De novo protein structure determination from near-atomic-resolution cryo-EM maps

Ray Yu-Ruei Wang[1,2], Mikhail Kudryashev[3,4], Xueming Li[5,8], Edward H Egelman[6], Marek Basler[3], Yifan Cheng[5], David Baker[2,7] & Frank DiMaio[2]

We present a *de novo* model-building approach that combines predicted backbone conformations with side-chain fit to density to accurately assign sequence into density maps. This method yielded accurate models for six of nine experimental maps at 3.3- to 4.8-Å resolution and produced a nearly complete model for an unsolved map containing a 660-residue heterodimeric protein. This method should enable rapid and reliable protein structure determination from near-atomic-resolution cryo-electron microscopy (cryo-EM) maps.

Model building is a key step in macromolecular structure determination. Whereas most atomic-resolution structures are solved using X-ray crystallography, single-particle cryo-EM has emerged as a powerful tool in determining electron density maps of large and high-symmetry particles to near-atomic resolution (3–5 Å; ref. 1). Recent advances even allow cryo-EM to reach these resolutions from smaller particles with low or no symmetry[2–4]. Despite these developments, little progress has been made in *de novo* model building into near-atomic-resolution cryo-EM density maps. Structural interpretation of cryo-EM maps typically starts with fitting an atomic X-ray or NMR structure into the map[5]. Recent work has shown that atomic-resolution models are achievable from near-atomic-resolution cryo-EM density, starting from a homologous structure of the correct topology[6]. However, when there are no previously solved structures of homologous proteins, *de novo* model building must be carried out. Currently, *de novo* structure determination requires manually building a backbone model into density and assigning sequence[2,7,8]. Although tracing the backbone into density at this resolution is generally straightforward, manually assigning sequence remains time consuming and error prone.

Automated protein model-building tools developed for X-ray crystallography[9,10] are widely used in structure determination from maps with resolution better than 3 Å. These methods separate backbone tracing and side-chain assignment, with density features largely guiding side-chain identification. Consequently, at resolutions worse than 3 Å, where side-chain density is mostly indiscernible, these approaches generally fail. Several *de novo* model-building methods targeted to cryo-EM have been developed for maps with resolution ranging from near-atomic (3–5 Å) to medium resolution (5–10 Å) (refs. 11,12). Although these methods are powerful in identifying the protein topology given a map, they have poor recovery, often <50%, of correct sequence registration[11,12].

Here we describe a novel *de novo* model-building approach for cryo-EM maps at 3- to 5-Å resolution. Our approach combines the agreement of sequence-derived predicted backbone conformations to local density with the agreement of the sequence to side-chain density in order to accurately assign sequence into density. On a benchmark set of nine experimental cryo-EM maps at near-atomic resolution with previously determined structure and a previously unsolved map for the 660-residue contractile sheath protein of the type VI secretion system from *Vibrio cholerae*, we show that high-accuracy models can be obtained without knowledge of detectable structural homologs.

Our approach for *de novo* interpretation of near-atomic-resolution density maps consists of three steps: (i) matching sequence-based local backbone conformations into the density map, (ii) identification of a maximally consistent subset of these fragment matches and assembly into a partial model and (iii) completion of the partial model using density-guided sampling and all-atom refinement (**Fig. 1a**). In the first step, for overlapping nine-residue windows of amino acid sequence, we identify segments ('fragments') of solved protein structures with similar local sequences and predicted secondary structures, analogous to the fragments used in Rosetta *de novo* structure prediction[13]. For each fragment, a translation-rotation search identifies placements with good map agreement after optimizing side-chain conformations; only a small subset of these placements are located near the native position (r.m.s. deviation (RMSD) < 2.5 Å). To identify these correct placements, we search a mutually compatible subset of fragment placements, using a score function that—in addition to preferring fragments that fit well into density—favors fragment pairs with (i) the same residue in the same place, (ii) residues nearby in sequence nearby in space and (iii) no two residues occupying the same space. Monte Carlo simulated annealing (MC-SA) guided by this score function finds the maximally

[1]Graduate program in Biological Physics, Structure and Design, University of Washington, Seattle, Washington, USA. [2]Department of Biochemistry, University of Washington, Seattle, Washington, USA. [3]Focal Area Infection Biology, Biozentrum, University of Basel, Basel, Switzerland. [4]Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Basel, Switzerland. [5]The Keck Advanced Microscopy Laboratory, Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, USA. [6]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, USA. [7]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [8]Present address: School of Life Science, Tsinghua University, Beijing, China. Correspondence should be addressed to F.D. (dimaio@u.washington.edu).

