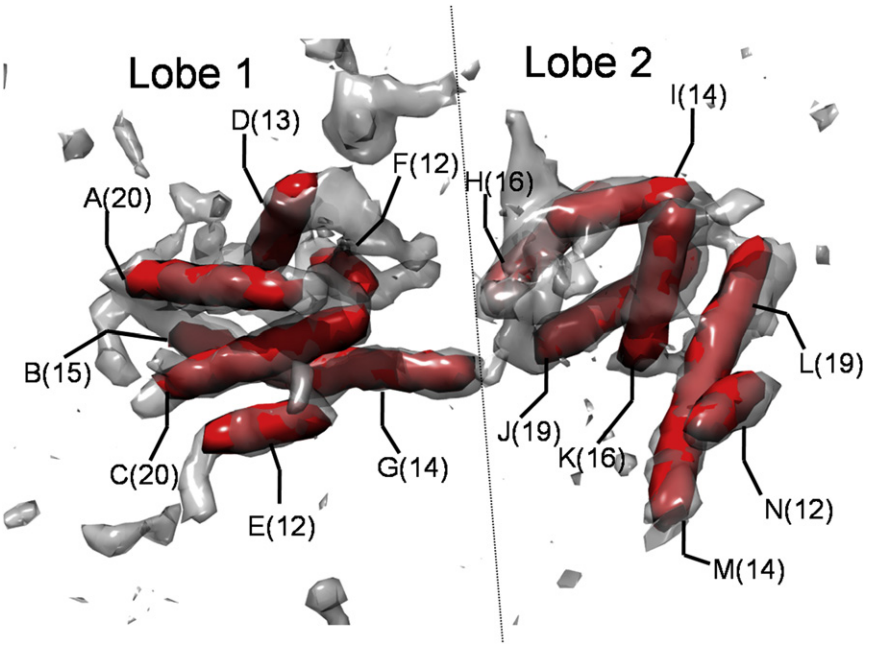

Figure 6. Experimental Cryo-EM Density Map of Adenovirus Protein IIIa Shown Segmented from an Adenovirus Reconstruction at 6.9 Å Resolution (FSC 0.5 Threshold)

Fourteen rods of minimum length 18 Å have been identified as α-helical regions (red). Each rod is labeled with a letter and the number of α-helical residues corresponding to its length. The EM-Fold assembly step involves placing α helices from the secondary structure prediction pool into the 14 identified density rods. The protein IIIa (gray) density has a two-lobe topology, with lobe 1 comprising rods A–G and lobe 2 comprising rods H–N. In the adenovirus capsid, lobe 1 is closer to the penton base.

models for protein IIIa and found that 33 of these have the N-terminal ~200 residues of protein IIIa positioned into lobe 1. A detailed analysis of this subset of models indicates that 14 models have the predicted α helix for residues 27–39, which includes the Trp at position 27, placed into rod E. We consider these 14 selected models the most likely models for protein IIIa. Within these 14 models, we note that four α helices (corresponding to residues 50–60, 70–83, 230–242, and 251–264) are placed into specific rods (G, B, H, and J, respectively) in all of the cases. Therefore, we assign these α helices, as well as the Trp-containing α-helix (rod E), as having a very high (>94%) confidence level. An additional six α helices are placed with > 50% repetition rate and thus are assigned a high (>82%) confidence level as labeled in the ROC curve in Figure 3B. A partial model of protein IIIa that contains these 11 confidently placed α helices is shown in rainbow in Figures 7A and 7B. The remaining three α helices are shown in gray and the loop regions are shown in white, indicating that their positioning within the density is more ambiguous. The number of confidently placed α helices puts this partial model into the confident region in Figure 4, further increasing the probability that it is correct. The proposed 50% confidence protein IIIa model is shown in context with penton base and two nearby peripentonal hexons (Figure 7C). Also, the agreement of the Trp (residue 27) side chain with the bump in rod E is shown in Figure 7D. Interestingly, one of the α helices placed with a high confidence level (rod L) contains a Tyr residue (Y369) in the middle of the α helix that corresponds to the density connection observed between rods K and L. On top of this another confidently placed α helix places Y299 in the middle of the connected density rod (rod K). This confidence assignment agrees perfectly with the observed density connection between rods K and L and gives further credence to our model. We anticipate that higher-resolution cryo-EM density revealing more of the side chains, combined with additional computational modeling, would resolve the remaining ambiguities in the protein IIIa fold model.

