Article

Resolution and Probabilistic Models of Components in CryoEM Maps of Mature P22 Bacteriophage

Grigore Pintilie,[1,*] Dong-Hua Chen,[1] Cameron A. Haase-Pettingell,[2] Jonathan A. King,[2] and Wah Chiu[1]
[1]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas; and [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT CryoEM continues to produce density maps of larger and more complex assemblies with multiple protein components of mixed symmetries. Resolution is not always uniform throughout a cryoEM map, and it can be useful to estimate the resolution in specific molecular components of a large assembly. In this study, we present procedures to 1) estimate the resolution in subcomponents by gold-standard Fourier shell correlation (FSC); 2) validate modeling procedures, particularly at medium resolutions, which can include loop modeling and flexible fitting; and 3) build probabilistic models that combine high-accuracy priors (such as crystallographic structures) with medium-resolution cryoEM densities. As an example, we apply these methods to new cryoEM maps of the mature bacteriophage P22, reconstructed without imposing icosahedral symmetry. Resolution estimates based on gold-standard FSC show the highest resolution in the coat region (7.6 Å), whereas other components are at slightly lower resolutions: portal (9.2 Å), hub (8.5 Å), tailspike (10.9 Å), and needle (10.5 Å). These differences are indicative of inherent structural heterogeneity and/or reconstruction accuracy in different subcomponents of the map. Probabilistic models for these subcomponents provide new insights, to our knowledge, and structural information when taking into account uncertainty given the limitations of the observed density.

INTRODUCTION

CryoEM is becoming an established method for producing density maps of increasingly complex macromolecular assemblies. A resolution is normally reported with each map, giving an indication of how much detail the map shows. High-resolution reconstructions (<4 Å) can provide detailed atomic-level models including backbone and side-chain atom placements. Medium-resolution reconstructions (4 to 10 Å) typically show secondary structures (e.g., α-helices and β-sheets). At lower resolutions (>10 Å), only individual proteins and their position with respect to each other may be coarsely discernible. Accurate estimation of resolution is thus quite important, as it gives an idea of what information can be extracted from the density map.

The calculation of resolution is based on computing a Fourier shell correlation (FSC) plot between two density maps, each map derived from half of the data (see earlier review (1)). When both halves of the data are aligned to the same reference, the resolution can be overestimated because of overfitting of noise (2); this has led to the gold-standard approach to be proposed, which recommends that both halves of the data be separated from the outset and aligned to independent references (3,4). It is also common that the area of interest (the molecular components) is masked out in the two reconstructions when calculating resolution, so as to eliminate the background where correlations are low. Such masking could introduce artificial correlations and thus affect the estimated resolution, though this can be avoided by using smooth masks (5).

Resolution can vary throughout the density map, and particularly from component to component. Two recent methods have looked at how to estimate local resolutions. One of them is based on FSC calculation of small extracted volumes around each voxel (*blocres*) (6), and another is based on fitting localized basis functions to densities around each voxel (*resmap*) (7). These methods provide a resolution estimate for each voxel in the map. In this study, we are interested in estimating resolution for each individual protein component within a large assembly. This analysis allows an investigator to zoom in one component at a time, for various research inquiries specific to each component. It involves applying masks around segmented components. We address the effect of using either the same or different masks in the two independent reconstructions (which has not been studied before, to our knowledge) as well as the effect of softness of the masks (which has been discussed earlier (5)).

To further analyze density maps at intermediate resolutions, atomic structures from x-ray crystallography are often rigidly docked and compared with map densities (8). In some cases, where there are differences between the docked model and cryoEM density, models are further flexibly fitted to find different conformations that match the cryoEM map.

Submitted July 30, 2015, and accepted for publication November 11, 2015.

*Correspondence: pintilie@bcm.edu

Dong-Hua Chen's present address is Department of Structural Biology, Stanford University, Stanford, California.

Editor: Andreas Engel

0006-3495/16/02/0827/13

http://dx.doi.org/10.1016/j.bpj.2015.11.3522

A main concern when using flexible fitting methods is that they could potentially overfit or unnecessarily distort the model, because of its many degrees of freedom, especially in the presence of noise. To reduce this effect, most methods constrain the model, e.g., via molecular dynamics and rigid bodies (FlexEM) (9), deformable elastic networks (DireX) (10), or molecular dynamics coupled with additional elastic constraints or flexible fitting (MDFF) (11).

In this article, we seek to address 1) how the resulting models can be validated, and 2) how multiple results can be combined to assess and relate the flexibly fitted model to the observed density, particularly in the medium-resolution range. This procedure is meant to be useable with other flexible fitting and modeling methods, and is not method-specific. Although we have used all three flexible fitting methods, in this study we only present and evaluate results produced with the MDFF method. In this particular case, MDFF allows a reasonable amount of flexibility while avoiding the need for defining rigid bodies (as in FlexEM), and allows parts of the model such as long loops to move more freely (elastic networks methods such as DireX tend to keep them closer to their initial conformation).

To address the question of validation or whether a model has been overfitted, one recent approach was to 1) remove data at high frequencies, 2) fit the model to the remaining low-resolution data, and 3) calculate the correlation between the model and the original map in the high-frequency band (12). A weakness of this approach is that, as pointed out by the authors, there can be significant cross talk between structure factors at all resolutions in a cryoEM density map. A more direct approach is to use independent reconstructions, which are now very common, first shown for modeling and refinement at medium-high resolutions using Rosetta (13). The main idea in this approach is to use only one of the independent reconstructions for modeling and refinement, and then use the second independent reconstruction for validation. We use the same approach here as a validation tool for loop modeling and flexible fitting with MDFF.

It is also quite common that long loops are not resolved in crystal structures because of their flexibility. Adding such loops into the model can be done with various homology-modeling software such as MODELER (14) and Rosetta (15), however these tools are not adapted to use densities in cryoEM maps for guidance in the initial step; though the density could be used subsequently to pick good candidates for refinement, this process can be cumbersome especially for longer chains. Gorgon and Pathwalker (16) can be used to more directly build loops in cryoEM density maps, however earlier placement of pseudo-atoms for each residue is needed, which may be hard at lower resolutions where the backbone density is not visible. In this study we present a new method, to our knowledge, to directly add long loops to a model based on segmented densities in medium-resolution cryoEM maps.

It remains an open question how accurate a model resulting from rigid docking and flexible fitting can be. Previous approaches have studied how different methods can be combined to improve the resulting model (17), or using ensembles and clustering to present different results (18). In our study, we explore a fundamentally different approach, which assumes that given low-resolution information, it may be impossible to distinguish between different possible models that match the density equally (or nearly equally) well. To this end, we propose the use of probabilistic models, which aim to capture the uncertainty in the resulting model because of limited and incomplete information in medium-low resolution density maps.

As an example, we apply resolution estimation, loop modeling, flexible fitting, and probabilistic modeling methods to a new asymmetric reconstruction of the P22 bacteriophage in mature form. This is an ideal specimen to demonstrate our approach because 1) it is a large map with multiple subcomponents (coat, portal, hub, tailspike, needle proteins, and packaged DNA); 2) there are known crystallographic structures of portal, hub, tailspike, and needle proteins; 3) the crystallographic models have several missing loops; and 4) flexible fitting in medium-resolution cryoEM maps does not produce unique results, and hence the results are more appropriately represented with probabilistic models.

The mature P22 bacteriophage virion consists of T $=$ 7 icosahedral arrangement of coat proteins; at one fivefold vertex, in the place of five coat proteins (one from each asymmetric unit) are the portal, hub, and tail assembly (19). The portal consists of 12 (gp1) proteins and the hub of 12 (gp4) proteins. Both portal and hub proteins are circularly arranged around the fivefold vertex, with pseudo C12 symmetry. The tail assembly includes six tailspike trimers, with each trimer consisting of three (gp9) proteins, and a needle consisting of three (gp26) proteins. Each tailspike protein has a N-terminal head-binding and a C-terminal adhesin domain.

The entire P22 phage has been imaged in both procapsid and virion forms by cryoEM and reconstructed with icosahedral symmetry to near-atomic resolutions (nongold standard) (20). From the density maps, Cα backbone models for the gp5 coat protein in both procapsid and virion states, were built de novo. Because of icosahedral averaging, portal, hub, and tailspikes are not seen in the icosahedral maps.

Asymmetric reconstructions of both the procapsid and virion have also been previously reported. The procapsid at 8.7 Å resolution (nongold standard) showed interacting shell, portal, and scaffolding proteins (20). The virion at 7.8 Å resolution (nongold standard), revealed coat, portal, hub, and tailspike components (21). In the latter, crystal structures of corresponding proteins were rigidly fitted into the map. Basic flexible fitting was applied to the portal protein showing a narrowing in the top of the barrel domain in the virion form, a possible structural mechanism for head-full sensing and stabilization of DNA once it has been inserted. In our study, we seek to add to this analysis by 1) estimating gold-standard resolutions through each individual component in an asymmetric reconstruction of the P22 virion using a different image processing protocol, 2) building more complete and validated models based on the observed density, and 3) looking for further biological insights based on new probabilistic models.

