Published in final edited form as:
Nat Methods. 2019 September ; 16(9): 911–917. doi:10.1038/s41592-019-0500-1.

Protein Secondary Structure Detection in Intermediate Resolution Cryo-EM Maps Using Deep Learning

Sai Raghavendra Maddhuri Venkata Subramaniya[1], **Genki Terashi**[2], **Daisuke Kihara**[*,2,1]

[1]Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA

[2]Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA

Abstract

An increasing number of protein structures have been solved by cryo-electron microscopy (cryo-EM). Although structures determined at near-atomic resolution are now routinely reported, many density maps are still determined at an intermediate resolution, where extracting structure information is still a challenge. We have developed a computational method, Emap2sec, which identifies the secondary structures of proteins (α helices, β sheets, and other structures) in an EM map of 5 to 10 Å resolution. Emap2sec uses a 3D deep convolutional neural network to assign secondary structure to each grid point in an EM map. We tested Emap2sec on 6.0 and 10.0 Å resolution EM maps simulated from 34 structures, as well as on 43 maps determined experimentally at 5.0 to 9.5 Å resolution. Emap2sec was able to clearly identify the secondary structures in many maps tested, and showed substantially better performance than existing methods.

Introduction

Cryo-electron microscopy (cryo-EM) has established its position in structural biology as an indispensable method for determining macromolecular structures due to recent technological breakthroughs[1,2]. Recent years have seen a steep increase in the number of biomolecular structures solved by cryo-EM, including those which were determined at a high, near-atomic resolution. On the other hand, there are still many structures being solved at intermediate resolutions, 5 to 10 Å, or even at lower resolutions. Of those deposited to Electron Microscopy Data Bank[3] (EMDB) in 2016 through 2018, over 50% were solved at intermediate resolution. Although the number of maps determined at near-atomic resolutions will undoubtedly increase, it is expected that a substantial number of maps will be still determined at intermediate resolutions over the coming decade, since the achievable map

[*]To whom correspondence should be addressed, dkihara@purdue.edu, Tel: 1-765-496-2284, Fax: 1-765-496-1189.

Author Contributions
D.K. conceived the study. S.R.M.V.S. designed the Emap2sec architecture with D.K. and G.T. and S.R.M.V.S. implemented it. The datasets were selected by S.R.M.V.S. and G.T. The experiments were designed by S.R.M.V.S. and D.K. and were carried out by S.R.M.V.S. S.R.M.V.S., G.T., and D.K. analyzed the results. The manuscript was drafted by S.R.M.V.S. D.K. administrated the project and wrote the manuscript.

Competing Financial interests: The authors declare no competing interests.

resolution is determined by various factors, including structural flexibility of protein chains, multiple functional states as well as noise from density images[4], implying that not all proteins may be solvable at high resolutions.

As the resolution of an EM map drops, structural interpretation of the map becomes apparently more difficult[5]. This is especially true in cases of de novo modeling, which is employed when the structure of the proteins have not been solved before or cannot be built by template-based modeling[6–10]. When a map is at a high resolution of around 2 Å, software originally designed for X-ray crystallography[11,12] can be used to build an atomic-resolution structure model. At around 4 Å resolution, a main-chain tracing method[13,14], such as MAINMAST[15,16], can be applied for modeling. For maps around 5 to 8 Å, some fragments of the secondary structure of proteins are typically visible, but a full trace of the main-chain is very difficult. To aid structural interpretation of maps in this resolution range, several protein secondary structure detection methods have been developed. One type of such methods identifies density regions that are typical of helices[17,18] or β sheets[19–21], and then builds protein models[22] or finds known structures in PDB[23] that match the identified secondary structures[24]. Another approach used machine learning methods to detect characteristic density patterns of secondary structures[25].

Here, we developed a new method, Emap2sec, for detecting protein secondary structures in EM maps of intermediate resolution. Emap2sec uses a deep convolutional neural network (CNN)[26,27] at the core of its algorithm. CNN is very suitable for protein local structure identification in 3D EM maps because the method “convolves” local map density features to images of a larger region so that the local structure detection is made in the context of a large region of the map. The performance of Emap2sec was tested on two datasets of EM maps: a dataset of 6.0 and 10.0 Å resolution EM maps simulated from each of 34 structures and a dataset of 43 experimental EM maps with resolution ranging from 5.0 to 9.5 Å. The overall accuracy at each amino acid level was 83.1 % and 79.8 % on average for the simulated maps at 6.0 Å and 10.0 Å, respectively. On the experimental map dataset, the accuracy was 64.4 % on average with the highest recording 91.6 %.