Conclusions

EM-Fold is a novel computational protein folding algorithm that assembles α-helical proteins guided by medium-resolution density maps. In a later stage, EM-Fold can be extended to include β strands in the assembly algorithm once more cryo-EM density maps allow an unambiguous identification of β strands. For future applications, manual identification of density rods will be replaced by an in-house algorithm that is currently under development. A benchmark on ten proteins shows a 100% success rate for the assembly of α helices when the correct secondary structure information is assumed. When predicted secondary structure information is used, which includes some incorrect information, the success rate drops to seven out of ten. Our results demonstrate that the 30% failure rate is linked to incorrect secondary structure prediction information, and future developments will include improving the secondary structure prediction input. This might be done by either improving the secondary structure prediction algorithms themselves or, as demonstrated here, by including more diverse predictions into a more complex pool of α helices prior to assembly. The final models generated by EM-Fold display rmsds in the range of 3.9 Å to 7.1 Å for the benchmark proteins. A complete model for rhodopsin with 7.9 Å rmsd could be built based on an experimental density map. These results demonstrate that de novo protein structure prediction can be extended to proteins well beyond 150 amino acids if the search is guided by medium-resolution density maps.

The iterative Rosetta refinement protocol did not completely succeed in refining the models to accurate atomic detail. Given the large size of the proteins this is not entirely surprising. However, portions of the final models, including specific α-helix-helix interfaces, do have correct atomic resolution detail. These partial native-like arrangements lead to an enrichment of correct topology models by energy. An improved iterative sampling protocol that includes the density map as an experimental restraint might allow refinement to atomic detail accuracy for complete models in the future.

EM-Fold has been applied to build a model of adenovirus protein IIIa, a protein for which we have a medium-resolution cryo-EM density map but no atomic resolution structure. Based