We first present a newly reconstructed asymmetric density map of the P22 phage in virion form, and we evaluate the gold-standard resolutions in its components such as portal, hub, tailspike, and coat proteins. The proteins for some components have crystallographic structures determined previously, though stretches of their amino acids in some loop regions were not resolved. We use loop modeling to build the connectivity of the missing residues under the restraints of the cryoEM densities. We also use MDFF to establish the models of the portal, hub, and tailspike proteins under the cryoEM density restraints. The results are validated and tested for overfitting by FSC plots between the resulting models and independent, masked density maps. In addition, multiple resulting models are combined into probabilistic models, which reveal the uncertainties in backbone atom positions throughout the model attributable to two factors: 1) the model is underconstrained by the observed density during modeling, and 2) the model's inherent rigidity given its structural composition. We finally discuss some new biological insights, to our knowledge, based on these probabilistic models.

MATERIALS AND METHODS

CryoEM and data processing

P22 virions were purified from *Salmonella typhimurium* using an established procedure (20). R1.2/1.3 400-mesh preirradiated copper Quantifoil grids (Quantifoil Micro Tools, Großlöbichau, Germany) were used for sample freezing on a Vitrobot Mark III (FEI, Hillsboro, OR). No continuous thin carbon support film was applied to the holey grids before specimen freezing. The data was collected on a Gatan 10 × 10 k CCD (US10000XP, 9 μm/pixel, model 990 (22)) with 2× hardware binning on a 300-kV JEM-3200FSC cryo-electron microscope (JEOL, Japan) with a 20-eV energy slit from the in-column energy filter and ~25 e/Å^2 dose per micrograph and with an effective magnification of 70,600× (yielding 2.55 Å/pixel on the specimen level) and the specimen temperature of ~101 K. Particle images were automatically boxed out using the program ethan (23). The total number of particle images was 79,731 with 512 × 512 box size. Contrast transfer function fitting was performed automatically using fitctf.py (24) and subsequently fine-tuned manually using ctfit in EMAN1 (25).

The symmetry-free reconstruction for P22 virion was done in a similar way as that for P22 procapsid (20) but with a gold-standard approach. Briefly, the particle images were first separated into odd/even subsets. Then initial model for each subset was built by EMAN1 program starticos or by multipath simulated annealing (26) and assuming the random orientations then EMAN1's make3d. Afterward the particle icosahedral orientations in each subset were gradually refined to subnanometer resolution using multipath simulated annealing's coarse search. The next steps are for symmetry-free reconstruction. The faint density at the special vertex representing the tailspikes, hub, and portal was segmented out from the icosahedral map and then sixfold averaged to seed as the initial model to aid in the subsequent steps for determining the location and orientation of the special vertex in each particle image (16,17). The reconstruction followed was done with C1 symmetry, producing a new density map including the special vertex. The choice of particle images for the symmetry-free reconstruction was based on the statistical significance (2σ) of the lowest-phase residual in the orientation search of each particle's special vertex based on the already-determined icosahedral orientations (20). The special vertex density segmented from the new density map was used as the new initial model for the next cycle of location and orientation for the unique vertex in each particle image. Several iterations were done for the C1 orientation search until the best resolution based on the gold-standard FSC was reached.

Segmentation

All maps were segmented using *Segger* (27), extracting the shell (gp5; 415 copies), portal (gp1; 12 copies), hub (gp4; 12 copies), tailspikes (gp9, 6 trimers), needle (gp26, 3 proteins), and DNA (Fig. 1). Densities for other components such as tube (gp10), pilot proteins, and portal plug are not analyzed further because of lack of known crystal models or convincing identifications.

When segmenting a density map, *Segger* first computes watershed regions (27). Each watershed region is a set of connected voxels that correspond to a single density maximum. The boundaries between regions delineate the lowest densities between adjacent maxima. Scale-space filtering is then used to group watershed regions, based on progressively low-passed filtered versions of the same map. This typically groups regions from the same protein together; however, manual regrouping is also typically needed to improve the segmentation, for example, using a known crystal model for guidance.

Icosahedron-corrected radial segmentation for the coat and DNA

In the P22 virion, the coat assumes an icosahedral shape. This was leveraged to more easily segment out the coat from the DNA and the rest of the complex. First, watershed regions were obtained for the entire map, while excluding portal, hub, and tailspike subcomponents. Then, the distance of each segment from the center of the virion was computed, and corrected for the icosahedral shape. The latter was done as follows: the vector from the center of the map to the center of each region was projected onto the nearest vector from the center of the map to the center of an icosahedral face. The magnitude of this vector was taken as the icosahedron-corrected radial distance. A histogram of these distances shows distinguishable peaks for the coat and the first few DNA shells. The segments corresponding to the coat were grouped based on this histogram. Slight manual modification was needed at the 12 fivefold vertices, where the shape of the coat is not perfectly icosahedral. This method is available in the *Segger* plugin for UCSF Chimera (28) (*rSeg* module).

Masking for FSC computation

To measure the FSC between different subcomponents (coat, portal, hub, and tailspike trimers) in the virion maps, each of the two independently reconstructed maps with two halves of particle images were masked with segmented regions corresponding to each subcomponent. When using the same mask in two maps with a sharp drop-off, correlations may be expected because of this mask boundary (Fig. 2 *B*). This can be avoided if 1) the masks have different boundaries between the two half maps, and 2) a soft mask is used (Fig. 2 *B*). These two conditions are met here, as follows:

1) When using *Segger* to segment a map, a threshold is applied to the map and only densities above this threshold are included in the segmented regions.

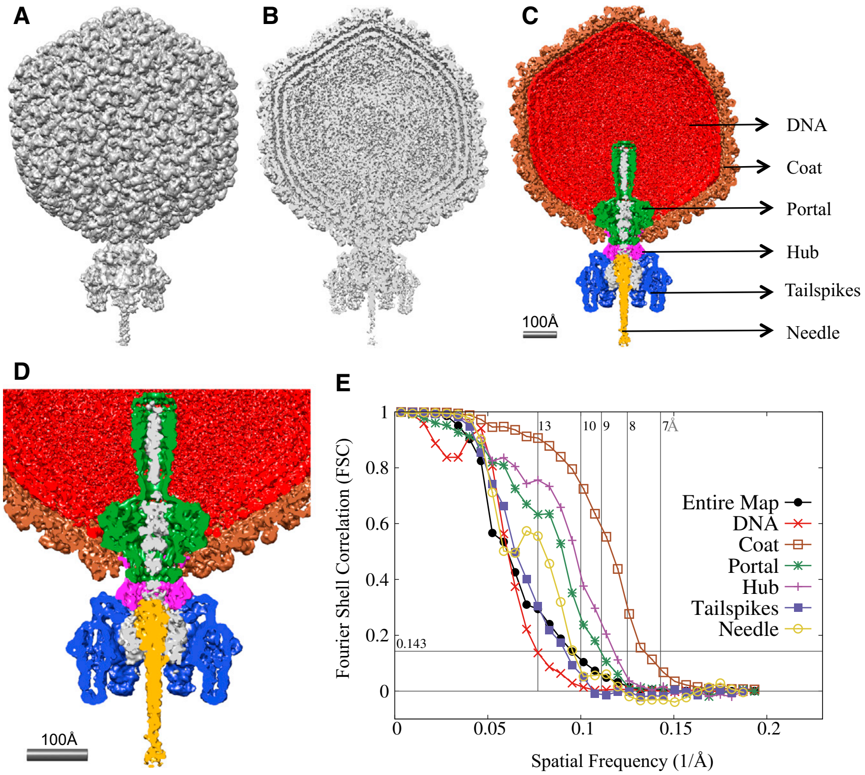

FIGURE 1 Entire P22 density map viewed as (*A*) isosurface and (*B*) isosurface that is sliced through the middle. (*C* and *D*) Segmented components including the coat (*brown*), portal (*green*), hub (*purple*), tailspikes (*blue*), needle (*gold*), and DNA (*red*). (*E*) Gold-standard FSC plots of the entire map and individual components. This figure shows the complex arrangement of various components in the mature P22 virion, and how resolution can vary significantly from component to component.

Because the densities in each map are slightly different, their densities at a given threshold are also different. Furthermore, the boundaries between segments are also different, as they follow the lowest-density contour between corresponding peaks, and these will differ in independent maps.

2) To create a soft mask from a segmented region, a sharp mask is first created by setting the value at all grid points inside the segment to 1 and all others to 0. This mask is then low-passed filtered before it is applied to the map. (The chimera Gaussian filter function is used for this purpose).

The soft mask is computed for each subcomponent, in each of the two independent maps, and applied to the corresponding map. This produces two (independent) masked maps for each subcomponent, which are then used as input into the FSC computation. This procedure is available in the *Segger* plugin for UCSF Chimera (extract module). The FSC computation was performed with EMAN2 (e2proc3d.py with -apix and -calcfsc options).

Furthermore, using the same principles, we also segmented and masked out individual coat, portal, hub, and tailspike proteins. The needle proteins, however, could not be individually segmented, as they are quite narrow and closely intertwined: at the observed resolution, each protein could not be separately delineated using the watershed regions from the nonsmoothed density map.