Results

The network architecture of Emap2sec

Emap2sec scans a given EM map with a voxel of 11^3 Å^3 volume. The voxel reads the density values, which are processed through the deep CNN, and outputs a detected secondary structure, α helix, β sheet, or other structures for the structure at the center of the voxel (Fig. 1a). We adopt a two-phase stacked neural network architecture where densities from local protein structures captured in the first phase are fine-tuned in the subsequent second phase by incorporating the context of neighboring voxels.

Fig. 1b and 1c illustrate the two-phase architecture of Emap2sec. The first phase of the network (Fig. 1b) contains five back-to-back convolutional layers, to which the input voxels of 11^3 Å^3 (i.e. 1331 density values) are fed. The convolutional layers capture local density patterns in an EM map as represented by pattern filters in the context of larger regions of the map. 32, 64, or 128 filters were used at each convolutional layer. Then, the output from the convolutional layers is processed through a max pooling layer and fully connected layers. Finally, a softmax layer outputs a probability value for each of α helix, β strand, or other structures. A stride of 1 was used uniformly in all the filters whereas a stride of 2 was used for the max pooling layer.

The second phase of the network (Fig. 1c) takes the predicted probability values from the first phase. An input of the second phase is the probability values for α helix, β strand, and other structures each in a voxel of 3^3 $Å^3$ size. Thus, there are $3^3 * 3 = 81$ values. This input is fed to a network of five fully connected layers followed by a softmax layer, which finally gives a prediction of the secondary structure with a probability value for the center of target voxel. The purpose of the second phase network is to smooth predictions by considering predictions made to the neighboring voxels. For experimental maps, we set the phase 2 network to only change between α helix and β strand or from other structures to α helix/β strand, because there are more other structures in the experimental maps than in simulated maps. The two-phase network architecture is inspired by a sequence-based protein secondary structure prediction, PSIPRED[28].

Secondary structure detection on simulated maps

We first evaluated the performance of Emap2sec on simulated EM maps computed from atomic-detailed protein structures. Parameters of the network were trained on a representative protein structure dataset obtained from the SCOPe protein domain structure database[29]. We tested Emap2sec on 34 simulated maps at two resolutions, 6.0 Å and 10.0 Å, which do not have overlap with data used for training (see Methods).

Table 1 summarizes the accuracy of the secondary structure identification for three levels, the voxel-, residue-, and segment-levels. For each voxel to which Emap2sec makes individual structure identification, the correct secondary structure is defined as the one from the closest Cα atom within 3.0 Å to the center of the voxel. Emap2sec made fairly accurate structure detections overall. For the 6.0 Å maps, Emap2sec achieved an average overall accuracy of 0.798. Among the three secondary structure classes, α helices were detected with the highest accuracy (0.848), followed by β strands with an accuracy of 0.828. At the residue-level (Q3), accuracy higher than the voxel-level was recorded. At the segment-level, the accuracy values reached even higher, 0.872 overall. Since the residue- and the segment-level assignments were made by the majority vote from the one-step more detailed-level (i.e. the voxel- and the residue-level, respectively), the results indicate that our method was very successful in capturing overall main-chain level protein structures in the maps.

We now compare the accuracy between maps simulated at 6.0 Å and at 10.0 Å (Table 1, Fig. 2a). Looking at the Q3 accuracy in Table 1, naturally the accuracy dropped for 10.0 Å maps; however, the margin of the drop was surprisingly small. The average Q3 values of 6.0 Å and 10.0 Å maps were 0.831 and 0.798, respectively, thus the decrease was only 0.033 (3.97% relative to the 6.0 Å maps). This result is consistent with Fig. 2a, where the Q3 values of individual maps are shown. Among the three secondary structure types, the smallest decrease of 0.023 (2.66%) was observed for α helices, indicating helices are the most tolerant to noise and visible in lower resolution maps.

In Fig. 2b, we show Q3 accuracy for individual fold classes for both 6.0 Å and 10.0 Å maps. Emap2sec performed best for the α class proteins while the accuracy for β class proteins was relatively low. We noticed that proteins in the α+β class had a higher accuracy than α/β class. This is partly because proteins in the α/β class had more residues in β strands than α +β proteins in this dataset. The average number of residues in β strands was 47.0 and 40.2 for the α/β and α+β class, respectively. Also, α+β proteins have helices and strands in spatially distinct regions by definition, which probably made identification easier for Emap2sec. The Q3 accuracy of α helices was on par, 0.854 and 0.848 for α/β and α+β while the β strand accuracy showed some gap, 0.788 and 0.840 for α/β and α+β, respectively.