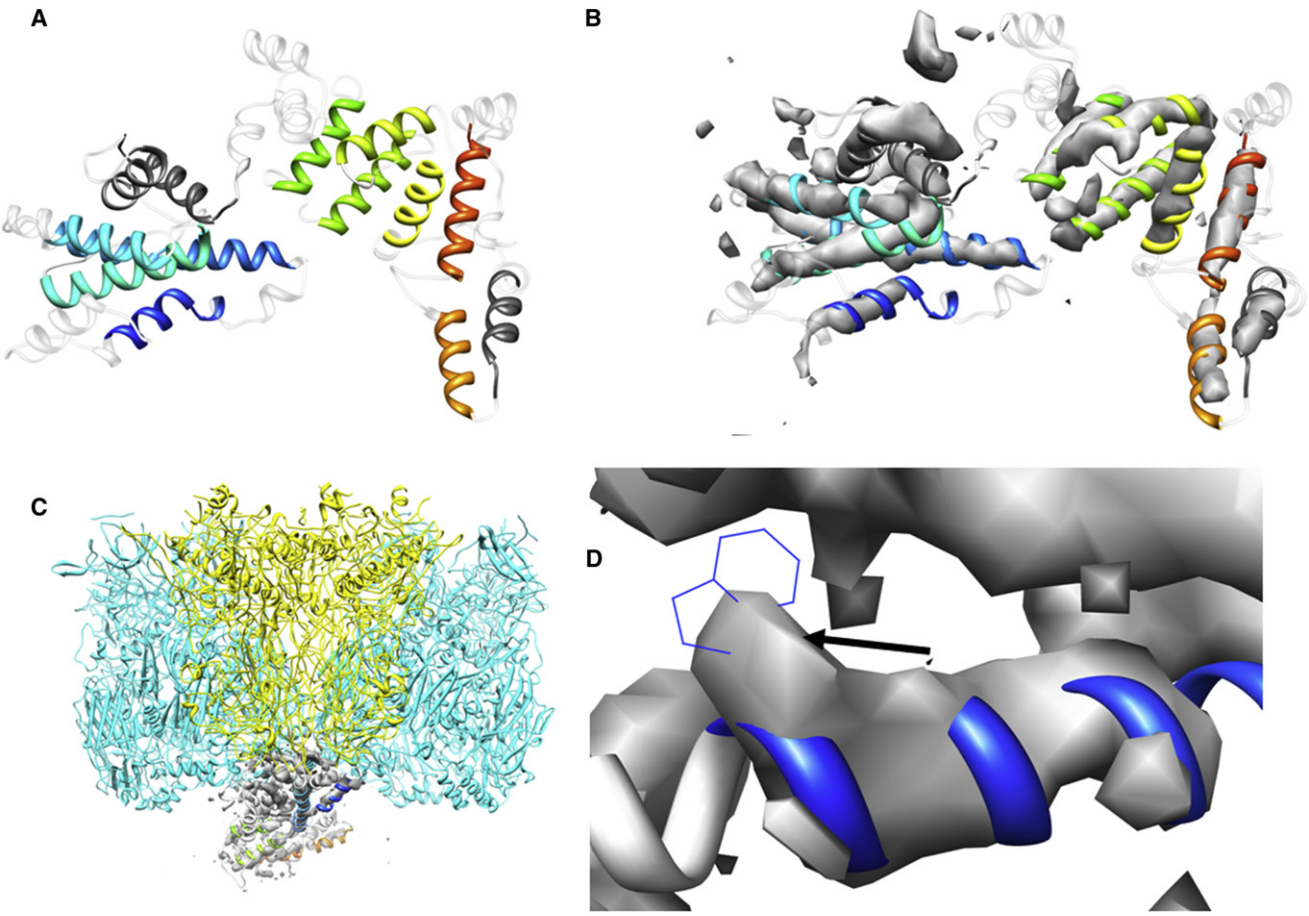

Figure 7. Model of Protein IIIa

(A) A reduced model of protein IIIa where only α helices that have been placed with at least 50% repetition rate are colored in rainbow. This topology agrees with the San Martin et al. (2008) results. A total of 11 of 14 α helices can be placed with a confidence of at least 82%. The remaining three α helices have been colored in gray whereas the loop regions are shown in white.

(B) Same as in (A), but shown but shown in density.

(C) Side view of partial model of protein IIIa (rainbow) in contact with penton base (yellow) and two peripentonal hexons (light blue).

(D) Density bump in rod E of the refined Ad35F density of protein IIIa that has been assigned to Trp27. The arrow marks the position of the side chain.

on the experimental constraints provided by N-terminal tagging (San Martin et al., 2008) as well as observed side-chain density in a refined cryo-EM density map, we were able to assign the lobe topology of the protein. We also used this experimental information as a filter to select the most likely fold models for protein IIIa produced by EM-Fold. We present a fold model for protein IIIa with 11 of the 14 α helices placed with a high level of confidence. Future improvements to the EM-Fold method will include improving secondary structure prediction, consideration of large side-chain information during the assembly stage, and simultaneous assembly of α helices and β strands.

EXPERIMENTAL PROCEDURES

Overall Protocol

The flowchart of the full assembly process is shown in Figure 1. The generation of a pool and identification of density rods is followed by the main assembly step in EM-Fold, a refinement step within EM-Fold, and loop and side-chain building in Rosetta. The assembly step builds α helices from the pool into the density rods. Three sequence-independent, computationally inexpensive, and therefore low-resolution scores are used to build a large number of initial models. The best scoring models from the assembly step are refined using sequence-dependent, medium-resolution scores and leaving the overall fold of the protein unchanged. The last step of the assembly protocol uses the existing Rosetta software (Rohl et al., 2004a; Sood and Baker, 2006) to model loops for the best-scoring models that emerged from the refinement step. Side chains are constructed using Rosetta relaxation and repacking strategies (Bradley et al., 2005). This is the computationally most expensive and highest resolution step of the model building process and is thus only applied to a handful of final models.

Secondary Structure Prediction Pool

To minimize secondary structure prediction inaccuracies, three different secondary structure pools (A, B, C) were investigated. Pool A uses the secondary structure prediction programs jufo (Meiler and Baker, 2003a; Meiler et al., 2001), psipred (Jones, 1999), and sam (Chandonia and Karplus, 1999; Karplus et al., 1997) to get three state predictions of the secondary structure of the benchmark cases. Sequences of more than nine amino acids predicted to be α-helical were considered to be a likely position of a non-short α-helix and were added to a "pool" of possible secondary structure elements. In addition to the individual predictions, a consensus secondary structure prediction was calculated by averaging jufo, sam, and psipred. Also, α helices longer than 21 residues were split into two, further expanding this pool.

In pool B, copies of the α helices from pool A were replaced with copies that are extended by one amino acid on both sides. Thus pool B has the same size as pool A, but all of the α helices are 2 residues longer. This procedure eliminated the bias in pool A toward α helices that are too short and reduces the per α-helix deviation from the correct secondary structure from 1.5 residues in pool A to 0.8 residues in pool B.

Pool C combines pools A and B and adds further versions of α helices extended by one amino acid either on the N terminus or the C terminus. As a result, the secondary structure element pool C has four versions of each α helix with different lengths available for assembly. The per α-helix deviation from the correct secondary structure in pool C is 0.4 residues. The length deviations of the elements that are closest in length and have maximal sequence overlap with the true α helices are reported in Table S2 for all three versions of the prediction pool.

