Published in final edited form as:

Biopolymers. 2012 September ; 97(9): 655–668. doi:10.1002/bip.22065.

Gorgon and Pathwalking: Macromolecular Modeling Tools for Subnanometer Resolution Density Maps

Matthew L. Baker^{1,*}, Mariah R. Baker¹, Corey F. Hryc^{1,2}, Tao Ju², and Wah Chiu^{1,2}

¹National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030

²Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030

³Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130

Abstract

The complex interplay of proteins and other molecules, often in the form of large transitory assemblies, are critical to cellular function. Today, X-ray crystallography and electron cryo-microscopy (cryo-EM) are routinely used to image these macromolecular complexes, though often at limited resolutions. Despite the rapidly growing number of macromolecular structures, few tools exist for modeling and annotating structures in the range of 3-10Å resolution. To address this need, we have developed a number of utilities specifically targeting subnanometer resolution density maps. As part of the 2010 Cryo-EM Modeling Challenge, we demonstrated two of our latest *de novo* modeling tools, *Pathwalking* and *Gorgon*, as well as a tool for secondary structure identification (SSEHunter) and a new rigid-body/flexible fitting tool in *Gorgon*. In total, we submitted 30 structural models from ten different subnanometer resolution data sets in four of the six challenge categories. Each of our utilities produced accurate structural models and annotations across the various density maps. In the end, the utilities that we present here offer users a robust toolkit for analyzing and modeling protein structure in macromolecular assemblies at non-atomic resolutions.

Introduction

From cell motility to signal transduction, large macromolecular assemblies are responsible for driving nearly all cellular events ¹⁻³. These complex machines usually contain tens to hundreds of individual subunits with varied morphologies and functions. As such, it is often difficult to analyze the structure and function of macromolecular assemblies, a critical task in discovering targets for disease prevention, improving human health and increasing our understanding of basic cellular processes.

Currently, the most common techniques for imaging and structure determination of entire macromolecular assemblies are X-ray crystallography and electron cryo-microscopy (cryo-EM) ⁴. Typically, X-ray crystallography has been used to solve the structures of single proteins or small protein complexes, though not all specimens are readily crystallized, such as large, transient and/or structurally dynamic assemblies. Since cryo-EM does not require crystallization ^{5,6}, it permits molecules/assemblies to be studied under near-native conditions, albeit at somewhat lower resolutions. However, in recent years, this resolution

*Corresponding author mbaker@bcm.edu Phone: 713-798-6989 Fax: 713-798-8682 .

difference has drastically reduced; a small but growing number of cryo-EM structures have been solved to better than 5 Å resolution⁷⁻²⁰. Today, nearly one-quarter of all cryo-EM density maps report subnanometer resolutions, while one-third of all X-ray crystallographic macromolecular structures greater than 150 KDa report 3-10 Å resolutions (Figure 1).

To give a conceptual feel for the effects of resolution in macromolecular density maps, the cryo-EM structures for GroEL at 18 Å, 10 Å, 8 Å and 4 Å resolution are shown superimposed on the atomic model from X-ray crystallography (Figure 2)^{14,21-23}. In the 18 Å resolution density map, the overall shape of the assembly can be seen as well as some initial delineation between the subunits. Demarcating the boundaries between subunits is fairly ambiguous until 10 Å. By 8 Å resolution, secondary structure elements (SSEs) begin to appear. Typically, α-helices appear as long density rods at ~9-10 Å resolution and begin to exhibit their characteristic helical pitch at ~6-7 Å resolution. β-sheets appear as flat planes, and strand separation does not occur until ~4.7 Å resolution. Feature recognition tools, such as Helixhunter²⁴, SSEHunter²⁵, Sheetminer²⁶ and Sheettracer²⁷, are often used to locate and place SSEs within cryo-EM density maps. Connections between structural features can often be identified at ~6 Å resolution, and by 4 Å resolution, bulky sidechains appear^{20,28,29}. However, sidechain density is generally not uniformly visible at this resolution.

Interpretation of a density map in terms of primary, secondary or even tertiary structure is critical, though atomic resolution data is not necessarily required to provide accurate structural models²⁹⁻³¹. Computational modeling approaches, such as comparative modeling and *ab initio* modeling have been adapted to utilize density maps at low-resolutions^{9,32,33}. With Modeller and Rosetta, for example, a density map can serve as a constraint for the refinement and selection of models³⁴⁻³⁷.

In the absence of a known or related atomic model, direct model building from density maps at near-atomic resolutions (3-6 Å) is still possible, even if often limited to just Cα backbone traces. In X-ray crystallography, extensions to several model building tools have been developed for these low-resolution density maps^{38,39}. Spawned by a growing number of near-atomic resolution cryo-EM density maps, new modeling software⁴⁰, based primarily on feature recognition, has been used to construct models *de novo*^{7,19}. An initial sequence to structure correspondence matches SSEs in the density with those predicted in the sequence⁴¹. Once an initial correspondence has been established, modeling tools, such as Gorgon⁴⁰, O⁴² or Coot⁴³, are then used to locate and trace the backbone path directly from the density map, similar to X-ray crystallography. However, the resulting models typically contain only Cα atoms, as sidechain density often is not uniformly resolved.

These types of modeling approaches in cryo-EM mirror similar developments in the early years of X-ray crystallography, in which little to no automation was present for generating structural models. As X-ray crystallography matured, tasks to solve individual problems were implemented and then integrated them into a structure determination pipeline, which included methods for skeletonization to trace a Cα backbone, analyzing local regions and their similarities to previously resolved structures from databases, comparing the density to secondary structure elements and tracing specific sequences through density⁴⁴⁻⁴⁷. Once an initial model was generated from the X-ray crystallographic density map, refinement procedures could then improve the model's fit to density and stereochemistry.

Today, a large set of robust X-ray crystallographic tools are available to model proteins structures at resolutions better than 3.5 Å, with a small collection of tools offering modeling capabilities out to 4 Å resolution⁴⁶. Conversely, cryo-EM has only recently seen its first set of near-atomic resolution density maps, and thus cryo-EM specific modeling tools are not nearly as common and lack the integrated pipeline for construction, refinement and

validation. For the 2010 Cryo-EM Modeling Challenge, we showcased two of our current modeling tools that attempt to deliver a more robust and unified approach for modeling subnanometer resolution cryo-EM density maps: *Pathwalking*⁴⁸ (*de novo* model building) and Gorgon (model building, SSE detection and rigid-body fitting)⁴⁰. Here, we describe our basic approach for each of these techniques and the resulting models from each.