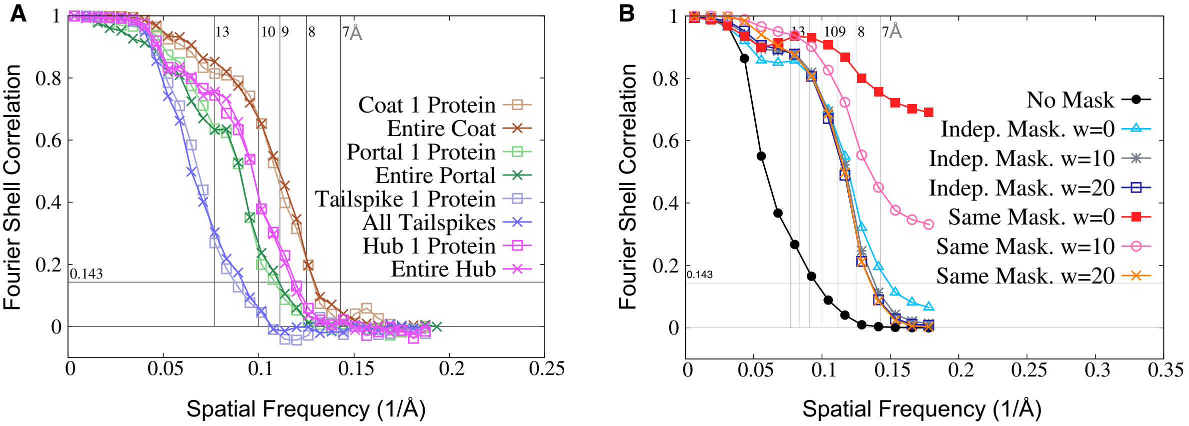

FIGURE 2 (*A*) Gold-standard FSC plots for entire components (coat, portal, tailspikes, and hub) and one protein from each component. Although the map is asymmetrically reconstructed, the components are approximately symmetric (coat-icosahedral, portal-C12, hub-C12, tailspike trimers-C6). Hence the FSC plots for entire components and for single proteins from each component are very similar. (*B*) The effect of mask smoothness and whether the same or different masks are used in gold-standard FSC plots. Using the same mask or a mask with a small smoothing width (w) in both independent maps can introduce artificial correlations at high frequencies. Using independent masks and larger smoothing widths removes this effect. To see this figure in color, go online.

Rigid docking of x-ray structures into density

Crystal structure models were rigidly docked into density maps, using the *Fit to Segments* module in the *Segger* plugin for UCSF Chimera. *Z*-score analyses were performed to ensure that the results are statistically significant (8). To compute the *z*-score, an exhaustive rotational search is performed, and the top score is compared with other random placements of the model in the map. The *z*-score indicates how many standard deviations the top score is above the average of all the other scores. Visual inspection also confirmed that the agreement between docked models and density was good, and any resolved helices in the map correspond to the helices seen in the crystal structure.

Density-guided loop modeling

To add missing loops to a crystal structure model, nearby densities are first identified using *Segger*. A number of points equal to the number of missing residues are then randomly distributed throughout this region. These points are then randomly connected into a chain, and their positions are relaxed gradually such that

- the connected points stay ~3.8 Å away from each other (the average distance between two Cα atoms);
- the nonconnected points push each other away to avoid overlaps
- the points are pushed slightly in the direction of the density gradient to keep the chain inside nearby density
- the points are pushed away from the rest of the model and
- the angle between two lines connecting any three adjacent points is pushed toward 180° (to avoid sharp turns in the chain.

After this relaxation, each point in the chain is replaced with a full atomic model of the corresponding residue, while trying to avoid collisions between the side-chain atoms and the rest of the model. This method is available in the *Segger* plugin for UCSF Chimera (*SegLoop* module).

Flexible fitting using MDFF

Visual molecular dynamics (VMD) (29) and MDFF (11) were used to set up and flexibly fit each model into the density map as described in the online tutorial (30). MDFF uses a full atomic force field to maintain good stereochemistry in the structure, while applying forces at each atom in the direction of the density gradient during the simulation. The forces make the structure gradually change to better fit the observed density. The atomic force field includes force fields that penalizes atom-atom clashes and maintains proper bond length, bond angle, and dihedral angles. In MDFF, the parameters for these forces come by default from the file *par_all27_prot_lipid_na.inp*, which are based on the CHARMM force field (31). Extra elastic restraints for secondary structure dihedrals, hydrogen bonds, *cis* peptide bonds, and chiral centers are added by MDFF to maintain proper secondary structure and prevent overfitting.

MDFF was applied to complete atomic models of the portal (gp1), hub (gp4), and tailspike (gp9) proteins. Portal and hub proteins were simulated together as a complex of 12 copies of portal (gp1) and 12 copies of hub (gp4) proteins (24 proteins in total). The tailspike proteins were simulated as a trimer of three gp9 proteins (there are six trimers per virion). The proteins were simulated together so as to replicate a more accurate environment in which each protein is found. Simulating a single protein from each component may make it appear more flexible than it really is, and nearby contacting proteins could potentially limit this flexibility and reflect the actual biological environment.

In both systems, 1×10^4 minimization steps were performed, followed by 6×10^5 molecular dynamics (MD) steps (1 MD step = 1 fs). During MD, atoms move under 1) the restraints mentioned above, 2) random forces that mimic solvent at a given temperature (300K), and 3) extra density forces applied at each atom in the direction of the gradient of the density. The extra density forces over time tend to move the model toward a different conformation that better matches the density map. Root-mean-square deviation plots showed that the model has stabilized by the end of the simulation (i.e., was no longer changing significantly as it did during the early simulation steps).

Validation and test for overfitting

For this test, only densities from one of the independent maps (map A) are used when adding missing loops and flexibly fitting models using MDFF, as described in the sections above. Then, the second independent map (map B) is used to validate and test for overfitting. The procedure is as follows:

1) The crystal structure is rigidly docked into map A, missing loops are added, and then the complete model is flexibly fitted to map A using MDFF.
2) For comparison and visualization purposes, the atomic model (before and after loop modeling and flexible fitting) of a single protein subunit was extracted and used to simulate a density map at a resolution of 5 Å (using the command molmap in Chimera).
3) Corresponding densities for a single protein were also segmented and extracted from both maps A and B.
4) FSC plots between the initial and final models of a single protein and map A and map B are calculated using EMAN2 e2proc3d.py method.

Probabilistic models

In a medium-resolution density map, each atom in a fitted model is not strongly restrained by the cryoEM density. Therefore, many atomic models with good stereochemistry are possible. Running MDFF on an initial model can produce a number of different structures, as randomness is introduced in each simulation by implicit (or explicit) solvent at a temperature of 300K. The uncertainty in the results is captured here by a probabilistic model. We are only interested in the position of the α-carbon atom in each residue, as at the resolutions seen in this density map, it is not possible to meaningfully restrain positions of side-chain and other atoms in each residue.

In an atomic structure, connected backbone atoms influence each other via indirect atomic bonds, van-der Waal forces, and electrostatic forces; thus the position of each backbone atom is dependent on the positions of all the other backbone atoms. The probability distribution function for the position of each backbone atom would be very complex if we tried to model this interdependence. For simplicity, we only assume that the position of each backbone atom can be modeled as an independent variable, and hence can be represented simply with a single normal (or Gaussian) probability function.

The probabilistic model is built by sampling. In this case we run the modeling and flexible fitting procedure 10 times for each protein subunit, first using only map A. The probabilistic model was built from these 10 results with the following steps:

1) For each residue's Cα atom, its average position among the 10 results is calculated.
2) The average model is determined to be the one out of 10 models that has the minimum sum of distances from each residue's Cα to the average Cα position (from step 1) for that residue.
3) The distance from the Cα atom position in each residue (in each of the 10 results) to the same residue in the average model is calculated.
4) The standard deviation around the average position at each Cα atom position is then calculated.

The loop modeling and flexible fitting was also run five times in map B. The first five results from map A and the five results in map B were then also combined into a probabilistic model. This could be useful in determining whether the probabilistic model can also assess uncertainty seen in independent reconstructions.

The following equations define the probabilistic model more concretely:

Let C_i^n be the three-dimensional position (x,y,z) of the Cα atom in residue i, resulting model n, with

a. $i = 1..M$, M is the number of amino acids/residues in the protein (this number varies from protein to protein),

b. $n = 1..N$, N is the number of resulting models after loop modeling and flexible fitting (here, $N = 10$).

Let $\overline{C_i}$ be the average position of the Cα atom in residue i among the 10 samples, i.e.

$$\overline{C_i} = \frac{\sum_{n=1}^{N} C_i^n}{N}.$$

3) The average or representative model is chosen to minimize the following sum-of-distances function, from the Cα position in each residue to the average Cα position for that residue among the 20 models:

a. the average model is $a = argmin(f(n))$ where

$$f(n) = \sum_{i=1}^{M} \left| C_i^n - \overline{C_i} \right|,$$

$$n = 1..N.$$

4) The standard deviation at each residue, i, is then defined as follows:

$$\sigma_i = \sqrt[2]{\frac{1}{N}\sum_{n=1}^{N}\left(C_i^n - C_i^a\right)^2}, \text{ for } i = 1..M.$$

The standard deviation at each residue is stored in the PDB file in the b-factor column. UCSF Chimera is used to render the probabilistic model using Tools → Depiction → Render by Attribute, varying the Colors and thickness (Worms) at each residue position based on the standard deviation. This method is available in the *Segger* plugin for UCSF Chimera (*ProMod* module).