The last panel in Fig. 2 illustrates the effect of the two-phase network architecture of Emap2sec (Fig. 2c). The results for maps at the two resolutions have the same trend: the phase 2 network improved the accuracy for α helices (red circles) and β strands (green downward triangles), and so did the overall accuracy (black circles). On average, the Q3 accuracies of overall, α helices, and β strands were improved by the phase 2 network by 0.011/0.011, 0.093/0.089, and 0.062/0.021, for 6.0/10.0 Å maps, respectively. Instead, the accuracy of other structures (yellow upward triangles) decreased for many maps, which indicates that the phase 2 network mainly changed assignments of other structures to either helices or strands. Supplementary Table 1 provides all accuracy values from the phase 1 and phase 2 networks for all the simulated maps.

The results shown so far are computed using the networks trained on 63 EM maps. We have also trained the networks on a larger dataset of 1963 maps (Supplementary Figure 1 and Supplementary Table 2) but only observed marginal improvement. Changes in the accuracies of individual maps is shown in Supplementary Figure 1.

Figure 3 shows examples of the structure identification by Emap2sec. The first four panels (a-d) are successful examples, one each for four fold classes, α, β, α/β, and α+β, respectively. Fig. 3a and 3c are maps simulated at 6.0 Å while 3b and 3d are at 10.0 Å. As visualized, the identified secondary structures shown in colored spheres agree with the true structure very well. In the structure of Fig. 3b, even small helices were correctly identified. In the α/β and α+β structures (Fig. 3c, 3d), helices and strands are clearly distinguished.

On the other hand, the two subsequent panels (Fig. 3e, 3f) show examples where a part of the identification was apparently wrong. For these two cases, extended loop regions that have a somewhat wriggled conformation were misidentified as α helices. Fig. 3g illustrates how the phase 2 network modifies the output from phase 1. The phase 1 output has more other structure assignments (green) even in helix regions, which were correctly changed to helix assignments in phase 2. Due to this modification by phase 2, the overall voxel-based accuracy improved from 0.705 to 0.810. The last panel (Fig. 3h) shows an example of results for 6.0 Å and 10.0 Å maps of the same protein. The overall F1-score for the 6.0 Å map was 0.848, which deteriorated to 0.779 for the 10.0 Å map. Noticeable differences were circled on the figure of the 10.0 Å map results: some parts of β sheets were not correctly recognized, and some over-detections of α helices were observed.

Structure detection in experimental maps

Next, we tested Emap2sec on a dataset of 43 experimental maps with resolutions ranging from 5.0 Å to 9.5 Å. The number of residues of the proteins in these maps ranges from 664 to 11,475 and the number of chains ranges from 1 to 62. Fig. 4a and 4b show the residue-based and the segment-based accuracy of α helices and β strands relative to the map resolution. At the residue-level (Fig. 4a), the accuracy did not show a strong dependency on the map resolution. High accuracies of over 0.8 (0.808 to 0.916) were observed for all the maps within the resolution range of 5.0 Å to 5.7 Å. On the other hand, it is noteworthy that high accuracy was also observed for maps at around 9.0 Å resolution, 0.802 (EMD-1655, determined at a map resolution of 9.0 Å) and 0.757 (EMD-1263, map resolution: 9.1 Å).

Compared with the simulated map cases, detecting structures in experimental maps is more difficult. For the simulated map dataset, the average residue-based accuracy was 0.831 and 0.798 for the resolution of 6.0 Å and 10.0 Å (Table 1), respectively. On the other hand, the average accuracy for experimental maps of 5.5 to 6.5 Å and those over 9.0 Å was 0.723 and 0.563, respectively, which is a decrease by 13 % and 29% from the simulated map counterpart.

Fig. 4b shows the accuracy change by the phase 2 network over the phase 1. Similar to results for the simulated map dataset (Fig. 2c), the accuracy for α helices and β strands were improved consistently for almost all the maps by phase 2. With the phase 2 network, the residue-based accuracy of α helices and β strands improved on average from 0.454 and 0.401 to 0.565 (24.4%) and 0.473 (17.80%), respectively.