EM-Fold Scoring Function

Three sequence-independent scores are used during the assembly of the fold: a loop, an occupancy, and a connectivity score. The loop score is a knowledge-based score that evaluates the likeliness of a certain C_α-C_α distance between terminal residues in an α-helix being bridged by a specific number of residues. It has a preference for short EUCLIDEAN distances between beginning and end of a loop (data not shown).

The occupancy score evaluates the length agreement of a density with an α helix that is placed in it (see Figure S1) with unfilled densities getting the maximum unfavorable score. Thus, the occupancy score drives the algorithm toward filling the density map completely.

The connectivity score is based on the assumption that, for short loops, a medium-resolution density map contains valuable information in the form of stronger density in the loop regions between density rods. The connectivity score employs a skeletonization algorithm (Ju et al., 2007) to find the highest intensity connection between all pairs of termini of density rods that are closer than 10 Å in space. This information is converted into a score that assesses whether the connection is a strong or a weak one (see Figure S2).

The connectivity score has been tested on the ten benchmark proteins. Within the ten proteins there are 65 pairs of density rods whose ends are closer than 10 Å. 25 of these pairs correspond to connected density rods. Figure 3C shows a ROC curve based on the strength of the connection. The area under the curve is 0.86, clearly showing the ability of the connectivity score to enrich for native connections. Out of 14 connections whose strength is more than one standard deviation above the average connection strength, 12 correspond to true connections.

EM-Fold Assembly Step

The sampling of conformational space is performed in a Monte Carlo algorithm in conjunction with the Metropolis criterion. When placing an α helix from the pool into a density two physical constraints are checked. The first is whether the length of the α helix fits the density within a deviation of 3 residues (corresponding to a maximum length deviation of 4.5 Å). This "length-tolerance-check" accounts for inaccuracies both in secondary structure prediction and in length determination of density rods. Second, it is checked whether the residues between the α helix and all previously placed α helices are sufficient to fill the gaps between α helices. The maximum loop length was set to 3.0 Å per amino acid plus an additional 6.0 Å per loop. If one of the constraints is violated, the move will be rejected because the resulting model would not agree with the density map. All placements that do not violate these constraints are evaluated by the three sequence-independent scores discussed above. Assuming that x density rods have been identified in the density map and the pool contains y α helices, there is a total of N_{pos} number of possibilities to place the α helices into the density rods:

$$N_{pos} = \binom{n}{k} k!2^k = \frac{n!}{(n-k)!}2^k,$$

where n = max(x;y) and k = min(x;y). This same equation is also used to calculate an overall confidence score for a partial model built by EM-Fold by reassigning n to the total number of α helices and k to the number of confidently placed α helices (with > 50% repetition rate).

The Monte Carlo moves (see Figure 8) that are used in the assembly step are: (B) adding an α helix from the pool to the model, (C) deleting an α helix from the model, (D) flipping the orientation of an α helix in the model, (E) swapping the positions of two α helices within the model, (F) swapping an α helix from the model with one from the pool, and (G) moving an α helix from the model to an empty density rod. The orientation of an α helix after any move that results in placement of a new α helix (moves B, E, F, and G) is arbitrary. A simulated annealing Monte Carlo Metropolis search is used where the temperature is decreased linearly from 0.25 to 0.08 over 2000 rejected steps. The weights of the scores are 1.0 (loop), 0.4 (occupancy), and 0.8 (connectivity). The final total scores range from −4.2 (2AX6) to −22.1 (1OUV). It is important to note that the temperature values are somewhat arbitrary and do not correspond to physiologically relevant temperatures.