Results

We submitted a total of 30 structural models and annotations from SSEHunter, Gorgon and *Pathwalking* for the 2010 Cryo-EM Modeling Challenge in four of the six challenge categories (SSE detection, rigid-body fitting, flexible fitting and backbone tracing). These submissions were from ten of the subnanometer resolution density maps, including Aquaporin (3Å resolution, PDB ID: 3M9I)⁴⁹, ϵ 15 (4.5Å, 7.3Å and 9.5Å resolution, EMDB IDs: 5003, 1557, 1176; PDB ID: 3C5B)^{11,50,51}, GroEL (4Å and 7.7Å resolution, EMDB ID: 5001, 1180; PDB ID: 3CAU, 2C7C, 2C7D)^{14,52}, the 70s *T. thermophilus* ribosome (6.4Å resolution, EMDB ID: 5030; PDB ID: 3FIN, 3FIC)⁵³, Rotavirus VP6 (3.8Å resolution, EMDB ID: 1461; PDB ID: 1QHD)⁵⁴ and Mm-cpn (4.3Å and 8Å resolution, EMDB ID: 5137, 5140; PDB ID: 3LOS, 3IYF)¹⁸. In the following sections, we describe the general algorithm for each of our tools and detail the resulting models/annotations in terms of their accuracy with relation to the known atomic model. For each C α backbone model, we report three scores in regard to model quality: the root-mean-squared distance (RMSD), nearest neighbor mean distance and the CLICK topology score⁵⁵. While the RMSD reflects the overall accuracy of the modeled structure in relation to the known structure, the CLICK webserver provides an alternative, topologically independent mechanism to identify structural similarity between two proteins. From a CLICK alignment of two protein structures, a topology score can be calculated to measure how similar the topologies of the matched structures are to one another based on matched sequence fragments. Topology score range from 0 to 1, where 1 indicates topologically identical structures and 0 indicates topologically dissimilar structures. The mean nearest neighbor distance is similar to the RMSD calculation but does not explicitly rely on sequence. Rather, distances are calculated based on the best pairs of aligned C α atoms irrespective of sequence, providing another measure of fold similarity. Together, these scores provide an effective mechanism for evaluating model accuracy at the sequence and fold levels.

It should be noted that for each of the following procedures, analysis was done a single segmented subunit unless otherwise noted. For the purpose of the challenge, segmentations were performed with UCSF Chimera using the “Zone” and “Volume Eraser” tools to help eliminate possible errors from improper segmentations⁵⁶. Furthermore, in practice, poor segmentations can significantly alter the results from the following density map analysis and modeling methods.

SSE detection with SSEHunter

Detecting SSEs is a key step in analyzing subnanometer resolution density maps and, when identified, can serve as important landmarks in model fitting, segmentation and *de novo* modeling. Building off of our original template-based method for α -helix detection in subnanometer resolution density maps, we developed SSEHunter to localize both α -helices and β -sheets²⁵. Algorithmically, SSEHunter employs three separate routines for detecting SSEs in a density map. First, a prototypical helix is used to exhaustively search a density map for helix-like regions. Second, a density skeleton, which preserves both the features and topology of the density map at a given threshold, is computed to identify sheet-like features^{57,58}. Finally, a set of pseudoatoms is placed within the density map and local geometry scores are calculated. Scores for these three routines are then combined and

mapped to the pseudoatoms (expressed in the B-factor column of a PDB file), which can then be rendered in visualization software, in Gorgon or other visualization software such as UCSF Chimera. The user then manually groups similarly scored pseudoatoms to form SSEs. For typically sized density maps of single proteins (128×128×128 voxels), this entire procedure takes less than 10 minutes.

For the challenge, we localized SSEs using SSEHunter for Aquaporin at 3Å resolution, ε15 at 4.5Å, 7.3Å and 9.5Å resolution, GroEL at 4Å and 7.7Å resolution, the 70s *T. thermophilus* ribosome at 6.4Å resolution, Rotavirus VP6 at 3.8Å resolution and Mm-cpn at 4.3Å and 8Å resolution. A gallery of the results superimposed on the known X-ray structure is shown in Figure 3 and summarized in Table 1. In total, SSEHunter was able to correctly identify ~90% of all α-helices with at least two-turns and ~77% of all β-sheets having more than two strands without any false positives. As size of SSE increased, the accuracy of detection increased. These results are equivalent to results from our previous evaluation of SSEHunter²⁵. In comparison to the other submitted SSE detection results, SSEs identified by SSEHunter were as good, or better, across the subnanometer resolution range. However, it should be noted that SSEHunter requires user intervention in the final assembly of SSEs, which may result in slightly different annotations by different users.

In terms of resolution, SSEHunter was able to reliably detect both α-helices and β-sheets from 3Å to nearly 10Å resolution. For ε15, a single gp7 subunit from cryo-EM density maps at three different resolutions (4.5Å, 7.3Å and 9.5Å) was analyzed with SSEHunter (Supplemental Figure 1). At 4.5Å resolution, all of the helices and β-sheets greater than two strands could be detected. At 7.3Å and 9.5Å resolution, five of the six α-helices and both sheets could still be detected. While not as clear as the 4.5Å resolution density map, SSEHunter split one of the two β-sheets into two separate sheets. Similar level of accuracy was obtained for the Mm-cpn (4.3Å and 7.8Å resolution) and GroEL (4.0Å and 7.7Å resolution) density maps.

Gorgon

Due to the complexity of *de novo* modeling at non-atomic resolutions, we created Gorgon (<http://gorgon.wustl.edu>), an interactive molecular modeling toolkit targeted towards density maps ranging from near-atomic to subnanometer resolutions (3-10Å)⁴⁰. Containing ~60,000 lines of code in C++ and Python, Gorgon features a GUI built around a novel *de novo* modeling protocol (reviewed in^{30,31}). Gorgon is distributed freely for Mac OSX, Windows and Linux.