RECEIVED 19 SEPTEMBER 2014; ACCEPTED 31 DECEMBER 2014; PUBLISHED ONLINE 23 FEBRUARY 2015; DOI:10.1038/NMETH.3287

consistent subset of fragment placements from this larger set. Fragment matching and MC-SA assembly are applied iteratively until >70% of the sequence has been assigned into density. Each iteration places fragments from previously unassigned sequence positions of the sequence into previously unoccupied regions in density (**Fig. 1b**). Finally, the partial model from the final iteration is completed through rebuilding and all-atom refinement using RosettaCM[14] guided by the experimental density data.

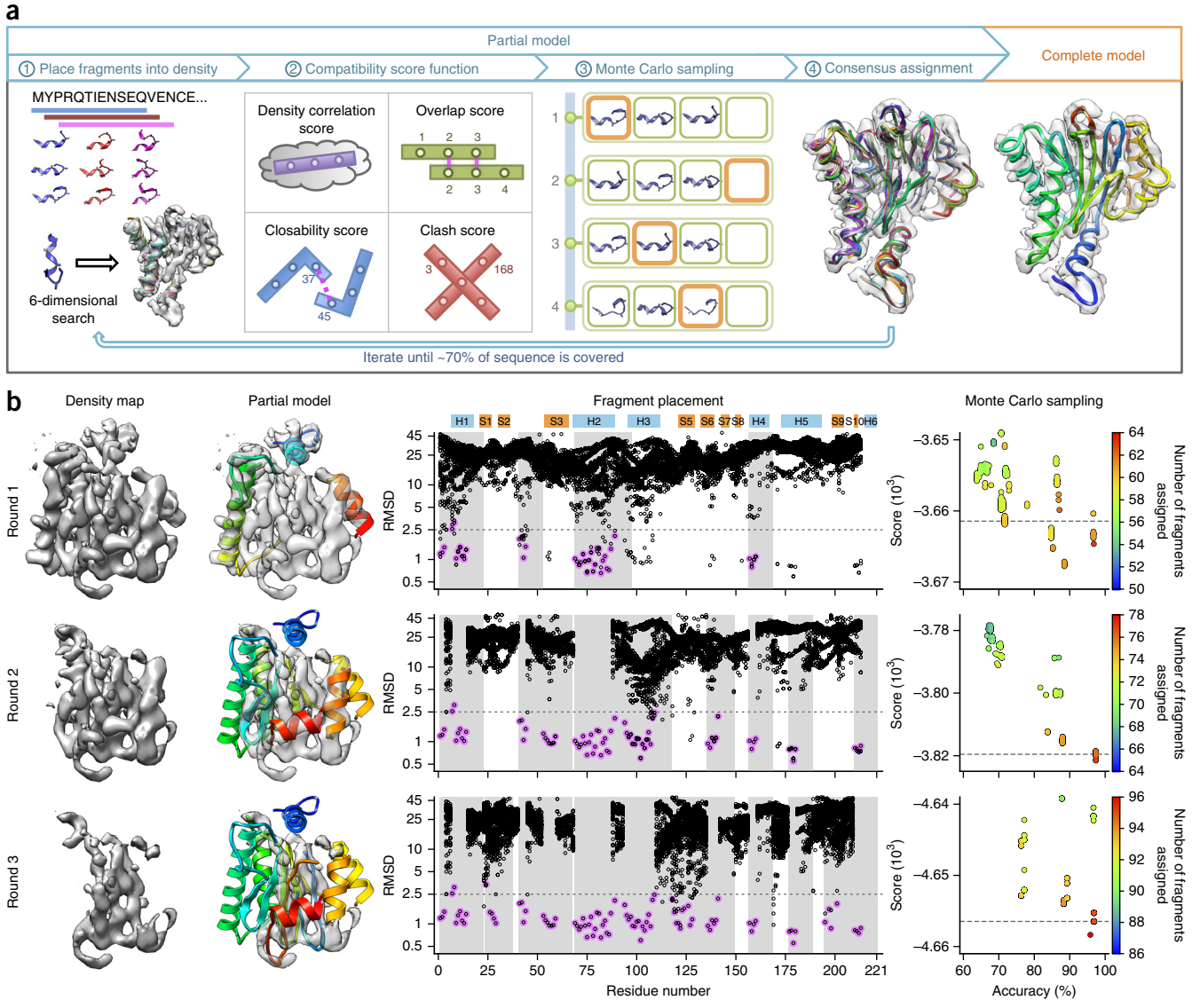

Figure 1 | Protocol overview. (**a**) First, for a nine-residue window centered on each position in the sequence, representative backbone conformations (fragments) are collected and docked into the density map. Second, the resulting fragment placements are then evaluated using a score function consisting of four terms: a density correlation term assessing the agreement of fragment and map; an overlap term favoring fragment pairs assigning the same residue to the same location; a closability term favoring fragment pairs close in sequence that are close in space; and a clash term preventing two residues from occupying the same place. Third, from the candidate placements (green squares), Monte Carlo simulated annealing finds a set of fragments (orange squares) optimizing the score function; a null placement (empty squares) may be assigned in positions where no good placements have been identified. Fourth, a partial model is assembled by combining fragment placements from multiple Monte Carlo trajectories. Steps 1–4 are carried out iteratively until ~70% of sequence is covered. Finally, unassigned regions in the final partial model are completed using density-guided loop sampling followed by all-atom refinement. (**b**) Model building for the 20S α-subunit in a 4.8-Å resolution cryo-EM map required three