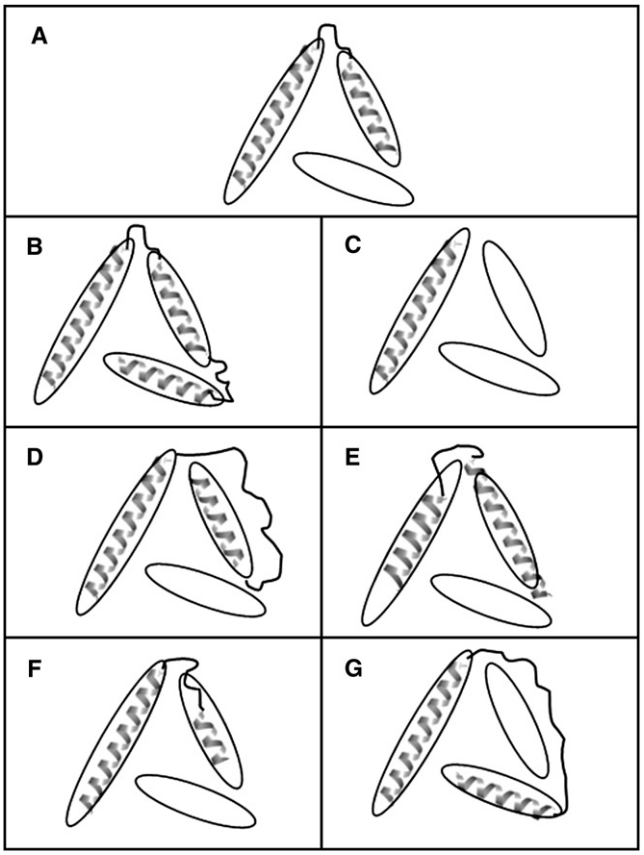

Figure 8. Schematic Representation of the Moves Used in the Assembly Step of the Protocol

(A) State of the model before the move.
(B) The add move adds α helix from the pool into an empty density rod.
(C) The delete move removes α helix from a density rod and returns it to the pool.
(D) The flip move rotates one α helix within a density rod by 180° perpendicular to its long axis.
(E) The swap move exchanges two α helices within density rods.
(F and G) The swap with pool move exchanges an α helix within a density rod with one from the pool. Move G removes α helix from its density rod and places it into another empty density rod.

EM-Fold Refinement Step

The lowest scoring models are used in a second medium resolution Monte Carlo refinement search. This refinement step uses different moves and scores than the previous assembly step. The moves constitute small perturbations of the model—shifts along the α-helical axis and rotations around the α-helical axis. A set of knowledge-based scores is used including an amino-acid-distance score, a neighbor-count score, a secondary-structure-element-packing score, a compactness-measure in form of a radius-of-gyration score,

and the loop and occupancy scores already used in the previous step. These scores are described in detail in Supplemental Data. The occupancy score avoids α helices sliding out of their density rods. This refinement step maintains the fold of the model but identifies correct α-helix-helix-interfaces. A simulated annealing Monte Carlo Metropolis search is used where the temperature is decreased linearly from 0.25 to 0.03 over 2000 rejected steps. The weights of the scores are 10 (loop), 4 (occupancy), 0.2 (aadist), 0.2 (neighbor count), 0.14 (radius of gyration), and 2 (ssepack). The final total scores range from −139 (2AX6) to −367 (1TBF).

Rosetta Loop and Side-Chain Building Step

For identification of the correct fold and for building a full-atom model of the protein, the Rosetta software (Bradley et al., 2005; Rohl et al., 2004a; Sood and Baker, 2006) was used. The backbone atoms of the residues that are missing in the EM-Fold models are built using the Rosetta cyclic coordinate descent loop-building protocol (Rohl et al., 2004a). The resulting models with loops are scored in the Rosetta force field and sorted according to their score. This score can discriminate the correct from non-native topologies as demonstrated in the benchmark. For the seven successful benchmark proteins, the ten best scoring topologies according to the Rosetta score were chosen and underwent an extensive refinement protocol within Rosetta. This protocol included building 1,000 EM-Fold-refined models per topology (10,000 models total). For each of the 10,000 refined models, 5 loop models were built in Rosetta (50,000 models total).

Eight rounds of iterative side-chain repacking and backbone relaxation in Rosetta followed (Bradley et al., 2005). All 50,000 models undergo round 1. Only models that stay within 2.5 Å of the starting structure and are within the best 10% scoring models according to the Rosetta full-atom energy are run through rounds 2–8. After the eighth round, the best 10% scoring models are analyzed according to their enrichment for the correct topology. The enrichment is computed as the ratio of relative frequency of correct topology models within the best 10% scoring models to relative frequency of correct topology models within all models.

Benchmark on Simulated Density Maps

The proposed EM-Fold search algorithm was benchmarked on ten proteins that were chosen to be mainly α-helical, exhibit nonredundant folds, possess 250 to 350 residues, and form 6 to 14 α helices of at least 12 residues in length (Table S1). Electron density maps for all ten benchmark cases were created from the coordinates. PDB2VOL of the Situs package (Wriggers and Birmanns, 2001) was used to simulate density maps with 6.9 Å resolution, a voxel spacing of 1.5 Å, and Gaussian flattening. Positions and lengths of the density rods were identified manually because available α-helix identification algorithms did not perform satisfactorily for either the simulated densities of the benchmark proteins or for the protein IIIa density. Errors in manual identification of α-helix length can be compensated by the length tolerance that is used in the assembly step. To test the influence of the resolution of the simulated medium-resolution density map, maps at 9.0 Å resolution were also simulated. Positions and lengths of the density rods were identified manually for the 9.0 Å resolution maps as well.