RESULTS

P22 cryoEM density maps and subcomponents

A total of 1,130 CCD frames were collected, and 79,731 P22 phage particles were picked and used to compute a density map, without imposing any symmetry. The published reconstruction method (26) was used. In addition, two density maps (maps A and B from halves of the same data set) were independently reconstructed, both without symmetry imposition. Gold-standard FSC comparison of these two maps estimates the resolution to be 10.5 Å (Fig. 1). One of the two maps is shown in Fig. 1, first as an entire map (Fig. 1 *A*), a slice through the middle of the map (Fig. 1 *B*), and finally with each segmented component highlighted with a different color (Fig. 1, *C* and *D*).

FSC plots for the entire virion and masked subcomponents are also shown in Fig. 1 *E*. From these plots, it can be seen that the coat protein is resolved to a higher resolution (7.6 Å) than the portal (9.2 Å), hub (8.5 Å), tailspike trimers (10.9 Å), and needle (10.5 Å) components. Because the particle orientations of the icosahedral components (coat proteins) were determined in the first step using the standard Fourier common-line criteria, their determinations are more accurate. The subsequent step in determining the positions and orientations of the portal vertex components may subject to different accuracy. This may contribute to better-resolved densities in the map corresponding to the coat protein. Other factor contributing to the different resolutions in different parts of the map is likely because of the varying conformational flexibility inherent in the protein components. The resolution of the entire virion is 10.5 Å, which is lower than that of most of the individual protein components, likely because the DNA density in the interior part of the particle has the worst resolution (13.1 Å). The protein components with the lower resolutions are the tailspikes (10.9 Å) and needle (10.5 Å).

No symmetry was imposed during reconstruction, however the components themselves are pseudo-symmetric. The coat has icosahedral symmetry, the hub, and portal are in C12, and the tailspike trimers are C6 symmetric. In Fig. S1 in the Supporting Material, the pseudo-symmetry can be seen in rotational cross correlation plots for the portal (C12), hub (C12), and tailspike trimers (C6) though no symmetry was imposed in the reconstruction. The plots indicate that the portal densities are extremely similar and have 12-fold symmetry (all 12 peaks are near 1.0 cross correlation). On the other hand, densities corresponding to the hub and tailspike trimers are less similar, with peaks closer to 0.9.

Individual proteins were also segmented out in both independent maps and compared by FSC plot (Fig. 2 *A*). The FSC plots for individual proteins are very similar to those for the respective entire component, indicating that each component indeed follows the above mentioned symmetries. The $FSC_{0.143}$ resolutions for each component complex compared with $FSC_{0.143}$ resolutions for a single protein are very similar: coat (7.6/7.7), portal (9.2/9.4), hub (8.5/8.4), and tailspike trimers (10.9/11.2), respectively. The map and segmented components have been deposited to the Electron Microscopy Data Bank (EMDB) under a single entry with accession number EMD-8005.

Effect of masking on FSC plots

The effect of masking with the same or independent masks and across various mask widths is illustrated in Fig. 2 *B*. It shows that high-frequency correlations can indeed be introduced when comparing independent maps masked with the same, nonsmoothed mask by FSC, and to a lesser degree also when comparing independent maps masked with different nonsmoothed masks. However, when smoothing the masks in either case, these high-frequency correlations are avoided. At small smoothing widths, high-frequency correlations because of masking appear to be dramatically lower when using different masks in each independent map.

Coat proteins

Fig. 3 *B* and Movie S1 show a single coat protein segmented from the map, along with a rigidly docked model that was

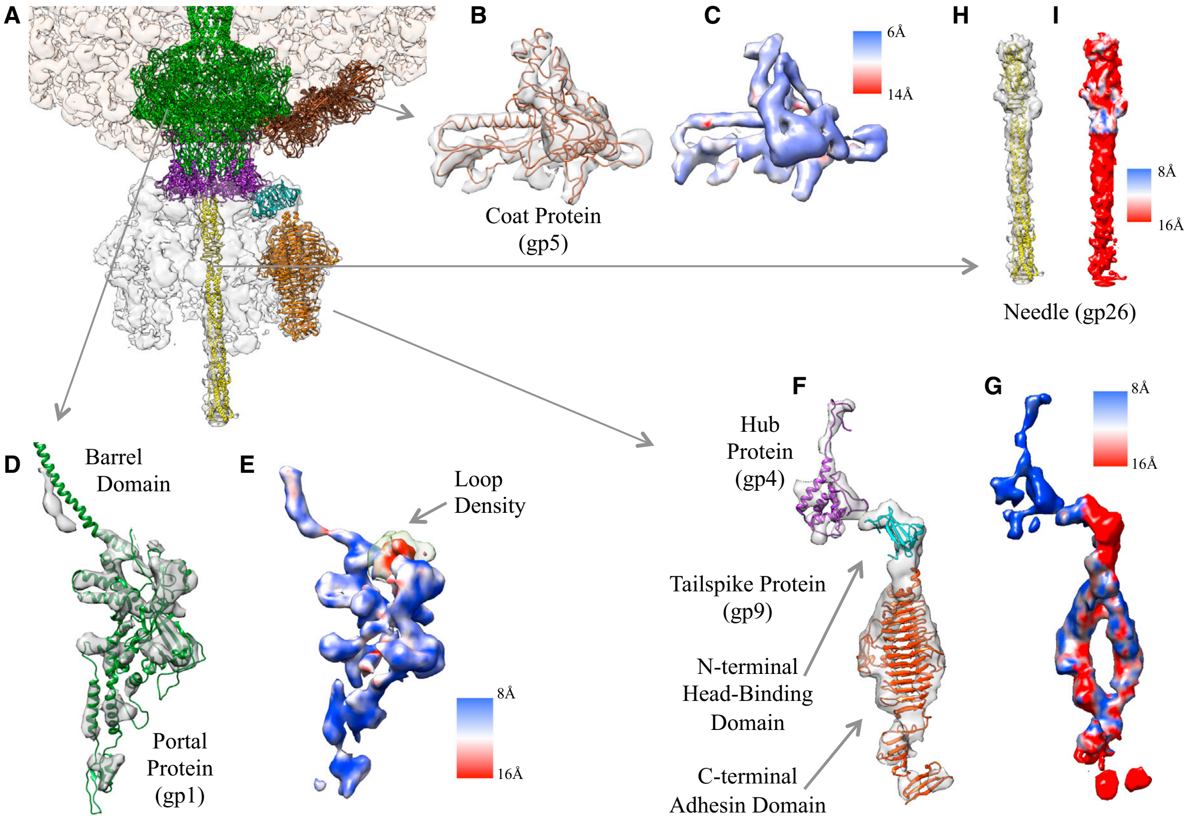

FIGURE 3 (*A*) Coat, portal, and tailspike densities with (*on the left*) docked models and (*on the right*) *resmap* resolution coloring (*blue* to *red*). (*B*, *D*, *F*, and *H*) Rigidly docked models of PDB structures and segmented densities of coat, portal, hub, tailspike, and needle proteins. Segmented densities are shown with a transparent gray surface and docked models are shown using a colored ribbon. PDB codes for docked models are as follows: coat, brown (PDB: 2XYZ); portal, green (PDB: 3LJ5); hub, purple (PDB: 3LJ4), tailspike N-terminal head-binding domain, cyan (PDB: 1LKT); tailspike C-terminal adhesin domain, orange (PDB: 1TYU); needle, yellow (PDB: 2POH). (*C*, *E*, *G*, and *I*) Surfaces of segmented densities of single proteins from each component, colored with per-voxel resolutions computed with *resmap*.

previously built from a higher-resolution icosahedral reconstruction of the P22 virion, PDB: 2XYZ (20). Alpha-helices are clearly seen in the density, as expected at the subnanometer-resolution (7.6 Å) map estimated for this segment (32). Per-voxel resolutions computed with *resmap* are visualized on the surface of the density in Fig. 3 *C* and appear to range mostly between 6 and 9 Å.

Portal proteins

The portal protein complex is made up of 12 copies of gp1, arranged circularly to form a channel through which DNA is inserted into the particle. Long α-helices form a barrel domain, which points toward the center of the capsid. A single portal protein is segmented as shown in Fig. 3 *D* and its surface colored with per-voxel resolutions computed with *resmap* is shown in Fig. 3 *E*. The resolution estimated in this segment is 9.2 Å, and *resmap* resolutions coloring the surface in Fig. 3 *E* appear to be in the range of ~8 to 10 Å.

Fig. 3 *D* shows a crystallographic model of a single portal protein (*green*) (33), rigidly docked into the cryoEM density. The density matches the crystal structure model very well, with some visible α-helices in the density closely matching those in the model. However, there is a significant difference at the bottom of the barrel domain, where the α-helix in the crystal model lies outside the cryoEM density. This difference was also pointed out in a previous study (21), and it reinforces the frequently encountered phenomenon that the crystal structure of a protein does not necessarily reflect its actual conformation(s) in a genuine biological environment.