iterations, illustrated in the three rows in the figure. The far left column shows the density map used for the corresponding iteration after density from the previous round's partial model was masked out. The next column shows the assembled partial models after Monte Carlo sampling (colored blue at the N terminus to red at the C terminus). Center, fragment placement results after translation and rotation search. The *x* axis covers the sequence of the protein, and each black point represents a single fragment placement; the *y* axis indicates the distance of the fragment placement from the native conformation. Pink points indicate fragments chosen to assemble the partial model, and the gray shading shows residues covered in the partial model. Secondary structural elements in the native protein are indicated above the plot: H, helix; S, strand. Right, the lowest-scoring Monte Carlo trajectories (below the dotted line) are combined to provide the starting model for the next iteration. Each point represents the fragment assignment of an independent search trajectory, colored by number of total fragments placed. The *x* axis indicates the percentage of fragments placed within 2.5-Å RMSD from the native configuration; the *y* axis shows the score with the fragment compatibility function. The horizontal dashed line shows the score cutoff used for partial model generation.

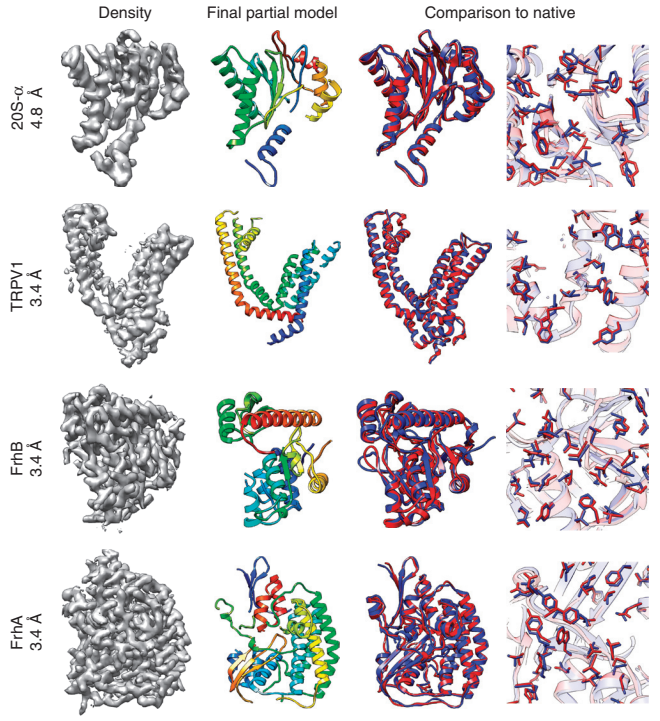

Figure 2 | High-accuracy model building in near-atomic-resolution cryo-EM maps. Far left, density maps used for *de novo* model building on 20S-α, TRPV1, FrhB and FrhA. Left, partial model at the final iteration. Right and far right, full-length RosettaCM models (red) superimposed with the experimentally determined structures (blue). Shown is the lowest-RMSD structure for each protein, determined from the ten models with the lowest electron density scores (right), with a close-up of the core showing that native core packing is recovered (far right).