Furthermore, it should be stressed that, independent of whether the density rods are identified manually or using automated software, there is always the possibility that density regions in medium resolution density maps that do not correspond to α helices are identified as α-helical regions. An example for this is a β-hairpin of at least four residues in each strand. Likewise, it is possible that an α-helical region in the protein is not identified as a density rod in the map (in the case of a more flexible α-helix for instance). In both cases EM-Fold is still capable of finding the correct topology, because the algorithm neither requires all identified rods to be filled with α helices, nor all predicted α helices to be placed in identified rods.

Benchmark on Experimental Density Map

EM-Fold was also benchmarked on the experimental cryo-EM density map of bovine metarhodopsin (EMDB Entry EMD-1079). The density map is reported to have a resolution of 5.5 Å and has a voxel size of (0.4 Å, 0.5 Å, 1.7 Å). A single subunit of the protein was segmented from the density map. Bovine rhodopsin has 349 residues and is highly α-helical (63% α-helical), with 8 α helices of 12 or more residues.

Protein IIIa Structure Elucidation

The adenovirus vector Ad35F was used in previous cryo-EM structural studies (Saban et al., 2006) and has been refined further with more data (7133 particle images) with the program Frealign (Grigorieff, 2007). A negative temperature factor of 450 $Å^2$ was applied to the final map and the structure was filtered at 5.1 Å using a filter with a cosine-shaped cut-off and a width of ~20 Fourier pixels. Ad35F is composed of the Ad5 capsid and the Ad35 fiber. The density for one copy of protein IIIa was segmented from an Ad35F reconstruction. The Ad5 protein IIIa has 585 residues. The 400 N-terminal residues are predicted to be mainly α-helical, whereas the remaining C-terminal residues are not predicted to have many secondary structural elements. In the density map, 14 density rods of at least 18 Å in length and 6–7 Å diameter (corresponding to α helices of at least 12 residues) were identified manually (Figure 6). A secondary structure element pool with a total of 257 α helices was built using the protocol described for pool C. A total of 100,000 models were built for protein IIIa according to the assembly procedure established for the benchmark set of proteins. The models were ranked by score. A total of 100 refined models were constructed for each of the top 150 models produced by the assembly step. The resulting 15,000 models were sorted by score and the top scoring model of each of the 150 topologies was selected for loop construction. A topological model built using EM-Fold is presented for the first 400 residues of protein IIIa.

EM-Fold Availability

EM-Fold is freely available for academic use. It will be made available as a part of the Biochemical Library that is currently being developed in the Meiler laboratory (www.meilerlab.org). In the meantime, an executable can be obtained by contacting the authors.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, six figures, and two tables and can be found with this article online at http://www.cell.com/structure/supplemental/S0969-2126(09)00223-8.

ACKNOWLEDGMENTS

This study was supported by grants from the National Science Foundation (0742762 to J.M.) and the National Institutes of Health (R01-GM080403 to J.M. and R01-AI42929 to P.L.S.). We thank the ACCRE staff at Vanderbilt for computer support. S.L., R.S., N.W., and M.K. developed methods; R.S. and S.L. developed de novo folding protocol; S.L. developed high-resolution refinement protocol, performed benchmark and experimental model building; P.L.S. and J.M. designed research; S.L., P.L.S., and J.M. wrote the paper.

Received: January 8, 2009
Revised: May 31, 2009
Accepted: June 2, 2009
Published: July 14, 2009

REFERENCES

Alexander, N., Bortolus, M., Al-Mestarihi, A., McHaourab, H., and Meiler, J. (2008). De novo high-resolution protein structure determination from sparse spin-labeling EPR data. Structure *16*, 181–195.

Baker, M.L., Ju, T., and Chiu, W. (2007). Identification of secondary structure elements in intermediate-resolution density maps. Structure *15*, 7–19.

Baker, M.L., Jiang, W., Wedemeyer, W.J., Rixon, F.J., Baker, D., and Chiu, W. (2006). Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density. PLoS Comput. Biol. *2*, e146.

Bonneau, R., Ruczinski, I., Tsai, J., and Baker, D. (2002a). Contact order and ab initio protein structure prediction. Protein Sci. *11*, 1937–1944.

Bonneau, R., Strauss, C.E., Rohl, C.A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., and Baker, D. (2002b). De novo prediction of three-dimensional structures for major protein families. J. Mol. Biol. *322*, 65–78.

Booth, C.R., Jiang, W., Baker, M.L., Zhou, Z.H., Ludtke, S.J., and Chiu, W. (2004). A 9 angstrom single particle reconstruction from CCD captured images on a 200 kV electron cryomicroscope. J. Struct. Biol. *147*, 116–127.