Hub proteins

The hub consists of 12 gp4 proteins, also circularly arranged like the 12 portal proteins. All 12 hub proteins are shown in Fig. 3 *A*, and a single hub protein along with segmented densities are shown in Fig. 3 *F* (*top*). The figure also shows a docked crystal model of this protein (*purple*) (33). The resolution of this component is estimated to be 8.5 Å, and as would be expected, some of the alpha helices can marginally be seen in the density (Movie S1). Fig. 3 *G* shows a surface of the segmented hub protein with colors reflecting *resmap* resolutions; these appear to vary between 8 and 9Å, consistent with the resolution estimated with the masked $FSC_{0.143}$ criterion.

Tailspike proteins

Six tailspike trimers are arranged circularly around the hub. Fig. 3 *A* shows crystallographic models representing one trimer docked inside the density. The tailspike protein has not been solved in its entirety using crystallography; instead its two domains were solved separately, and hence the entire tailspike protein can be found as two separate PDB files: 1LKT and 1TYU (34,35). Each of the two domains was rigidly fitted as a trimer, as shown in Fig. 3 *A*. The fits are good, as evidenced by good *z*-scores and visual match between the fitted structures and the observed density. Secondary structure elements do not appear to be resolved in this component, which is to be expected given its resolution of 10.9 Å calculated by gold standard $FSC_{0.143}$.

Fig. 3 *F* shows single chains from each of the two crystal structures, representing the N-terminal head-binding domain and the C-terminal adhesin domain of a single tail-spike protein. Because the two domains come from different crystal structures, the two chains are not connected and a gap can be seen between them. The missing residues are added so as to create a complete tailspike (gp9) protein as described in a subsequent section. Fig. 3 *G* shows the surface of a single gp9 protein (under the hub/gp4 protein), with coloring representing *resmap* resolutions, which appear to vary between ~8 and 16Å.

Needle proteins

The needle proteins take the form of a trimer, where all three proteins have a mostly α-helical structure curled up into a triple helix. The docked crystal model, PDB: 2POH (36), is shown in Fig. 3 *H*, along with a surface enclosing density segmented for these proteins. These densities could not be segmented into separate proteins, because the resolution is quite low for this component (10.5 Å), and the three protein chains are tightly wound around each other into a triple-helix. Coloring of the surface with resolutions from *resmap* (Fig. 3 *I*), shows resolutions varying between 10 and 16Å.

Loop modeling

A 28 amino-acid (AA) loop (residues 464–492) of the portal protein (gp1), is missing in the crystal structure, likely because this loop is disordered and thus has many possible conformations. The cryoEM map shows some densities, which can be attributable to this loop (Fig. 3 *E*). This putative density is not noise because it can be seen in both independent maps at a lower (but above-noise) threshold level. Because this loop is quite long, resulting loop conformations can be varied. We built 10 different loop models (Movie S2), and each of the 10 full portal protein models was flexibly fitted using MDFF (also described in more details in Methods, and shown in Movie S3).

The same procedure was used to build three smaller missing loops in the tailspike protein (gp9). One of the loops (residues 110–112) was added to connect one chain from the head-binding domain trimer (PDB: 1LKT), to one chain from the adhesin-domain trimer (PDB: 1TYU). The two domains of the same protein were solved using x-ray crystallography separately, and hence there are two PDB files, one for each domain. The other two missing loops (residues 401–406 and 509–513) were also added to the tail domain (PDB: 1LKT). Thus we produced a complete model of the tailspike protein based on two independent crystal structures for its two domains and the cryoEM density.

Flexible fitting: Validation and test for overfitting

MDFF applies forces to each atom position in the model in the direction of the gradient, and hence the resulting model tends to match the density better than the initial model after the simulation. However this could also result in overfitting as the full atomic model has many degrees of freedom (three per atom, represented as each atom's position in space). The main parameter that can be varied with MDFF is the gradient scale, which determines the strength of the force applied on the model in the direction of the density gradient. The value recommended by the MDFF tutorial is 0.3; larger values can result in overfitting, as the forces because of the density gradient on each atom become correspondingly larger. We tested values of 0.1 to 0.9 and 500.

The protocol we used to validate the resulting model and to test for overfitting is illustrated in Fig. 4. The process involves 1) masking the density of interest (in this case one protein from the portal) from both independent maps, resulting in maps A and B; 2) rigid docking of the crystallographic model to map A; 3) loop modeling and flexible fitting based on map A; and 4) comparison of the resulting model with map B by computing the FSC between them.

The initial model and cryoEM densities are shown in Fig. 5 *A*. Below them, three FSC plots are shown: protein map A to protein map B, initial model to map A, and initial model to map B. The FSC of map A to map B (the signal curve) shows the structural information that we have about the portal protein. The FSC plot of initial model to map A is well below the signal curve, because the crystal structure of the model is not the same as the structure seen by cryoEM. The FSC plot of the initial model to Map B is very similar to the FSC plot of the initial model to Map A.

Fig. 5 *B* and Movie S3 show one of the resulting models after flexible fitting with MDFF, using a (moderate) gradient scale of 0.3. Visually, the model fits the density better in several parts, primarily at the start of the barrel domain, shown with an arrow and text in Fig. 5 *B*. This change is similar to that reported previously, as a narrowing of the

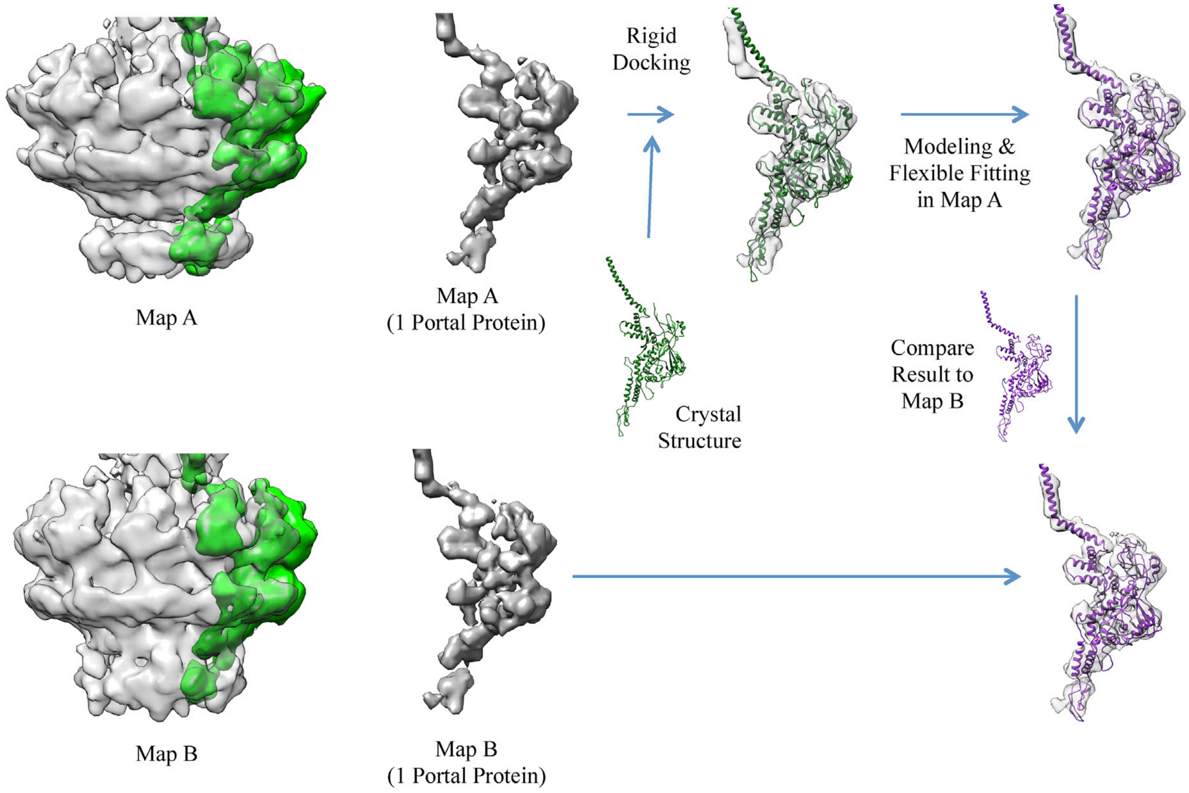

FIGURE 4 Flexible fitting validation protocol using two independent maps (A and B). On the far left, entire portals extracted from the two independent maps are shown, with a single protein colored in green. Map B is aligned to map A, though in the images the densities are shown separately. The protocol involves docking of a crystal structure and modeling based on map A (*top row*), then evaluating the resulting model by FSC plot to both maps A (*top row*) and B (*bottom row*). To see this figure in color, go online.

barrel part of the portal (21). The FSC plot for the fitted model to map A in Fig. 5 *B* is now much closer to the signal curve, though it has not overpassed it (hence no overfitting has occurred). The FSC plot for the fitted model to map B is still very similar to the FSC plot of the fitted model to map A, meaning that the fitted model is still valid in the context of the observed density from the independent map B (i.e., the independent map B is used to cross-validate the resulting model). Similar results were obtained with gradient scales of 0.1 to 0.9 (data not shown), in that the fitted model appeared to stay valid and did not overfit to density; only minor differences in the final model and FSC plots were observed.