We tested our method on a benchmark set of nine proteins whose structures have been determined through cryo-EM (TMV, TRPV1, FrhB and BPP1) or X-ray crystallography (20S-α, FrhA, FrhG, VP6 and STIV). These proteins range in size from 155 to 397 residues, include different fold types and have experimental cryo-EM maps varying in resolution range from 3.3 to 4.8 Å (**Supplementary Table 1**). For each map, a single subunit was first segmented from the entire density map. Fragments from proteins with similar topology or sequence were excluded while constructing the fragment libraries. In seven of the nine cases, partial models from the final iteration of the *de novo* building step were within 1.1- to 2.3-Å Cα RMSD from the experimental structures (**Fig. 2** and **Supplementary Table 1**), six of which were more than 70% complete. These partial models were then completed and refined using RosettaCM, yielding models with 1.3- to 2.2-Å Cα RMSD (2.0- to 3.1-Å all-atom RMSD) from the experimentally determined structures. In some cases, when completing partial models, RosettaCM was able to fix errors resulting from the initial fragment placement (**Supplementary Fig. 1**). In contrast, Buccaneer[9]—a widely used model-building method from X-ray crystallography—although able to trace portions of the backbone for all targets, correctly identified more than 5% of the sequence in only three cases and never identified more than 50% (**Supplementary Table 2**).

Among the proteins in the benchmark set, TRPV1 (ref. 2) and FrhB[7] were proteins with new folds solved recently by manually building models into cryo-EM density. Our method automatically obtained completed models with 1.4-Å Cα RMSD model for TRPV1 and 1.7-Å Cα RMSD for FrhB. To test the resolution limit at which *de novo* structure determination is possible, we used a previously unpublished 4.8-Å-resolution map from the 20S proteasome α-subunit (20S-α). At this resolution, the α-helix pitch is somewhat visible; however, β-strand separation is only barely resolved (**Fig. 1b**). With our approach, the final partial model had 196 of 221 residues placed, with just 1.3-Å RMSD to the crystal structure (**Figs. 1** and **2**, and **Supplementary Table 1**). Using RosettaCM to build a completed model, we obtained a 1.2-Å Cα RMSD model (2.0-Å all-atom RMSD). Despite the lack of side-chain density details, side-chains in the core of the protein showed very good agreement with the crystal structure (**Fig. 2**).

As described above, we iterated fragment matching and assembly steps to improve coverage of the sequence assignment. This is because near-native placements of some fragments did not score well enough to be carried over to MC-SA assembly, even though they adopted near-native conformations (**Supplementary Table 3**). In each iteration, models are assembled from the consensus assignment of the lowest-scoring 5% of trajectories; these regions are locked, the corresponding density is masked out, and another round of fragment search and MC-SA is carried out. In all cases except one (TMV), more than one iteration was required to obtain a partial model with at least 70% of the sequence placed (**Supplementary Table 1**). For example, 20S-α took three rounds to reach this level of coverage; the partial model after one round had only 34% of the sequence placed (**Fig. 1b** and **Supplementary Table 1**). Sequence positions at S3, S6 and S7 were correctly traced only in the second round, and S1, S2, S5, S9 and S10 only in the third (**Fig. 1b**).

There were three cases (BPP1, STIV and VP6) in which our approach was unable to automatically determine accurate full-length structures. This was clearly identifiable by the poor coverage of the models after a single round of modeling (**Supplementary Table 1**). There are two main reasons for such failures. If a large portion of the protein does not have fragments that adopt near-native conformations, it is not possible to accurately assign positions for these residues into the map. BPP1 is one such case: almost half the sequence has no accurate fragments (**Supplementary Table 2**). Second, β-sheet assembly from

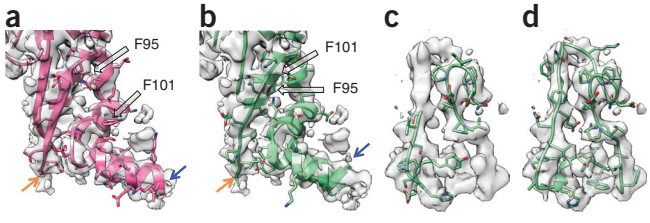

Figure 3 | Blind structure determination of the VipA/VipB complex from *V. cholerae*. (**a**,**b**) An error in the manually traced model (pink, **a**) is corrected by our method (green, **b**). The arrows in black show the positions of two residues in both models (F95 and F101), highlighting the six-residue registration shift between the models. Orange and blue arrows indicate the beginning and end of the region with the sequence registration discrepancy. (**c**) Partial trace generated by our method in a region where manual tracing was impossible. (**d**) The full-length RosettaCM model at the same region shows good agreement with the map.

fragments is difficult owing to the conformational variability of sheets compared to helices. STIV and VP6 are such cases in which we failed to accurately build sheets (**Supplementary Fig. 2**). These failures suggest possibilities for future method improvement.