Bottcher, B., Wynne, S.A., and Crowther, R.A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. Nature *386*, 88–91.

Bowers, P.M., Strauss, C.E., and Baker, D. (2000). De novo protein structure determination using sparse NMR data. J. Biomol. NMR *18*, 311–318.

Bradley, P., Misura, K.M., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868–1871.

Chandonia, J.M., and Karplus, M. (1999). New methods for accurate prediction of protein secondary structure. Proteins *35*, 293–306.

Conway, J.F., Cheng, N., Zlotnick, A., Wingfield, P.T., Stahl, S.J., and Steven, A.C. (1997). Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. Nature *386*, 91–94.

Dal Palu, A., He, J., Pontelli, E., and Lu, Y. (2006). Identification of alpha-helices from low resolution protein density maps. Comput. Syst. Bioinformatics Conf. *2006*, 89–98.

Grigorieff, N. (2007). FREALIGN: High-resolution refinement of single particle structures. J. Struct. Biol. *157*, 117–125.

Hanson, S.M., Dawson, E.S., Francis, D.J., Van Eps, N., Klug, C.S., Hubbell, W.L., Meiler, J., and Gurevich, V.V. (2008). A model for the solution structure of the rod arrestin tetramer. Structure *16*, 924–934.

Jiang, W., Baker, M.L., Ludtke, S.J., and Chiu, W. (2001). Bridging the information gap: Computational tools for intermediate resolution structure interpretation. J. Mol. Biol. *308*, 1033–1044.

Jiang, W., Baker, M.L., Jakana, J., Weigele, P.R., King, J., and Chiu, W. (2008). Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy. Nature *451*, 1130–1134.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

Ju, T., Baker, M.L., and Chiu, W. (2007). Computing a family of skeletons of volumetric models for shape description. Comput. Aided Des. *39*, 352–360.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. (1997). Predicting protein structure using hidden Markov models. Proteins *1997* (Suppl 1), 134–139.

Kovacs, J.A., Yeager, M., and Abagyan, R. (2007). Computational prediction of atomic structures of helical membrane proteins aided by EM maps. Biophys. J. *93*, 1950–1959.

Li, J., Edwards, P.C., Burghammer, M., Villa, C., and Schertler, G.F. (2004). Structure of bovine rhodopsin in a trigonal crystal form. J. Mol. Biol. *343*, 1409–1438.

Lindert, S., Stewart, P.L., and Meiler, J. (2009). Hybrid approaches: applying computational methods in cryo-electron microscopy. Curr. Opin. Struct. Biol. *19*, 218–225.

Ludtke, S.J., Baker, M.L., Chen, D.H., Song, J.L., Chuang, D.T., and Chiu, W. (2008). De novo backbone trace of GroEL from single particle electron cryomicroscopy. Structure *16*, 441–448.

Martin, A.G., Depoix, F., Stohr, M., Meissner, U., Hagner-Holler, S., Hammouti, K., Burmester, T., Heyd, J., Wriggers, W., and Markl, J. (2007). Limulus polyphemus hemocyanin: 10 angstrom cryo-EM structure, sequence analysis, molecular modelling and rigid-body fitting reveal the interfaces between the eight hexamers. J. Mol. Biol. *366*, 1332–1350.

Meiler, J., and Baker, D. (2003a). Coupled prediction of protein secondary and tertiary structure. Proc. Natl. Acad. Sci. USA *100*, 12105–12110.

Meiler, J., and Baker, D. (2003b). Rapid protein fold determination using unassigned NMR data. Proc. Natl. Acad. Sci. USA *100*, 15404–15409.

Meiler, J., and Baker, D. (2005). The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. J. Magn. Reson. *173*, 310–316.

Meiler, J., Muller, M., Zeidler, A., and Schmaschke, F. (2001). Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J. Mol. Model. *7*, 360–369.

Min, G., Wang, H., Sun, T.T., and Kong, X.P. (2006). Structural basis for tetraspanin functions as revealed by the cryo-EM structure of uroplakin complexes at 6-A resolution. J. Cell Biol. *173*, 975–983.

Misura, K.M., and Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. Proteins *59*, 15–29.

Misura, K.M., Chivian, D., Rohl, C.A., Kim, D.E., and Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc. Natl. Acad. Sci. USA *103*, 5361–5366.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr. Opin. Struct. Biol. *15*, 285–289.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. Nature *450*, 259–264.

Rohl, C.A., and Baker, D. (2002). De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J. Am. Chem. Soc. *124*, 2723–2729.