Fig. 5 *C* shows the model after flexible fitting with MDFF, but using an extremely high gradient scale of 500. At such a large gradient scale, very large forces are applied at each atom in the direction of the density gradient. Comparing the resulting model with map A, the FSC plot is now very much above the signal curve (especially at higher frequencies), meaning that the fitted model has been overfitted (mostly to high-frequency noise). Moreover, the FSC plot for the fitted model to map A is also no longer similar to the FSC of the fitted model to map B, hence the fitted model no longer agrees with the (cross-validating) independent data set (map B).

Probabilistic models

A probabilistic model of a protein consists of 1) the average model (the one model from the 10 resulting models with minimum sum of distances from each Cα position to the average Cα position among the 10 models), and 2) the standard deviations at each Cα position among the 10 results. The standard deviation reflects the likelihood or probability that the position of the Cα atom in each residue is in a given area of space, based on the observed density in which the models are generated, and the modeling procedure itself.

Based on the 10 samplings for the portal protein, we tested the assumption that distances from the Cα atom in each residue to the position of the Cα atom in the same residue in the average model follow a normal distribution. The assumption was tested using a normal probability plot (37) (Fig. S2). The plots show that this assumption is valid based on the obtained samplings.

Probabilistic models of the portal, hub and tailspike proteins

Probabilistic models for the portal, hub, and tailspike proteins are shown in Fig. 6, *A* and *C*. The standard deviation at each backbone atom is coded in the following two

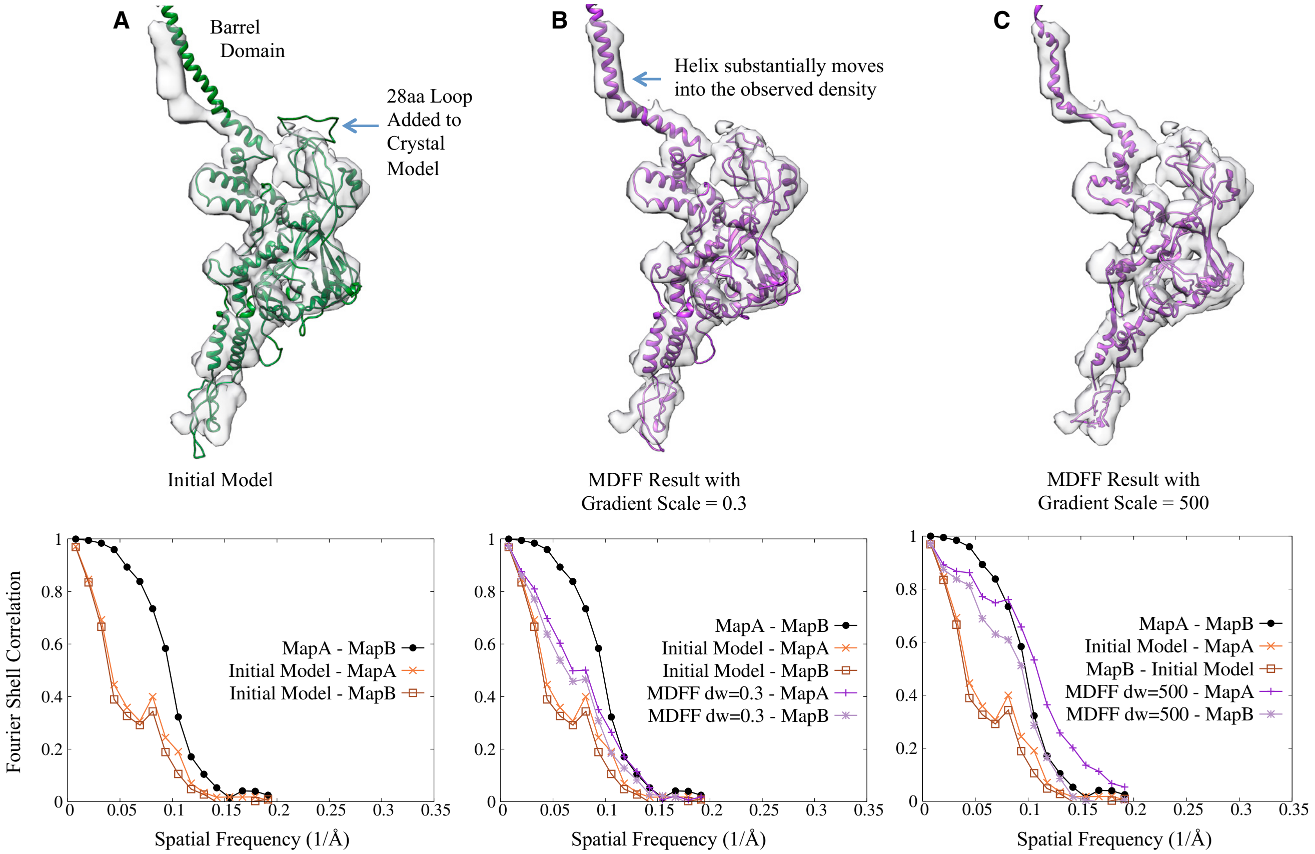

FIGURE 5 The top row shows segmented densities for a single protein from the portal as a transparent surface, the initial, rigidly docked model (*A*) with green ribbon, and the model after flexible fitting with MDFF with two different gradient weights of 0.3 (*B*) and 500 (*C*). On the bottom row, the FSC between simulated maps for each model and maps A and B is plotted, along with FSC between map A and map B (a single protein was smooth-masked in both maps). After flexible fitting with gradient scale of 0.3, the FSC between the model and map A and map B improves but stays below the signal plot between map A and map B, hence no overfitting took place. Using an extremely large gradient scale of 500, the FSC between the fitted model and map A (in which it is fitted) increases more dramatically, but the FSC between the fitted model and map B stays below the signal plot, indicating overfitting took place. The fitted model in this case also looks extremely distorted. To see this figure in color, go online.

ways: 1) by thickness, with thicker tube corresponding to higher standard deviations; and 2) by color, with blue representing lower standard deviations and red representing higher standard deviations. Movie S4 shows the 10 models for the portal protein built with loop modeling and flexible fitting. Movie S5 illustrates the probabilistic model of the portal protein rotating about the vertical axis.

In the figures and movies, the probabilistic models built from the 10 results in map A are shown. The probabilistic models built using five results from map A and five results from map B are nearly identical, i.e., the standard deviations at each backbone atom are extremely similar. Thus uncertainties because of differences between the two independent maps did not appear to influence the probabilistic models to a larger degree than the modeling process itself did.

In the resulting probabilistic models, parts of the models with secondary structure elements appear to have lower standard deviations, whereas more flexible loop regions appear to have higher standard deviations. In the portal protein, the highest standard deviations are in the long 28AA loop that was missing in the crystal model (Fig. 6 *A*). This is not surprising, because the density is not well resolved in that area, and does not tightly constrain the long, flexible loop. The standard deviations in the portal protein range between 0.3 and 13.7 Å. On the other hand, standard deviations in the hub and tail proteins range between 0.2 and 6.6 Å. The probabilistic models of the portal, hub, and tailspike proteins have been deposited to the Protein Data Bank (PDB) under accession number PDB: 5GAI. The entry contains the average model and the standard deviations computed for each residue stored in the B-factor column.

DISCUSSION

CryoEM density maps of large molecular complexes can have varying resolutions across different components. Estimating resolutions of subcomponents can be very useful in understanding the structural information provided by cryoEM reconstructions. In this study, we showed how these resolutions can be estimated by using gold-standard FSC

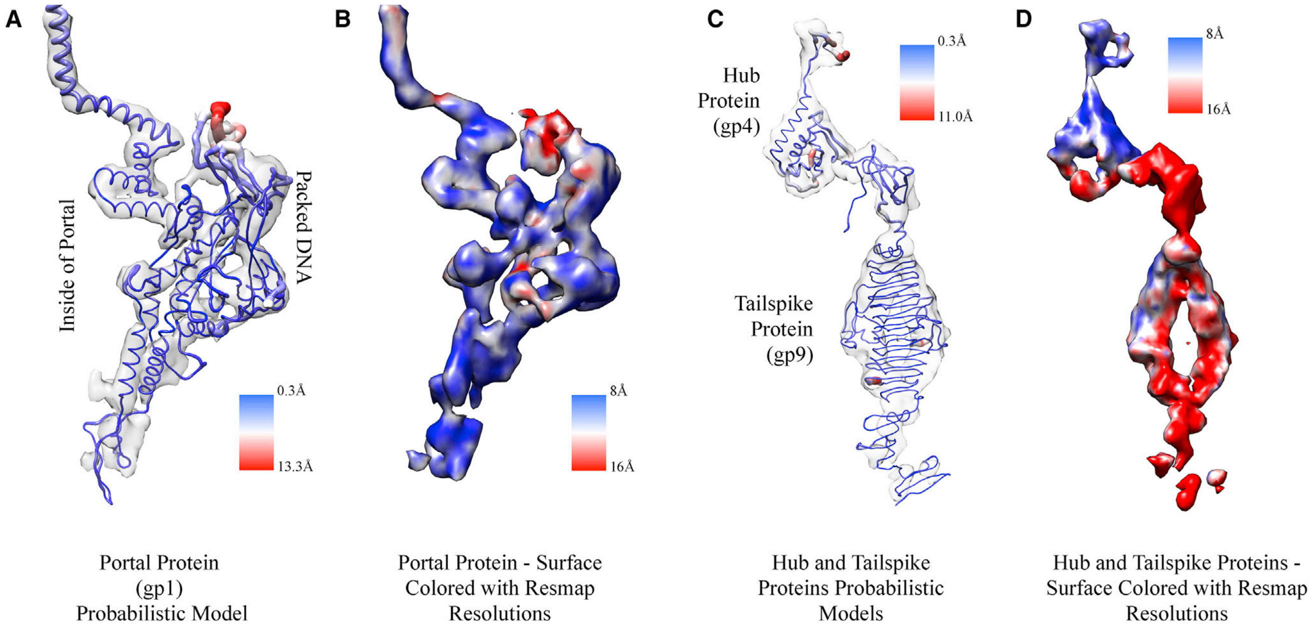
A
Inside of Portal
Packed DNA
0.3Å
13.3Å
Portal Protein (gp1) Probabilistic Model
B
8Å
16Å
Portal Protein - Surface Colored with Resmap Resolutions
C
Hub Protein (gp4)
Tailspike Protein (gp9)
0.3Å
11.0Å
Hub and Tailspike Proteins Probabilistic Models
D
8Å
16Å
Hub and Tailspike Proteins - Surface Colored with Resmap Resolutions