We applied our method on a newly reconstructed cryo-EM map of the contractile sheath proteins of the type VI secretion system from *V. cholerae* (EMD-2699) at 3.5-Å resolution, with no detectable homologs of known structure. The asymmetric unit contained a heterodimer (VipA/VipB) with 660 residues total. After manually segmenting the map, eight iterations of our protocol generated a partial model with 466 residues placed. In parallel, the map was manually traced with the aid of Buccaneer, which placed a total of 513 residues. There was good overall agreement between the two models: over 394 residues, the Cα RMSD was 1.1 Å. However, there were 35 residues for which sequence registration was shifted by six positions between the models (**Fig. 3a**,**b**). The segment was flanked by disordered residues; this combined with the poor local resolution made sequence assignment particularly difficult. The sequence assignment made by our method showed better agreement with the density map than the hand-traced model in this region (**Supplementary Fig. 3**). We used RosettaCM to assemble full-length structures starting from both configurations. Among the low-energy models RosettaCM generated, the segment assigned by the automated method was exclusively chosen, suggesting our assignment was more energetically favorable and hence correct. Additionally, our approach was able to assign sequence in regions where the manual model did not (**Fig. 3c**). Combining our model with the manual model in RosettaCM, we were able to build a full-length model for the heterodimer complex (**Fig. 3d** and ref. 15). The blind case shows that our approach is tolerant to errors in segmentation; although our manual segmentation was imperfect, structure determination was still successful.

The key concept we use here, that local amino acid sequences have preferences for certain backbone conformations, has previously been used to predict structures of small proteins (<100 residues)[16] *de novo* and larger proteins using sparse backbone-only NMR data[17]. However, no previous approach in protein structure modeling has used this concept in conjunction with experimentally determined local Cartesian-space restraints to restrict conformational space. The method described here should provide a general framework for the use of these types of sparse experimental constraints in protein structure determination.

Several improvements will increase both the applicability and accuracy of our *de novo* approach. Our tests assumed a map in which the asymmetric unit was segmented. Although manual segmentation is often straightforward (as in the blind case), it may prove difficult in highly intertwined structures. Further improvements of the method on all-β proteins are also necessary: strand-pairing bonuses in the scoring function combined with more aggressive fragment optimization into density should improve accuracy with all-β proteins. Our approach is amenable to integrating additional structural information: known structures of components are easily incorporated, experimentally derived pairwise distance restraints may guide conformational sampling, and Cα traces can be provided by users. Our method should streamline the protein structure determination process from cryo-EM maps at near-atomic resolution, reducing human effort and errors due to human biases. The *de novo* protein structure determination method described here is freely available for academic use through the Rosetta software suite, available at https://www.rosettacommons.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. EMDataBank: 20S proteasome density map at 4.8 Å, EMD-6219.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank K. Laidig and D. Alonso for setting up and managing the computational resources. This work was supported by the US National Institutes of Health under award numbers R01GM092802 (R.Y.-R.W. and D.B.), EB001567 (E.H.E.) and R01GM098672 (Y.C.); the Swiss systems biology initiative SystemsX.ch grant CINA (M.K.); the University of California, San Francisco, Program for Breakthrough Biomedical Research (Y.C.); and Howard Hughes Medical Institute (D.B.). We thank Y. Liu for the use of his clustering algorithm.