Operationally, Gorgon requires the user to first identify SSEs within a density map. This is accomplished using our aforementioned SSEHunter routine, which is now incorporated into Gorgon. Combined with SSEs predicted from the sequence and a novel density-based skeleton^{57,58}, an automated graph-matching method is used to establish a one-to-one mapping of SSEs, giving rise to an initial path through the map⁴¹. Once a path is established, interactive routines in Gorgon can guide the user to individually place Cα atoms or “sketch” entire loops in the density map. With a complete graphical interface and automation of some complex *de novo* modeling tasks, model building for typically sized proteins takes as little as an afternoon.

For this challenge, five structures were modeled with the Gorgon utilities: Aquaporin-1 at 3Å resolution, rotavirus VP6 at 3.8Å resolution, GroEL at 4Å resolution, Mm-cpn 4.3Å resolution and ε15 gp7 at 4.5Å resolution (Figure 5). For each structure, RMSDs ranged from 3.34Å to 7.89Å and nearest neighbor distances (Cα positions) were 1.58-2.67Å, indicating overall similar structures. The overall topologies of the Gorgon-generated models

were nearly identical to those of the published structures, with CLICK topology scores ranging from 0.76 to 1.00 (Table 2).

Typically the Gorgon models had deviations in regions where structural features were less well-resolved. In $\epsilon 15$ gp7, the long, extended E-loop had weak density connectivity and resulted in less-reliable placement of Ca atoms, resulting in a 7.89Å RMSD, though the Ca mean difference was only 2.67Å, indicating the overall fold was captured during model building though errors in sequence assignment may have been present. Maps with large well-defined SSEs, like rotavirus VP6 and GroEL, had the least amount of difference from the published crystal structure with RMSDs of 3.34Å and 4.70Å and nearest neighbor mean differences of 1.58Å and 2.21Å, respectively.

When comparing the Gorgon models to the published structure, differences were generally attributed to “register shifts” of 1-3 amino acids. These register shifts, where the amino acid assignment was shifted with respect to the known structure, can be seen in examining the RMSDs versus the nearest neighbor mean distances. In these examples and the later *Pathwalking* examples, relatively high RMSDs coupled with low nearest neighbor mean distances indicate that the models captured the overall structure but differed in sequence assignment. In examining the models individually, register shifts generally started at the SSEs. As only SSEs, and not visible sidechain density, is initially used to anchor a backbone trace in Gorgon, model building is susceptible to errors in secondary structure prediction, which is, at best, only about 80% accurate⁵⁹. Typically, errors in secondary structure prediction occur near the ends of the SSEs, creating too short or too long predicted SSEs. When building a *de novo* model in Gorgon, such errors can then be propagated throughout the structure. However, in some structures a secondary shift can compensate in a later SSE. An example, in the Aquaporin model, a helix was shifted by two amino acids. The following two helices were then shifted back by one amino acid each, bringing the Ca trace back into register with the known structure. While the resulting model was topologically similar, the shift introduced a relatively large error (6.49Å RMSD with the known structure). Where visible sidechain densities can serve as markers during model building with Gorgon and help to prevent register shifts and minimize modeling error.

A single experienced user constructed each of the aforementioned Gorgon models. In an attempt to assess model variations from different users, a relatively novice user was asked to construct a model for Mm-cpn. Both users’ models were nearly topologically identical (CLICK topology score of 0.79 and 1.00) to the X-ray structure and had RMSDs of 5.28Å and 7.87Å (experienced and novice user, respectively). Superimposition of the two Mm-cpn Gorgon models showed that the models were topologically identical (CLICK topology score of 1.0) to each other. The majority of the differences between the two models were in the assignment of loops and strands in the apical domain of Mm-cpn; though these errors did not affect the overall structure of the protein.

In addition to model building, Gorgon contains several other structural analysis tools, including a rigid-body fitting routine. For rigid-body fitting in subnanometer resolution density maps, we implemented a novel fitting routine that established correspondences between SSEs in an atomic model (probe) and those detected in the density volume (target) using a clique-based algorithm⁶⁰. In Gorgon, our α -helix only fitting routine requires only seconds to compute even in the largest and most complicated data sets. For the challenge, we performed rigid-body docking with Gorgon on $\epsilon 15$ gp7 at 4.5Å resolution, GroEL at 7.7Å resolution, the 70s *T. thermophilus* ribosome at 6.4Å resolution, rotavirus VP6 at 3.8Å resolution and Mm-cpn at 4.3Å and 8Å resolution. In all cases, the fitted structures were similar to fitting results from other software and within 2Å RMSD to the known position of

the structure in the density map. As such, this rapid fitting of a known structure to a density map can serve as an initial localization of the structure.

Besides being relatively fast (a few seconds per fit for each of the challenge data sets), our fitting routine is capable of fitting and evaluating multiple subunits in a single density map. This was highlighted in the fitting of several individual protein subunits to the ribosome density map. SSEs were first identified in the *T. thermophilus* ribosome density map (a total of 73 α -helices in the 30S subunit). Using these SSEs for fitting, eight chains were fit independently (run time of ~1sec per chain) to the density map (Figure 5). The results obtained for the fittings were nearly identical when compared to our exhaustive cross-correlation based routine, Foldhunter²⁴, which took ~15 minutes per subunit on a modern desktop workstation.

In addition to rigid-body fitting, the same basic fitting algorithm can be applied to sets of SSEs that undergo non-rigid, hinge-like motions, making it possible to describe large deformations for domains in a protein structures. This provides an initial trajectory for flexibly fitting a probe structure. For the challenge, a prototype flexible fitting routine was used on the closed and open states of Mm-cpn at 4.3Å and 8Å resolution, respectively. Merely a proof of principle at the time of the challenge, we were able to demonstrate that our clique-based approach of SSE matching could accurately define subsets of matched helices that underwent similar deformations. In flexibly fitting the closed-state structure of Mm-cpn to the open-state density map, only the positions of the SSEs were considered and deformed. Compared to other flexible fitting routines, our method required only seconds to compute but, aside from α -helix movements, poor results were produced. A more recent version of this approach has been improved using elastic deformation technique guided by density skeletons and now considers all atoms in the protein structure (data not shown). This version will be available in upcoming releases of Gorgon.