FIGURE 6 (*A* and *C*) Segmented protein densities are shown with a transparent surface, whereas probabilistic models are shown as a worm model. Ribbon thickness and color correspond to standard deviations at each backbone residue. (*B* and *D*) The segmented protein densities for the proteins are also shown color-coded with *resmap* resolutions for comparison. Although higher standard deviations appear to correspond to lower resolutions for the portal protein, in the case of the hub and tailspike proteins, standard deviations are lower even at lower resolutions. Thus, standard deviations are not only influenced by resolution alone, but also by structural composition and the fitting method.

plots and comparing masked regions in the map. The software *resmap* (7) was also used to calculate per-voxel resolutions through the entire map; the surfaces colored with these values show reasonable consistency with the values calculated with the $FSC_{0.143}$ criterion. It is not clear at this point whether resolutions from these two methods can be more directly or quantitatively compared, and this remains a topic of potential future interest. In our view, it is very useful to be able to apply both methods for different purposes; gold standard-FSC is useful for having a meaningful number for the entire map and for segmented components, whereas looking at voxel-based resolution as given by *resmap* is useful in looking at how the resolution varies throughout each individual component.

It is very common to interpret medium-resolution cryoEM density maps, as presented here, by docking and flexible fitting of crystal models. In previous work, we focused on validating rigid body docking of crystal models into medium-resolution density maps (8). In this study, we focused on validation of the model after flexible fitting, by using two independent reconstructions. Only one of the reconstructions was used for flexible fitting, and then the other independent reconstruction was used to validate and test for overfitting. From our tests, avoiding overfitting with MDFF in particular seems to be tied directly to the gradient-scale parameter used. When using a small gradient-scale parameter, overfitting was not detected, and the resulting model's resemblance to the cross-validating, independent data set was good. However, when using a very large gradient scale, the resulting model appeared distorted (Fig. 4 *C*), and our tests showed that it was overfitted (mostly to noise) and also inconsistent with the independent, cross-validating data set (the independent reconstruction that was not used while fitting).

Because many feasible results can be obtained after flexible fitting in medium-resolution density maps, we have shown how probabilistic models can capture the uncertainty of a model, given the medium-resolution map in which it is interpreted. Interestingly, the uncertainties do not appear to be only tied to the resolution of the density. For example, as Fig. 6 shows, although the tailspike (gp9) protein appears at lower resolution than the hub (gp4) protein, it actually has similar standard deviations across multiple flexible fitting results. This may be because the extensive secondary structure elements in the tailspike (gp9) protein make it less flexible. Thus, the uncertainty or standard deviation at each residue is not only influenced by the resolution of the density, but also by the flexibility of the protein allowed to it during modeling.

The flexibility of a protein being modeled is dictated by 1) how much flexibility the modeling method allows, which may restrain it in certain ways for various reasons, e.g., to prevent overfitting; 2) secondary structures and intraprotein contacts; and 3) contacts between it and any adjacent proteins it may be in complex with. Points 1 and 2 in the context of MDFF were described in the methods. In terms of point 3, ideally the entire complex would be used for building the probabilistic models. Although this was impractical in this case as some of the proteins structures are not known, and the packaged DNA itself is extremely large, we attempted to simulate as much of each subcomponent as possible: the entire portal and hub were simulated together (24 proteins in total), and an entire tailspike trimer was simulated (three proteins), as each trimer appears to have little if any contact with other trimers.

It is important to note that the uncertainty reflected by a probabilistic model is not meant to reflect the accuracy of the resulting models. Although the initial model, typically obtained by x-ray crystallography, can be accurate to a much higher atomic resolution, here it is fitted to less accurate, medium-resolution density maps. An interesting question in the field is whether the modeling methods themselves (MD and flexible fitting in particular) can provide models that are indicative of native-like conformations with accuracy beyond what is seen in the density map. A more rigorous proof of this hypothesis requires further experimental and computational validation.

With that caveat in mind, in the case of P22 phage in this study, some interesting new insights, to our knowledge, can be deduced from probabilistic models. For example, the portal protein appears to have smaller standard deviations in the core region (annotated in Fig. 6 *A*), where it has to stay rigid as DNA gets pushed into the virion, whereas larger standard deviations are seen in the outer parts and the long loop region, where it has to interact with varied conformations of packaged DNA. On the other hand, the probabilistic models for the pseudo-C12-symmetric hub, which connects the portal to the six tailspike timers, appears to display lower standard deviations (Fig. 6 *C*). A rotational cross correlation plot (Fig. S2) shows the 12 hub proteins themselves may be somewhat different from each other, adjusting accordingly so that they can bind tightly to each pseudo-C6-symmetric tailspike trimer. The tailspike protein probabilistic model also appears to show lower standard deviations (Fig. 6 *C*), despite the corresponding density appearing to have lower resolution. This may mean that although they are flexible to move as a rigid body with respect to the stable hub they connect to (leading to lower resolvability across many averaged particles), they likely stay rather rigid in their overall conformation while doing so.

To conclude, cryo-EM continues to reveal interesting close-to-native organizations of complex protein assemblies, such as in the example used here, the mature form of the P22 bacteriophage. Estimating resolution of various components allows us to better understand the accuracy of the reconstruction with respect to each functional unit within the complex. Independent iterations of loop modeling and flexible fitting, ensuring overfitting did not occur, were combined into probabilistic models, which capture not only the uncertainty in the observed density with which they are built but also the structural properties of the proteins themselves. This procedure for analyzing density maps will be useful in other cases as well.

SUPPORTING MATERIAL

Two figures and five movies are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)04706-2.

AUTHOR CONTRIBUTIONS

D.H.C. collected the data and determined the 3D cryoEM density maps; G.P. did the resolution estimation and modeling of subcomponents. J.A.K. and C.H.P. provided the samples. W.C. conceived the research. All authors wrote the manuscript.

ACKNOWLEDGMENTS

This research has been supported by NIH grants (P41GM103832, PN2EY016525, and R01GM079429). Molecular graphics and analyses were performed with the UCSF Chimera package, along with the Segger extensions as mentioned throughout the manuscript. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).

REFERENCES

1. Liao, H. Y., and J. Frank. 2010. Definition and estimation of resolution in single-particle reconstructions. *Structure.* 18:768–775.
2. Grigorieff, N. 2000. Resolution measurement in structures derived from single particles. *Acta Crystallogr. D Biol. Crystallogr.* 56:1270–1277.
3. Henderson, R., A. Sali, ..., C. L. Lawson. 2012. Outcome of the first electron microscopy validation task force meeting. *Structure.* 20:205–214.
4. Scheres, S. H. W., and S. Chen. 2012. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods.* 9:853–854.
5. Chen, S., G. McMullan, ..., R. Henderson. 2013. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy.* 135:24–35.
6. Cardone, G., J. B. Heymann, and A. C. Steven. 2013. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* 184:226–236.
7. Kucukelbir, A., F. J. Sigworth, and H. D. Tagare. 2014. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods.* 11:63–65.
8. Pintilie, G., and W. Chiu. 2012. Comparison of Segger and other methods for segmentation and rigid-body docking of molecular components in cryo-EM density maps. *Biopolymers.* 97:742–760.
9. Topf, M., K. Lasker, ..., A. Sali. 2008. Protein structure fitting and refinement guided by cryo-EM density. *Structure.* 16:295–307.
10. Schroder, G. F., A. T. Brunger, and M. Levitt. 2007. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure.* 15:1630–1641.
11. Trabuco, L. G., E. Villa, ..., K. Schulten. 2008. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure.* 16:673–683.
12. Falkner, B., and G. F. Schröder. 2013. Cross-validation in cryo-EM-based structural modeling. *Proc. Natl. Acad. Sci. USA.* 110:8930–8935.
13. DiMaio, F., J. Zhang, ..., D. Baker. 2013. Cryo-EM model validation using independent map reconstructions. *Protein Sci.* 22:865–868.
14. Eswar, N., B. Webb, ..., A. Sali. 2007. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* http://dx.doi.org/10.1002/0471250953.bi0506s15.