AUTHOR CONTRIBUTIONS

R.Y.-R.W. performed the research and drafted the manuscript. F.D. and D.B. supervised the research and edited the draft. M.K., E.H.E. and M.B. provided the helical reconstruction map of the VipA/VipB complex and compared the automated model with their manually traced model initially. X.L. and Y.C. provided the 20S map at 4.8-Å resolution. All authors edited the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hryc, C.F., Chen, D.H. & Chiu, W. *Curr. Opin. Virol.* **1**, 110–117 (2011).
2. Liao, M., Cao, E., Julius, D. & Cheng, Y. *Nature* **504**, 107–112 (2013).
3. Bai, X.C., Fernandez, I.S., McMullan, G. & Scheres, S.H. *eLife* **2**, e00461 (2013).
4. Lu, P. *et al. Nature* **512**, 166–170 (2014).
5. Rossmann, M.G., Bernal, R. & Pletnev, S.V. *J. Struct. Biol.* **136**, 190–200 (2001).
6. DiMaio, F. *et al. Nat. Methods* doi:10.1038/nmeth.3286 (23 February 2015).
7. Allegretti, M., Mills, D.J., McMullan, G., Kühlbrandt, W. & Vonck, J. *eLife* **3**, e01963 (2014).
8. Amunts, A. *et al. Science* **343**, 1485–1489 (2014).
9. Cowtan, K. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1002–1011 (2006).
10. Terwilliger, T.C. *et al. Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
11. Baker, M.R., Rees, I., Ludtke, S.J., Chiu, W. & Baker, M.L. *Structure* **20**, 450–463 (2012).
12. Lindert, S. *et al. Structure* **17**, 990–1003 (2009).
13. Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E. & Baker, D. *PLoS ONE* **6**, e23294 (2011).
14. Song, Y. *et al. Structure* **21**, 1735–1742 (2013).
15. Kudryashev, M. *et al. Cell* doi:10.1016/j.cell.2015.01.037 (in the press).
16. Bradley, P., Misura, K.M. & Baker, D. *Science* **309**, 1868–1871 (2005).
17. Raman, S. *et al. Science* **327**, 1014–1018 (2010).

ONLINE METHODS

Map preparation. For all benchmark targets, the cryo-EM maps were segmented into single-subunit guided by native structures using UCSF Chimera's "zone" tool at a distance of 4 Å. The cryo-EM maps and the corresponding deposited native structures used are listed in **Supplementary Table 1**.

Matching fragments into density. For each nine-residue window of amino acid sequence, we used the standard Rosetta fragment picker[13] to collect libraries of representative backbone conformations from proteins of known structure on the basis of similar sequence and predicted secondary structure. Fragments from proteins of known structure homology (PSI-BLAST *E*-value <0.05) to the benchmark proteins were excluded while constructing the fragment libraries. A sequence-derived fragment library given a protein sequence was curated with 25 backbone conformations per sequence position.

We used backbone information given a fragment to first identify the likely locations and orientations in the density map using six-dimensional (6D) translation-rotation search. The density map was subdivided into a regular 3D grid, and the search fragment was translated to each grid point in turn. At each grid point, the spherical harmonic decomposition of model and map density was used to rapidly search all rotations of a backbone fragment against regions of experimental density[18]. To further speed up matching, we carried out this rotation search only at regions of high density (mean density *Z* score >1 in a sphere around each grid point). For each fragment, the top 2,000 placements were collected using the approximated correlation score between backbone configurations and density[19], giving 50,000 candidate placements per sequence position.

Side-chain information was then used to further refine the placements and identify the most likely placements where both backbone and physically realistic side-chain conformations have good agreement to the local density. At each sequence position, the 50,000 backbone placements were then further refined with rotamer optimization and rigid-body minimization using Rosetta. After this optimization, 2,500 placements for each sequence position were selected for each sequence position using the Rosetta full-atom density correlation score[19]. These fragments were clustered (with 2-Å RMSD cluster radius), and the member with the lowest density score was taken from each cluster. Finally, if there were more than 50 clusters, only 50 fragments were carried over to model assembly.

Evaluating compatible set of fragments. From these fragment placements, we next want to select a mutually compatible set. We assessed this compatibility using a scoring function with four terms:

$$\begin{aligned}\text{score}_{\text{total}}(\boldsymbol{F}) = {} & w_{\text{dens}} \sum_{f_i \in \boldsymbol{F}} \text{score}_{\text{dens}}(f_i) \\ & + w_{\text{overlap}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{overlap}}(f_i, f_j) \\ & + w_{\text{close}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{close}}(f_i, f_j) \\ & + w_{\text{clash}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{clash}}(f_i, f_j)\end{aligned}$$