Pathwalking

Pathwalking, built on a set of computational tools in EMAN2⁶¹, provides an alternative *de novo* modeling approach to targeting near-atomic resolution density maps⁴⁸. Where Gorgon uses the positions of SSE to anchor an initial sequence to structure assignment, *Pathwalking* has the unique advantage of being sequence and template “free”, meaning that the primary sequence or any other structural information is not considered for the initial construction of a model. The only constraint in *Pathwalking* is that one subunit/domain must be extracted from the entire macromolecular assembly.

The overall *Pathwalking* procedure is similar to a “connect-the-dots” style puzzle and can be broken down into three discrete steps. Briefly, a set of pseudoatoms is first populated within the density map at evenly spaced intervals using a k-means clustering approach. The pseudoatoms will form the nodes/C α atoms along a set of paths that satisfy a set of polypeptide constraints: every pseudoatom is connected to two other pseudoatoms (except the N- and C-terminus), all pseudoatoms must be included and deviation from the observed 3.8Å C α -C α bond distance must be minimized. In *Pathwalking*, we model the construction of a path after the Traveling Salesman problem (TSP), a classical problem in computational mathematics that attempts to find the minimum travel distance through a given set of points⁶². For tracing a protein backbone, a TSP solver can be used directly to find the optimal path among the pseudoatoms, whereby instead of minimizing distance traveled, we attempt to minimize their deviation from the prototypical C α -C α distance (3.8Å). As such, the TSP solver is passed a set of distances “errors” between all points rather than a complete all-to-all distance matrix. The resulting paths represent “first-approach” models, which are defined to be topologically correct models but not fully stereochemically or density-refined. Finally, the primary sequence is threaded on to the model automatically. Even for the novice

user, the entire procedure takes less than one day to build and refine an entire protein structure.

For the challenge, we computationally placed pseudoatoms and calculated *Pathwalking* paths for the following structures: Aquaporin at 3Å resolution, rotavirus VP6 at 3.8Å resolution, GroEL at 4Å resolution, $\epsilon 15$ at 4.5Å resolution, and the 70s *T. thermophilus* ribosome at 6.4Å resolution. A gallery of models can be seen in Figure 5 and a summary of the results in Table 2.

The final *Pathwalking* models for Aquaporin, rotavirus VP6 and a GroEL monomer matched the fold of the known protein. For these data sets, the C α RMSDs were 4.63Å for Aquaporin, 7.47Å for rotavirus VP6 and 7.48Å for GroEL. Despite large RMSDs, the models for each of these proteins were topologically equivalent to that of the known structure, as evaluated by the CLICK topology score (CLICK topology scores of 1.00, 0.93 and 0.79, respectively). While the atomic model of $\epsilon 15$ is not available, the initial model produced by *Pathwalking* was similar to the original hand-built *de novo* model. The *Pathwalking* model had a relatively high RMSD, 9.02Å, when compared to the hand-built *de novo* model but was topologically equivalent with a CLICK score of 0.75.

In the 6.4Å resolution structure of the 70S *T. thermophilus*, a model for chain B in the 30S subunit of the ribosome was modeled using the *Pathwalking* protocol. At this resolution, β -strands were not explicitly resolved, however loops and helices were generally well-defined. The deficiency in β -strands resolution resulted in a relatively high RMSD of 9.86. However, a CLICK topology score of 0.94 suggests that the *Pathwalking* was able to detect and model a topologically correct structure.

Like the Gorgon models, the nearest neighbor mean distances were considerably smaller than the RMSDs, ranging from 1.45Å to 2.31Å. Again, this suggests that *Pathwalking* was able to capture the overall structure of the protein in question, though, as evidenced by the high RMSDs, register shifts were present.

Discussion

With an increasing number of subnanometer resolution structures from cryo-EM and X-ray crystallography, a need for reliable and accurate modeling tools has emerged. Our first attempt at addressing these needs was with our *de novo* model building tools for near-atomic resolution density maps, which relied on the presence of structural landmarks (*i.e.* SSEs) to seed backbone tracing. These tools formed the core of our Gorgon toolkit, which enhance and automate many of the manual processes in *de novo* model building. While Gorgon builds models based on a computed SSE correspondence, our *Pathwalking* utilities construct first-approach models without requiring any sequence or *a priori* knowledge. As part of the 2010 Cryo-EM Modeling Challenge, we demonstrated the efficacy of these tools in modeling protein structure at subnanometer resolutions, as well as the fitting of atomic models and identification of SSEs in these cryo-EM density maps.

While the focus of the Cryo-EM Modeling Challenge was to showcase the current collection of modeling and analysis tools, it also provided a platform for the developers of the software to compare their results to other best-in-class computational tools. Unfortunately, at this time our model submissions were the only ones in the *de novo* modeling category, making it impossible to compare the results of our efforts with other groups participating in the Challenge.

Fitting atomic models

Over the past decade, feature detection routines have been developed. The various tools, each with their own set of input parameters, scoring routines and outputs, are routinely applied to subnanometer resolution cryo-EM density maps^{30,31}. For the modeling challenge, we utilized our SSEHunter routine in Gorgon to identify SSEs in subnanometer resolution density maps. For the purpose of this work, we were not exploring any new techniques or applications of our utility; the detection of SSEs was used primarily for model fitting and building purposes.

Rigid-body and flexible fitting routines have grown increasingly popular in analyzing cryo-EM density maps. These routines range from relatively simple cross-correlation based approaches, to complex simulations that use molecular dynamics to deform a protein structure to better fit the density map (reviewed in^{63,64}). Rather than using an exhaustive approach to fit a model to a density map, our method utilizes SSE locations in the density map to find a good initial fitting solution, which can then be refined by more exhaustive approaches. In this approach, we maximally consider the relevant structural features of subnanometer resolution density maps. Because we reduce the fitting step to consider just SSEs, the effective search space is smaller and can be re-formulated as a clique-based matching procedure. This results in short computational times (1-2 sec for a rigid-body fit) and can be adapted for fitting multiple structures to a density simultaneously. Additionally, our clique-based fitting routine can also locate multiple subsets of SSE agreement, meaning that the groups of SSEs from the probe structure may fit SSEs in the target density map differently. This provides a measure of local deformation that can be applied to flexibly fit the protein structure to the density map based on SSE agreement. At the time of the challenge, we only had a crude proof of concept implementation of our flexible fitting routine. In ongoing work, we have now developed a more complete routine for flexible fitting that is both accurate and fast (less than 1 min for typical size proteins).