Rohl, C.A., Strauss, C.E., Chivian, D., and Baker, D. (2004a). Modeling structurally variable regions in homologous proteins with rosetta. Proteins *55*, 656–677.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004b). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Rossmann, M.G. (2000). Fitting atomic models into electron-microscopy maps. Acta Crystallogr. D Biol. Crystallogr. *56*, 1341–1349.

Ruprecht, J.J., Mielke, T., Vogel, R., Villa, C., and Schertler, G.F. (2004). Electron crystallography reveals the structure of metarhodopsin I. EMBO J. *23*, 3609–3620.

Rux, J.J., Kuser, P.R., and Burnett, R.M. (2003). Structural and phylogenetic analysis of adenovirus hexons by use of high-resolution x-ray crystallographic, molecular modeling, and sequence-based methods. J. Virol. *77*, 9553–9566.

Saban, S.D., Silvestry, M., Nemerow, G.R., and Stewart, P.L. (2006). Visualization of alpha-helices in a 6-angstrom resolution cryoelectron microscopy structure of adenovirus allows refinement of capsid protein assignments. J. Virol. *80*, 12049–12059.

San Martin, C., Glasgow, J.N., Borovjagin, A., Beatty, M.S., Kashentseva, E.A., Curiel, D.T., Marabini, R., and Dmitriev, I.P. (2008). Localization of the N-terminus of minor coat protein IIIa in the adenovirus capsid. J. Mol. Biol. *383*, 923–934.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. (2005). Progress in modeling of protein structures and interactions. Science *310*, 638–642.

Serysheva, I.I., Ludtke, S.J., Baker, M.L., Cong, Y., Topf, M., Eramian, D., Sali, A., Hamilton, S.L., and Chiu, W. (2008). Subnanometer-resolution electron cryomicroscopy-based domain models for the cytoplasmic region of skeletal muscle RyR channel. Proc. Natl. Acad. Sci. USA *105*, 9610–9615.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins *34*, 82–95.

Sood, V.D., and Baker, D. (2006). Recapitulation and design of protein binding peptide structures and sequences. J. Mol. Biol. *357*, 917–927.

Tama, F., Miyashita, O., and Brooks, C.L., 3rd. (2004a). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. J. Mol. Biol. *337*, 985–999.

Tama, F., Miyashita, O., and Brooks, C.L., 3rd. (2004b). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. J. Struct. Biol. *147*, 315–326.

Topf, M., and Sali, A. (2005). Combining electron microscopy and comparative protein structure modeling. Curr. Opin. Struct. Biol. *15*, 578–585.

Topf, M., Baker, M.L., Marti-Renom, M.A., Chiu, W., and Sali, A. (2006). Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. J. Mol. Biol. *357*, 1655–1668.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. Structure *16*, 295–307.

Trabuco, L.G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure *16*, 673–683.

Velazquez-Muriel, J.A., and Carazo, J.M. (2007). Flexible fitting in 3D-EM with incomplete data on superfamily variability. J. Struct. Biol. *158*, 165–181.

Velazquez-Muriel, J.A., Valle, M., Santamaria-Pang, A., Kakadiaris, I.A., and Carazo, J.M. (2006). Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. Structure *14*, 1115–1126.

Villa, E., Sengupta, J., Trabuco, L.G., LeBarron, J., Baxter, W.T., Shaikh, T.R., Grassucci, R.A., Nissen, P., Ehrenberg, M., Schulten, K., and Frank, J. (2009). Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. Proc. Natl. Acad. Sci. USA *106*, 1063–1068.

Volkmann, N., and Hanein, D. (1999). Quantitative fitting of atomic models into observed densities derived by electron microscopy. J. Struct. Biol. *125*, 176–184.

Wriggers, W., and Birmanns, S. (2001). Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. J. Struct. Biol. *133*, 193–202.

Wriggers, W., Milligan, R.A., and McCammon, J.A. (1999). Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. J. Struct. Biol. *125*, 185–195.

Yu, X., Jin, L., and Zhou, Z.H. (2008). 3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. Nature *453*, 415–419.

Zhang, X., Settembre, E., Xu, C., Dormitzer, P.R., Bellamy, R., Harrison, S.C., and Grigorieff, N. (2008). Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. Proc. Natl. Acad. Sci. USA *105*, 1867–1872.

Zhang, X., Walker, S.B., Chipman, P.R., Nibert, M.L., and Baker, T.S. (2003). Reovirus polymerase lambda 3 localized by cryo-electron microscopy of virions at a resolution of 7.6 angstrom. Nat. Struct. Biol. *10*, 1011–1018.

Zhou, Z.H. (2008). Towards atomic resolution structural determination by single-particle cryo-electron microscopy. Curr. Opin. Struct. Biol. *18*, 218–228.