The term $\text{score}_{\text{dens}}$ measures the fit of a fragment to density and is based on the density correlation between the fragment after side-chain rotamer optimization and the experimental map[19]. The other three terms, $\text{score}_{\text{overlap}}$, $\text{score}_{\text{close}}$ and $\text{score}_{\text{clash}}$, assess the compatibility of a pair of fragments:

$$\text{score}_{\text{overlap}}(f_i, f_j) = \sum_{\substack{C\alpha_i, C\alpha_j \in f_i, f_j \\ \text{res}(C\alpha_i) = \text{res}(C\alpha_j)}} \frac{2}{1 + \exp\left(-8\left(C\alpha_i - C\alpha_j - 3\right)\right)} - 1$$

$$\text{score}_{\text{close}}(f_i, f_j) = \begin{cases} -1, \|f_i - f_j\| < \text{maxdist}(|i-j|) \\ 1, \|f_i - f_j\| \geq \text{maxdist}(|i-j|) \end{cases}$$

$$\text{score}_{\text{clash}}(f_j, f_j) = \sum_{\substack{C\alpha_i, C\alpha_j \in f_i f_j \\ |\text{res}(C\alpha_i) - \text{res}(C\alpha_j)| \geq 3}} \begin{cases} 1, \|C\alpha_i - C\alpha_j\| \leq 2.0 \\ 0, \|C\alpha_i - C\alpha_j\| > 2.0 \end{cases}$$

The term $\text{score}_{\text{overlap}}$ gives a bonus to pairs of fragments that place the same residue nearby, with a larger bonus for more overlapping residues; $\text{score}_{\text{close}}$ penalizes pairs of fragments that put residues close in the sequence further apart than maxdist, the maximum observed distance of residues at a particular sequence separation; finally, $\text{score}_{\text{clash}}$ penalizes fragment pairs with two residues occupying the same place.

Simulated annealing with Monte Carlo sampling. Monte Carlo simulated annealing (MC-SA) sampling was used to search for a set of fragments that are mutually compatible. Each sequence position is initially assigned one random (out of 50 possible) fragment placements or a 'null placement' that handles the possibility that there may be no good fragment placements at a particular sequence position. Each step in the trajectory replaces the fragment at a particular position subject to the Metropolis criterion using the $\text{score}_{\text{total}}$. For pairwise score terms, precomputing all pairwise scores allows for fast score evaluation of a fragment assignment. To control precision versus coverage, we assign a density score, $\text{dens}_{\text{null}}$, to the null placement; lower values lead to reduced coverage but more precision in fragment placement. All experiments in the paper used $\text{dens}_{\text{null}} = -150$. Finally, simulated annealing was carried out by slowly reducing the temperature from 500 to 1 in 200 increments with 5,000 moves each. Total run time was approximately 10 min per trajectory.

Iterative assembly of models. In many cases, there are a few similar fragment assignments with roughly equivalent scores. To identify all of these alternate models, we run 2,000 MC-SA trajectories. We use this ensemble to find a high-confidence partial model to carry into the next round. From the lowest scoring 5% of trajectories, we assemble a backbone model by identifying all residues that are placed in the same position (with 3-Å RMSD tolerance) and taking the average backbone coordinate at each residue position. If less than 70% of backbone residues have been assigned, we iterate fragment matching and MC-SA assembly.

The subsequent iteration of fragment matching was carried out by first masking out density that has been assigned in the backbone model from the previous iteration and then placing fragments only from sequence not yet assigned into density.

Completing models with RosettaCM. The final step in our approach is to rebuild the final set of unassigned residue positions in the partial models using RosettaCM[14], a comparative modeling method. Unassigned sequence positions in each partial model are rebuilt in the same manner as unaligned regions in comparative modeling. RosettaCM is guided by the cryo-EM density maps in completing partial models by adding a score term assessing agreement of a model to experimental density during model building and refinement with Rosetta's physically realistic all-atom energy function. For each partial model, 1,000 full-length models are generated. The best 20% by Rosetta energy are selected, and of those, the ten models with best fit to the density are selected.

In four of the cases from our test set, this led to models that had RMSDs similar to or slightly higher than the partial model from the final iteration, which is expected because the unbuilt parts are mostly loops or regions with less-resolved density. However, in two cases—FrhA and 20S-α—we saw an improvement in overall RMSD. For FrhA, this improvement was particularly striking: the Cα RMSD decreased from 2.3 Å to 1.3 Å. **Supplementary Figure 1** illustrates some improvements in the structure: RosettaCM corrected several loop residues incorrectly placed into density from the previous MC-SA assembly step. As indicated in **Supplementary Table 1**, this rebuilding is consistent and robust, with minimal structural deviation over the ten lowest-scoring models.