De novo modeling with Pathwalking and Gorgon

From the Challenge data sets and prior benchmarks for the individual tools, Pathwalking and Gorgon are capable of producing topologically correct backbone models^{40,48}. When compared to the known structures, models built from near-atomic resolution density maps typically have between ~3-10Å RMSDs (C α atoms only). While these values may appear high, the fold of the protein is generally captured as evidenced by the relatively high CLICK topology scores and the ~2Å nearest neighbor mean distance. These types of initial models provide a basic understanding of the protein structure, though they lack the ability to precisely position exact amino acids, particularly at lower resolutions. Often, these models are accurate enough, that when combined with further optimization steps, can provide a more realistic depiction of protein structure, possible functional mechanisms or interfaces with other proteins.

In the context of model accuracy, one important caveat must be mentioned again. Our current set of tools work primarily on density maps of individual proteins, i.e. a single protein subunit must first be segmented from the entire density map. Missing portions or extra density may result in incorrect models.

The second consideration in our *de novo* model building approach is map quality. As all density maps vary in composition and quality, it is almost impossible to assign strict limits for our approach. Our modeling tools work best on well-segmented, low-noise density in which structural features can be resolved visually. Regardless of reported resolution, even the most experienced users may not be able to build a reliable model in poorly resolved regions of density maps. From our testing of the individual utilities, we can surmise that

de novo modeling is generally possible at 3-6Å resolution, and possibly even to maps as low as 8Å resolution, depending on the resolvability of features in the density map.

Regardless of the tool, a user must visually evaluate any potential model in the context of the density map. A “good” model should provide the following: connect all Ca/pseudoatoms such that each is visited only once, not contain intersecting path segments, have reasonable connectivity (i.e. good bond distances and angles), have connections between Ca/pseudoatoms that are within/bounded by the density map and have “realistic” structural features. Additionally, our current set of tools only considers a single protein, and thus, does not provide any mechanism for assessing interfaces with neighboring subunits in an entire complex. Models must eventually be evaluated in the context of the entire assembly, in which subunit interfaces can be optimized and clashes minimized.

In comparing *Pathwalking* and Gorgon, both techniques produce equivalent, topologically correct folds (Table 2, Supplemental Table 1, Figure 6). When compared to the known structures, the Gorgon *de novo* models generally had slightly better RMSDs with both models having nearly identical CLICK topology scores. The major source of the high RMSDs in either of these approaches was primarily due to sequence registration errors. The implication here is that our modeling methods produced an overall correct trace through the protein even if the sequence assignment was not perfect. This type of error can often be corrected with additional optimization steps, as demonstrated in our previous work³⁴.

In the results for the challenge, we performed *de novo* modeling using Gorgon and *Pathwalking* separately. However, in practice both approaches can be combined for building and validating *de novo* models. In particular, the *Pathwalking* utilities can be used to assess *de novo* model constructed with Gorgon and report potential alternative topologies. Here, a model is first constructed with Gorgon and then passed to *e2pathwalker.py* in EMAN2. This utility, which invokes the TSP solver, is then used to generate a gallery of possible models by introducing slight perturbations in Ca positions and path. In our experience, unambiguously determined structures will have a limited number of alternative paths, while poorly determined models will produce hundreds of alternative paths.

Conclusion

As discussed, our feature identification and *de novo* modeling utilities are capable of rapidly computing first-approach models for individual subunits in large macromolecular complexes. These tools are still under active development and many new features will be added soon, including constraining models based on Ca Ramachandran plots⁶⁵, filtering models based on SSE assignment and improved density skeleton path detection. In their current form, they provide an effective mechanism for model construction of individual protein subunits. Coupled with improved tools for map filtering, segmentation and analysis, we believe our computational tools could become an important part of modeling the growing number of near-atomic and subnanometer resolution density maps by cryo-EM and X-ray crystallography.

Materials and Methods

SSE detection with SSEHunter

Before performing any fitting and modeling procedures, SSEs were detected in the density map using SSEHunter. For the modeling protocols, a single protein was first segmented from the intact assembly using Chimera (Zone option with known structural model). The segmented density map was then normalized with EMAN's *proc3d* utility⁶⁶.

SSEHunter has two graphical user interfaces: one in Chimera and one in Gorgon. For the purpose of the challenge, we used the version in Gorgon. First, a binary skeleton was generated at an isosurface value that showed complete connectivity of the density; this is later used to locate SSEs and calculate possible backbone paths between them. SSEs were then identified in the density with the use of SSEHunter from within Gorgon. This SSE identification procedure generates a set of pseudoatoms that are colored based on the structural features of the atom location, where red pseudoatoms correspond to helical regions and blue pseudoatoms correspond to β -sheet regions. Color intensity reflects the confidence of the SSE assignment. Related pseudoatoms were then grouped together manually in local regions corresponding to the SSEs. Grouping these pseudoatoms resulted in a VRML representation of the SSEs within the density.

In the case of near-atomic resolution density maps, a low-pass filter to 5Å resolution was first applied to the density map using EMAN's *proc3d* routine⁶⁶. This helped eliminate "noise" from the segmentation routine.

Model building in Gorgon

The following methods for constructing models without the aid of a previously resolved structure with Gorgon were completed for the five aforementioned cryo-EM density maps ranging from 3-5Å resolution. As described above, a single protein was first segmented, followed by SSE identification with SSEHunter in Gorgon. Before calculating a SSE correspondence and placing C α atoms within the density map, a secondary structure prediction was run on the primary sequence of the protein of interest using The PSIPRED Protein Structure Prediction Server⁶⁷. These predicted SSEs were then matched to the identified SSEs in the density map using the "Find SSE Correspondences..." tool in Gorgon. Multiple inputs were required in generating the correspondence, including the skeleton, sequence, secondary structure prediction and the SSEs from SSEHunter. A resulting SSE correspondence was then calculated; multiple SSE correspondences were returned and scored from best to worst. In the examples for the challenge, the correct SSE correspondence was always among the top SSE correspondences. From the correspondence, C α atoms were then placed into the density map using the "Semi-Automatic Atom Placement Tool..." in Gorgon. This tool uses the skeleton path connecting SSEs as a guide for users to interactively assign C α positions. Initially, C α atoms were placed in helices, followed by loops and strands. Individual atoms were placed one-at-a-time, using bond length (3.5Å) as a constraint. Where visible, sidechain density was used to help anchor the trace. When the complete model was constructed, a single round of atom placement optimization was done manually to ensure that all atoms fit within the density, that proper bond length was used, and that clashes were removed. This complete modeling building process took less than one day per structure.