15. DiMaio, F., M. D. Tyka, …, D. Baker. 2009. Refinement of protein structures into low-resolution density maps using Rosetta. *J. Mol. Biol.* 392:181–190.

16. Baker, M. L., M. R. Baker, …, W. Chiu. 2012. Gorgon and pathwalking: macromolecular modeling tools for subnanometer resolution density maps. *Biopolymers.* 97:655–668.

17. Pandurangan, A. P., S. Shakeel, …, M. Topf. 2014. Combined approaches to flexible fitting and assessment in virus capsids undergoing conformational change. *J. Struct. Biol.* 185:427–439.

18. Farabella, I., D. Vasishtan, …, M. Topf. 2015. TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. *J. Appl. Crystallogr.* 48:1314–1323.

19. King, J., D. Botstein, …, E. Lenk. 1976. Structure and assembly of the capsid of bacteriophage P22. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 276:37–49.

20. Chen, D.-H., M. L. Baker, …, W. Chiu. 2011. Structural basis for scaffolding-mediated assembly and maturation of a dsDNA virus. *Proc. Natl. Acad. Sci. USA.* 108:1355–1360.

21. Tang, J., G. C. Lander, …, J. E. Johnson. 2011. Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in P22. *Structure.* 19:496–502.

22. Bammes, B. E., R. H. Rochat, …, W. Chiu. 2011. Practical performance evaluation of a 10k × 10k CCD for electron cryo-microscopy. *J. Struct. Biol.* 175:384–393.

23. Kivioja, T., J. Ravantti, …, D. Bamford. 2000. Local average intensity-based method for identifying spherical particles in electron micrographs. *J. Struct. Biol.* 131:126–134.

24. Yang, C., W. Jiang, …, W. Chiu. 2009. Estimating contrast transfer function and associated parameters by constrained non-linear optimization. *J. Microsc.* 233:391–403.

25. Ludtke, S. J., P. R. Baldwin, and W. Chiu. 1999. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* 128:82–97.

26. Liu, X., R. H. Rochat, and W. Chiu. 2010. Reconstructing cyanobacteriophage P-SSP7 structure without imposing symmetry. *Protoc. Exch.* 10

27. Pintilie, G. D., J. Zhang, …, D. C. Gossard. 2010. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* 170:427–438.

28. Pettersen, E. F., T. D. Goddard, …, T. E. Ferrin. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25:1605–1612.

29. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38.

30. MDFF Tutorial. http://www.ks.uiuc.edu/Training/Tutorials/science/mdff/tutorial_mdff-html/. Accessed: December 17, 2015.

31. Becker, O. M., A. D. MacKerell, Jr., …, M. Watanabe. 2001. Computational Biochemistry and Biophysics. Marcel Dekker, New York.

32. Chang, J., X. Liu, …, W. Chiu. 2012. Reconstructing virus structures from nanometer to near-atomic resolutions with cryo-electron microscopy and tomography. *Adv. Exp. Med. Biol.* 726:49–90.

33. Olia, A. S., P. E. Prevelige, Jr., …, G. Cingolani. 2011. Three-dimensional structure of a viral genome-delivery portal vertex. *Nat. Struct. Mol. Biol.* 18:597–603.

34. Steinbacher, S., S. Miller, …, R. Huber. 1997. Phage P22 tailspike protein: crystal structure of the head-binding domain at 2.3 A, fully refined structure of the endorhamnosidase at 1.56 A resolution, and the molecular basis of O-antigen recognition and cleavage. *J. Mol. Biol.* 267:865–880.

35. Steinbacher, S., U. Baxa, …, R. Huber. 1996. Crystal structure of phage P22 tailspike protein complexed with Salmonella sp. O-antigen receptors. *Proc. Natl. Acad. Sci. USA.* 93:10584–10588.

36. Olia, A. S., S. Casjens, and G. Cingolani. 2007. Structure of phage P22 cell envelope–penetrating needle. *Nat. Struct. Mol. Biol.* 14:1221–1226.

37. Chambers, J. M., W. S. Cleveland, …, P. A. Tukey. 1983. Graphical Methods for Data Analysis. Chapman and Hall/CRC, Belmont, CA.

Biophysical Journal, Volume *110*

Supplemental Information

Resolution and Probabilistic Models of Components in CryoEM Maps of Mature P22 Bacteriophage

Grigore Pintilie, Dong-Hua Chen, Cameron A. Haase-Pettingell, Jonathan A. King, and Wah Chiu

Resolution and Probabilistic Structural Models of Subcomponents Derived from CryoEM Maps of Mature P22 Bacteriophage

Grigore Pintilie[1], Dong-Hua Chen[1,*], Cameron A. Haase-Pettingell[2], Jonathan A. King[2], Wah Chiu[1]

Supporting Material

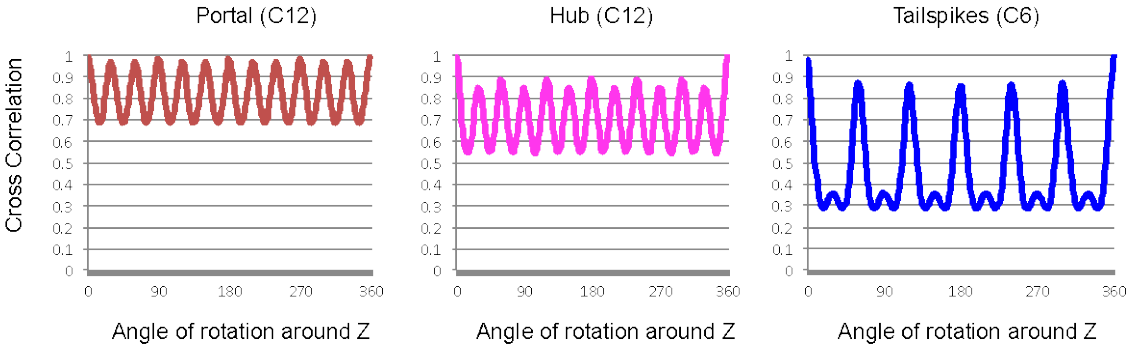

Figure S1. Rotational cross-correlation through individual components extracted from the entire map – portal, hub and tailspikes. The portal and hub have C12 symmetry, thus 12 peaks are seen the CC plot as the densities are rotated through 360° around the z-axis. For the portal, every 3rd peak is practically at CC~=1.0, and all others are just slightly less; overall the 12 portal proteins are very similar. For the hub, another interesting pattern can be seen: every second peak (except the one at Angle=0) has CC=~0.89, and all others have CC~=0.85. Because the peak at Angle=0 is the highest, this means that each protein in the hub can be somewhat different from all others. This may be explained by the fact that the C12 hub has to connect the C12 portal to the C6 tailspike trimers. The C6 tailspike trimers' density has 6 major peaks in the CC rotational plot – each peak except at Angle=0 has CC ~=0.88. Thus, the 6 trimers are similar but not exactly the same in conformation.

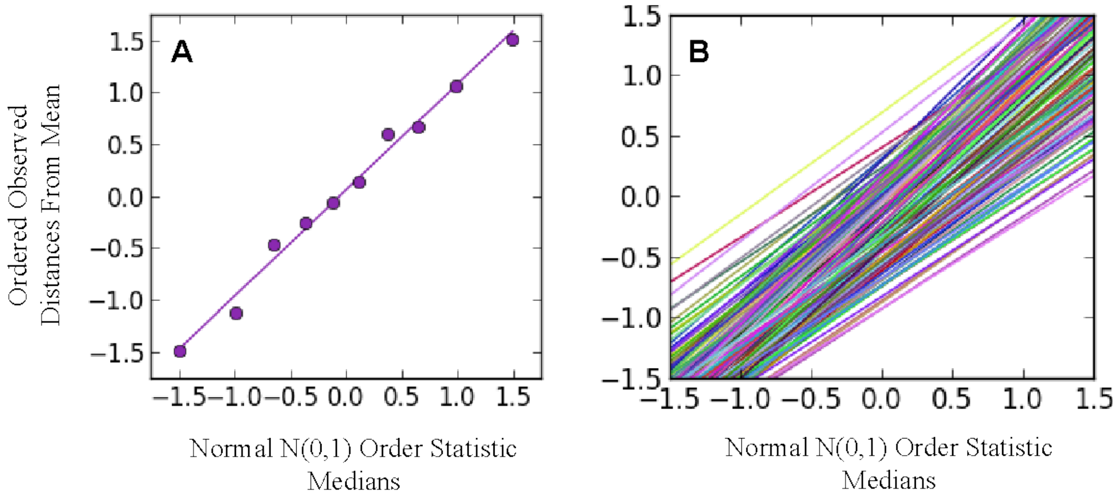

Figure S2. Normal probability plot to test whether backbone atoms positions can be represented with a normal distribution. (A) A shows a normal probability plot for the backbone atom position of a single residue, showing data points (circles), and line fitted to the points. (B) Lines fitted to sample points for all residues. Lines that have ~45° slope indicate that the samples fall very close to a normal distribution.