Model building with Buccaneer. Model building with Buccaneer[9] used the same segmented maps and was provided the same sequences as was our approach. Reflection data were computed from the cryo-EM maps using phenix.map_to_structure_factors[20]. SIGF was set to F/10 for all reflections using SFTOOLS from the CCP4 Program Suite v6.4.0 (ref. 21). A map padding of 5 Å was added to the border to ensure no effects from periodicity on model building. We ran Buccaneer from the CCP4 Program Suite v6.4.0 with mostly default setting: five cycles of building/refinement were carried out using the correlation target function during model building, with "use R-free" disabled.

20S map reconstruction. *Thermoplasma acidophilum* 20S proteasome was expressed and purified from *Escherichia coli* according to the established protocols[22]. A drop of 2 µL of purified 20S proteasome at a concentration of ~0.9 µM sample was applied to glow-discharged Quantifoil holey carbon grids (Quantifoil, Micro Tools GmbH) and plunge frozen using a Vitrobot Mark III (FEI). Grids of frozen hydrated samples were imaged using a FEI TF30 Polara electron microscope (FEI) equipped with a field mission electron source and operated at an accelerating voltage of 300 kV. Images were recorded at a nominal magnification 20,000× using a Gatan K2 Summit camera operated in super-resolution counting mode with a calibrated physical pixel size of 1.96 Å at 20,000×. A 10-s exposure time at a dose rate of ~10 counts/pixel/s leads to a total dose ~30 e^{-}/Å^2. The defocus was in the range of ~0.8–1.9 µm. The CTFFIND3 (ref. 23) was used to determine the defocus values. Half of the images with substantial drift and bad Thon rings were discarded. Side-view particles of 20S proteasome were picked automatically using FindEM[24]. All picked particles were first subject to a standard procedure of multiple rounds of multireference alignment and classification[25]. Particles within bad 2D classes were removed. All remaining particles were subject to further manual inspection, and more bad particles were removed. The final data set contained 79,801 particles from 157 images of 20,000× magnification.

GeFREALIGN[26] was used to refine and determine the 3D reconstructions with a D7 symmetry following a frequency-limited refinement procedure[27,28]. Note that no motion correction was carried out for this data set. The atomic structure of archaeal 20S proteasome (PDB code: 3C92) filtered to 15 Å was used as initial model. The final 3D reconstruction has a resolution of ~4.8 Å using Fourier shell correlation 0.143 criteria[29]. This resolution is beyond 80% of camera's physical Nyquist limit. Structure features in the amplitude-sharpened map confirm this claimed resolution.

Method availability. The *de novo* protein structure determination method described here is freely available for academic use through the Rosetta software suite (weekly releases on or after 15 February 2015), available at https://www.rosettacommons.org/.

18. DiMaio, F.P., Soni, A.B., Phillips, G.N. Jr. & Shavlik, J.W. *Int. J. Data Min. Bioinform.* **3**, 205–227 (2009).
19. DiMaio, F., Tyka, M.D., Baker, M.L., Chiu, W. & Baker, D. *J. Mol. Biol.* **392**, 181–190 (2009).
20. Adams, P.D. *et al. Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
21. Winn, M.D. *et al. Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
22. Chen, X., Wang, Q., Ni, F. & Ma, J. *Proc. Natl. Acad. Sci. USA* **107**, 11352–11357 (2010).
23. Mindell, J.A. & Grigorieff, N. *J. Struct. Biol.* **142**, 334–347 (2003).
24. Roseman, A.M. *J. Struct. Biol.* **145**, 91–99 (2004).
25. Frank, J. *et al. J. Struct. Biol.* **116**, 190–199 (1996).
26. Li, X., Grigorieff, N. & Cheng, Y. *J. Struct. Biol.* **172**, 407–412 (2010).
27. Li, X. *et al. Nat. Methods* **10**, 584–590 (2013).
28. Scheres, S.H. & Chen, S. *Nat. Methods* **9**, 853–854 (2012).
29. Rosenthal, P.B. & Henderson, R. *J. Mol. Biol.* **333**, 721–745 (2003).

doi:10.1038/nmeth.3287