Pathwalking

As described, for each density map a single subunit was isolated from the complete density map manually using UCSF Chimera. SSEs were also detected with SSEHunter. Pseudoatoms corresponding to the exact number of expected amino acids were automatically populated in the density maps using *e2segment3d.py* with minimum and maximum distances ranging from 2.2 to 4.2Å. For the challenge, we first populated the density map with pseudoatoms corresponding to the C α positions of the helices as detected by SSEHunter. The final number of pseudoatoms placed corresponded to the total number of amino acids in the protein. Path determination was carried out using the LKH-TSP solver⁶² in *e2pathwalker.py* with minimum and maximum distances set at 3.2 and 4.5Å.

The initial paths through the pseudoatoms were examined in the context of the density map and detected SSEs. When noted, adjustments of pseudoatom locations were performed to improve path geometry and eliminate density outliers. Iterative path determination and optimization (1-3 iterations), beginning with pseudoatoms in SSEs and followed by loops, was performed to improve the agreement to the density map. For the final models the primary sequence was threaded onto the model using *e2seq2pdb.py* and further evaluated with respect to the known structure.

Model Accuracy

For each of the Gorgon and Pathwalking models, we report three scores in regard to model quality. First, we report RMSD as measure of agreement between our model and the known structure. This is calculated in Chimera using the Matchmaker option where corresponding atoms are compared in the sequence. Second, we report a protein topology similarity from CLICK webserver⁵⁵. The CLICK topology score ranges from 0 to 1.0, where 1.0 is an exact topological match. Finally, we report a nearest neighbor distance, which measure the mean distance between aligned C α atoms irrespective of sequence¹⁴.

Rigid Body and Flexible Fitting

Both rigid body and flexible fitting were done using the “Fit to Density...” option in Gorgon with the default fitting parameters. The density map, identified SSEs from SSEHunter and the probe structure were given as inputs. Resulting fits were shown in the “alignments” tab of the “Fit to Density...” widget. In the case of flexible fitting, the top fit was selected and the flexible fitting was performed using the default options in the “flexible fitting” tab of the “Fit to Density...” widget. Results were examined visually and then saved as C α models in Gorgon.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported by grants from NIH through the National Center for Research Resources (P41RR002250), National Institute of General Medical Science (R01GM079429) and National Science Foundation (IIS-0705644, IIS-0705474). M. R. Baker is supported by a postdoctoral training fellowship from the National Library of Medicine Training Program in Computational Biology and Biomedical Informatics provided by the Keck Center and Gulf Coast Consortia (T15LM007093).

References

- (1). Alberts B, Miake-Lye R. Cell. 1992; 68:415–20. [PubMed: 1531448]
- (2). Alberts B. Cell. 1998; 92:291–4. [PubMed: 9476889]
- (3). Sali A. Structure. 2003; 11:1043–7. [PubMed: 12962622]
- (4). Chiu W, Baker ML, Almo SC. Trends Cell Biol. 2006; 16:144–50. [PubMed: 16459078]
- (5). Baumeister W, Steven AC. Trends Biochem Sci. 2000; 25:624–31. [PubMed: 11116190]
- (6). Frank J. Biopolymers. 2003; 68:223–33. [PubMed: 12548625]
- (7). Chen DH, Baker ML, Hryc CF, Dimaio F, Jakana J, Wu W, Dougherty M, Haase-Pettingell C, Schmid MF, Jiang W, Baker D, King JA, Chiu W. Proc Natl Acad Sci U S A. 2011; 108:1355–60. [PubMed: 21220301]
- (8). Chen JZ, Settembre EC, Aoki ST, Zhang X, Bellamy AR, Dormitzer PR, Harrison SC, Grigorieff N. Proc Natl Acad Sci U S A. 2009; 106:10644–8. [PubMed: 19487668]
- (9). Cheng L, Zhu J, Hui WH, Zhang X, Honig B, Fang Q, Zhou ZH. J Mol Biol. 2010; 397:852–63. [PubMed: 20036256]

- (10). Cong Y, Baker ML, Jakana J, Woolford D, Miller EJ, Reissmann S, Kumar RN, Redding-Johanson AM, Batth TS, Mukhopadhyay A, Ludtke SJ, Frydman J, Chiu W. *Proc Natl Acad Sci U S A*. 2010; 107:4967–72. [PubMed: 20194787]
- (11). Jiang W, Baker ML, Jakana J, Weigele PR, King J, Chiu W. *Nature*. 2008; 451:1130–4. [PubMed: 18305544]
- (12). Liu H, Jin L, Koh SBS, Atanasov I, Schein S, Wu L, Zhou ZH. *Science*. 2010; 329:1038–43. [PubMed: 20798312]
- (13). Liu X, Zhang Q, Murata K, Baker ML, Sullivan MB, Fu C, Dougherty MT, Schmid MF, Osburne MS, Chisholm SW, Chiu W. *Nat Struct Mol Biol*. 2010; 17:830–6. [PubMed: 20543830]
- (14). Ludtke SJ, Baker ML, Chen D-HH, Song J-LL, Chuang DT, Chiu W. *Structure*. 2008; 16:441–8. [PubMed: 18334219]
- (15). Miyazawa A, Fujiyoshi Y, Unwin N. *Nature*. 2003; 423:949–55. [PubMed: 12827192]
- (16). Sachse C, Chen JZ, Coureux P-DD, Stroupe ME, Fändrich M, Grigorieff N. *J Mol Biol*. 2007; 371:812–35. [PubMed: 17585939]
- (17). Yu X, Jin L, Zhou ZH. *Nature*. 2008; 453:415–9. [PubMed: 18449192]
- (18). Zhang J, Baker ML, Schröder GF, Douglas NR, Reissmann S, Jakana J, Dougherty M, Fu CJ, Levitt M, Ludtke SJ, Frydman J, Chiu W. *Nature*. 2010; 463:379–83. [PubMed: 20090755]
- (19). Zhang R, Hryc CF, Cong Y, Liu X, Jakana J, Gorchakov R, Baker ML, Weaver SC, Chiu W. *EMBO J*. 2011; 30:3854–63. [PubMed: 21829169]
- (20). Zhang X, Jin L, Fang Q, Hui WH, Zhou ZH. *Cell*. 2010; 141:472–82. [PubMed: 20398923]
- (21). Ranson NA, Farr GW, Roseman AM, Gowen B, Fenton WA, Horwich AL, Saibil HR. *Cell*. 2001; 107:869–79. [PubMed: 11779463]
- (22). Sanz-García E, Stewart AB, Belnap DM. *J Struct Biol*. 2010; 171:216–22. [PubMed: 20353825]
- (23). Stagg SM, Lander GC, Pulokas J, Fellmann D, Cheng A, Quispe JD, Mallick SP, Avila RM, Carragher B, Potter CS. *J Struct Biol*. 2006; 155:470–81. [PubMed: 16762565]
- (24). Jiang W, Baker ML, Ludtke SJ, Chiu W. *J Mol Biol*. 2001; 308:1033–44. [PubMed: 11352589]
- (25). Baker ML, Ju T, Chiu W. *Structure*. 2007; 15:7–19. [PubMed: 17223528]
- (26). Kong Y, Ma J. *J Mol Biol*. 2003; 332:399–413. [PubMed: 12948490]
- (27). Kong Y, Zhang X, Baker TS, Ma J. *J Mol Biol*. 2004; 339:117–30. [PubMed: 15123425]
- (28). Zhang J, Baker ML, Schröder GF, Douglas NR, Reissmann S, Jakana J, Dougherty M, Fu CJ, Levitt M, Ludtke SJ, Frydman J, Chiu W. *Nature*. 2010; 463:379–83. [PubMed: 20090755]
- (29). Zhou ZH. *Curr Opin Struct Biol*. 2008; 18:218–28. [PubMed: 18403197]
- (30). Baker ML, Baker MR, Hryc CF, Dimaio F. *Methods Enzymol*. 2010; 483:1–29. [PubMed: 20888467]
- (31). Baker ML, Zhang J, Ludtke SJ, Chiu W. *Nat Protoc*. 2010; 5:1697–708. [PubMed: 20885381]
- (32). Baker ML, Jiang W, Wedemeyer WJ, Rixon FJ, Baker D, Chiu W. *PLoS Comput Biol*. 2006; 2:e146. [PubMed: 17069457]
- (33). Serysheva II, Ludtke SJ, Baker ML, Cong Y, Topf M, Eramian D, Sali A, Hamilton SL, Chiu W. *Proc Natl Acad Sci U S A*. 2008; 105:9610–5. [PubMed: 18621707]
- (34). DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D. *J Mol Biol*. 2009; 392:181–90. [PubMed: 19596339]
- (35). Topf M, Baker ML, John B, Chiu W, Sali A. *J Struct Biol*. 2005; 149:191–203. [PubMed: 15681235]
- (36). Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A. *J Mol Biol*. 2006; 357:1655–68. [PubMed: 16490207]
- (37). Zhu J, Cheng L, Fang Q, Zhou ZH, Honig B. *J Mol Biol*. 2010; 397:835–51. [PubMed: 20109465]
- (38). Cohen SX, Morris RJ, Fernandez FJ, Ben Jelloul M, Kakaris M, Parthasarathy V, Lamzin VS, Kleywegt GJ, Perrakis A. *Acta Crystallogr D Biol Crystallogr*. 2004; 60:2222–9. [PubMed: 15572775]
- (39). Cowtan K. *Acta Crystallogr D Biol Crystallogr*. 2006; 62:1002–11. [PubMed: 16929101]

- (40). Baker ML, Abeysinghe SS, Schuh S, Coleman RA, Abrams A, Marsh MP, Hryc CF, Ruths T, Chiu W, Ju T. *J Struct Biol.* 2011; 174:360–73. [PubMed: 21296162]
- (41). Abeysinghe S, Ju T, Baker ML, Chiu W. *Comput Aided Design.* 2008; 40:708–720.
- (42). Jones TA, Zou JY, Cowan SW, Kjeldgaard M. *Acta Crystallogr A.* 1991; 47(Pt 2):110–9. [PubMed: 2025413]
- (43). Emsley P, Lohkamp B, Scott WG, Cowtan K. *Acta Crystallogr D Biol Crystallogr.* 2010; 66:486–501. [PubMed: 20383002]
- (44). Jones TA, Thirup S. *EMBO J.* 1986; 5:819–22. [PubMed: 3709525]
- (45). Jones TA, Kjeldgaard M. *Methods Enzymol.* 1997; 277:173–208. [PubMed: 18488310]
- (46). Jones TA. *Acta Crystallogr D Biol Crystallogr.* 2004; 60:2115–25. [PubMed: 15572764]
- (47). Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. *Acta Crystallogr D Biol Crystallogr.* 2004; 60:2240–9. [PubMed: 15572777]
- (48). Baker MR, Rees I, Ludtke SJ, Chiu W, Baker ML. *Structure.* 2012 in press.
- (49). Hite RK, Li Z, Walz T. *EMBO J.* 2010; 29:1652–8. [PubMed: 20389283]
- (50). Booth CR, Jiang W, Baker ML, Zhou ZH, Ludtke SJ, Chiu W. *J Struct Biol.* 2004; 147:116–27. [PubMed: 15193640]
- (51). Zhang J, Nakamura N, Shimizu Y, Liang N, Liu X, Jakana J, Marsh MP, Booth CR, Shinkawa T, Nakata M, Chiu W. *J Struct Biol.* 2009; 165:1–9. [PubMed: 18926912]
- (52). Ranson NA, Clare DK, Farr GW, Houldershaw D, Horwich AL, Saibil HR. *Nat Struct Mol Biol.* 2006; 13:147–52. [PubMed: 16429154]
- (53). Schuette JC, Murphy FV, Kelley AC, Weir JR, Giesebrecht J, Connell SR, Loerke J, Mielke T, Zhang W, Penczek PA, Ramakrishnan V, Spahn CM. *EMBO J.* 2009; 28:755–65. [PubMed: 19229291]
- (54). Zhang X, Settembre E, Xu C, Dormitzer PR, Bellamy R, Harrison SC, Grigorieff N. *Proc Natl Acad Sci U S A.* 2008; 105:1867–72. [PubMed: 18238898]
- (55). Nguyen MN, Tan KP, Madhusudhan MS. *Nucleic Acids Res.* 2011; 39:W24–8. [PubMed: 21602266]
- (56). Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. *J Comput Chem.* 2004; 25:1605–12. [PubMed: 15264254]
- (57). Abeysinghe, SS.; Baker, ML.; Chiu, W.; Ju, T. *IEEE International Conference on Shape Modeling and Applications*; 2008. p. 63-71.
- (58). Ju T, Baker ML, Chiu W. *Comput Aided Des.* 2007; 39:352–360. [PubMed: 18449328]
- (59). Cole C, Barber JD, Barton GJ. *Nucleic Acids Res.* 2008; 36:W197–201. [PubMed: 18463136]
- (60). Abeysinghe S, Baker ML, Chiu W, Ju T. *Computer Graphics Forum (Proceedings of Pacific Graphics 2010).* 2010
- (61). Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, Ludtke SJ. *J Struct Biol.* 2007; 157:38–46. [PubMed: 16859925]
- (62). Helsgaun K. *European Journal of Operational Research.* 2000; 126:106–130.
- (63). Förster F, Villa E. *Methods Enzymol.* 2010; 483:47–72. [PubMed: 20888469]
- (64). Rossmann MG, Morais MC, Leiman PG, Zhang W. *Structure.* 2005; 13:355–62. [PubMed: 15766536]
- (65). Kleywegt GJ. *J Mol Biol.* 1997; 273:371–6. [PubMed: 9344745]
- (66). Ludtke SJ, Baldwin PR, Chiu W. *J Struct Biol.* 1999; 128:82–97. [PubMed: 10600563]
- (67). McGuffin LJ, Bryson K, Jones DT. *Bioinformatics.* 2000; 16:404–5. [PubMed: 10869041]

Figure 1. Subnanometer resolution density maps. A plot of the deposited macromolecular structures in the Protein Data Bank >150KDa (blue) and the EM DataBank (green) are shown sorted by resolution. Nearly 25% of all cryo-EM and ~33% of all X-ray crystallographic density maps report resolutions between 3Å and 10Å.

Figure 2. Density map resolution. Cryo-EM reconstructions of GroEL at 4 different resolutions are shown on the top row. The corresponding resolutions and EMDB ID numbers are shown below. A single subunit from the X-ray crystal structure of GroEL (1SS8) is shown superimposed on the cryo-EM density map in the bottom row.

Figure 3. SSE detection. The results for SSE detection are shown. Helices are represented as green cylinder, β-sheets as cyan planes and the density map is shown as a transparent isosurface.

Figure 4.

De novo models. A gallery of *de novo* models is shown. In the first column, the density maps are shown. In the second and third column, the resulting models from Gorgon and Pathwalking are shown, respectively. For Mm-cpn, only Gorgon was used to construct the model. The two results, grouped together in the box, were constructed by two different users in Gorgon. In the final column, the known structure is shown. All structures are rainbow colored from N- (blue) to C- (red) terminus.

Figure 5. Fitting with Gorgon. Results for the fitting of eight protein subunits in *T. thermophilus* ribosome density map are shown in (A). In (B), the known positions of the protein subunits in the ribosome density map are shown in grey with the Gorgon fits superimposed. The Gorgon “Fit to density” interface is shown in (C).

Figure 6. *Pathwalking* versus Gorgon model building. The results for *Pathwalking* (green) and Gorgon (blue) model building for rotavirus VP6 are shown overlaid on the crystal structure (red). Two zoomed in regions of the models are shown overlaid on the density map to demonstrate the accuracy of model construction.

Table 1

SSEHunter results. A summarized table for all of the SSEHunter results is shown. The number of detected α -helices greater than five amino acids and the number of β -sheets greater than two strands is reported. In Aquaporin, seven of the eight helices were detected. SSEHunter rendered a single helix for two helices in the central portion of the Aquaporin map. For the *T. thermophilus* ribosome results, only the α -helices in the 30S subunit were considered. In $\epsilon 15$ gp7 at 7.3Å and 9.5Å resolution, one of the two β -sheets was represented as two separate β -sheet region in SSEHunter.

Data set	Reported Resolution (Å)	α -helices (detected/actual)	β -sheets (detected/actual)
Aquaporin	3.0	7/8	0/0
Rotavirus VP6	3.8	9/9	2/2
GroEL	4.0	17/17	2/3
	7.7	15/17	2/3
30S Ribosome	6.4	73/84	N/A
Mm-cpn	4.3	15/15	3/4
	8.0	12/15	2/4
$\epsilon 15$ gp7	4.5	6/6	2/2
	7.3	5/6	2/2
	9.5	5/6	2/2

Table 2

De novo results on subnanometer resolution density maps. A summarized table of all *de novo* models generated with Gorgon and *Pathwalking* is presented. For each data set, the number of amino acids for each protein, resolution, C α RMSD, and the topology score from the CLICK webserver are reported. For the Mm-cpn models, two users constructed C α models using Gorgon, designated by the *. Listed first are the results from an experienced Gorgon user, followed by the novice user.

Data set	Method	Reported Resolution (Å)	Number of Amino Acids	C α RMSD (Å)	Nearest neighbor distance (Å)	Topology score
Aquaporin	Gorgon	3.0	220	6.49	2.48	0.76
	Pathwalking	3.0	220	4.63	2.16	1.00
Rotavirus VP6	Gorgon	3.8	397	3.34	1.58	1.00
	Pathwalking	3.8	397	7.47	1.58	0.93
GroEL	Gorgon	4.0	523	4.70	2.21	1.00
	Pathwalking	4.0	523	7.48	2.31	0.79
Ribosome chain B	Pathwalking	6.4	221	9.86	1.45	0.94
Mm-cpn*	Gorgon	4.3	510	5.28	2.09	0.79
	Gorgon	4.3	510	7.87	2.15	1.00
ϵ 15 gp7	Gorgon	4.5	335	7.89	2.67	1.00
	Pathwalking	4.5	335	9.02	2.24	0.75