The results shown for the experimental maps were obtained by using Emap2sec trained on experimental maps. Adding simulated maps in the training set to increase the training data size did not make consistent improvement of detection results. Adding the simulated maps improved the voxel-based F1-score for 14 maps (32.6%) but deteriorated for 22 maps (51.2%) (no change for 7 maps). As a consequence the overall average voxel-based F1-score slightly deteriorated (Supplementary Figure 2). Thus, probably because of different nature of two types of maps, adding simulated maps in training did not help much for structure detection in experimental maps.

We discuss several illustrative examples in Fig. 5. The first four panels (5a – 5d) are cases where Emap2sec performed well. Fig. 5a and 5b are α helix-rich and β strand-rich complex structures, respectively. The dominance of α helices in Fig. 5a and β strands in Fig. 5b was vividly captured, which yielded high accuracy values (see the figure caption). In addition, β sheets in Fig. 5a and α helices in fig. 5b, which share a small portion in the maps, were accurately detected. The next example (Fig. 5c), archaeal 20S proteasome, is a more difficult case where α helices and β strands exist in almost the same amounts in the structure, 38.4% and 29.9% of all the residues, respectively. Despite the more complex structure, Emap2sec was able to detect the secondary structures distinctively as highlighted in the zoomed window and successfully captured the overall architecture of the structure that has layers of α helices and β sheet domains. Fig. 5d is another example of structures with a similar amount of α helices (39.8%) and β strands (22.7%) except that the map resolution is lower, at 8.34 Å. Although the overall accuracy was lower than the previous example (Fig. 5c), β strand-rich domain on the left in the figure and α helix-rich domain and structural details were well captured, including locations of β sheets in the α helix-rich domain and a single helix that bridges the two large domains (the zoomed window). The next one is an example where the detection did not work very well (Fig. 5e). For this map determined at 9.1 Å, the overall Q3 accuracy was relatively low, 0.565, partly because many α helices on the left side of the structure were misrecognized as β strands. It turned out that it was because, as shown in the zoomed window, many such helices have lower density than β stands, often even almost outside the contour level, which is opposite to the usual case where α helices have higher density than β strands.

In Fig. 5f applied Emap2sec to an EM map determined at 7.94 Å that has no structural assignment (EMD-4361, mLRRC8A/C volume-gated anion channel[30]). Although this map does not have associated PDB files in EMDB, it turned out that the authors also deposited two additional higher resolution EM maps (EMD-4366 determined at 5.01 Å and EMD-4367 at 4.25 Å) that have associated PDB entries. When compared with those PDB structures, Emap2sec's detection was quite correct, including β propeller structures on the top (two right panels) and β sheets at the bottom (the LRRC8A subunits) in the side-view (two left panels. α helices surrounding the β sheets are outside the author-recommended contour level hence not analyzed by Emap2sec). Thus, Emap2sec was able to perform structural assignment even with the map at 7.94 Å resolution.

Detected secondary structure information will be useful for assigning subunit structures in an EM map. Supplementary Figure 3 shows two such examples. In the first map, HIV capsid proteins in complex with antibodies (EMD-8693), the detected secondary structures would help assigning β-sandwich structures of Fab proteins as well as three other α-helical chains that are distinctively and accurately detected in the map. Similarly, in the second example, a map of Tor2 in complex with Lst8 (EMD-3229), the locations of two chains of Lst8, which has a round-shaped β-class structure, are clearly visible in the detected structure.

Comparison with related works

Emap2sec showed substantially better performance than two widely used protein structure modeling methods, Phenix[12] and ARP/wARP (ver. 8.0)[31] on the dataset of simulated maps at 6.0 Å and 10.0 Å (Supplementary Table 4 and Supplementary Figure 4).

Emap2sec was also compared with two similar methods, HelixHunter[17], which uses a probe helix to identify helices in a map, and a more recent method by the Jing He group[25], which uses CNN. Overall Emap2sec showed better performance than these two existing methods on the simulated maps used in their original papers (Supplementary Table 5 and 6).

Discussion

We developed Emap2sec, a novel approach that can detect protein secondary structures in an EM density map of intermediate resolution. This work shows that structure information of proteins can be obtained from an intermediate resolution EM maps more vividly and accurately than conventional methods partly owing to a recent advanced image processing algorithm. Expanding the approach to handle other molecules, such as DNAs, RNAs, lipids as well as large posttranslational modifications, is left for future work. Emap2sec will be able to push the limit of extracting structure information and aid for structural modeling and will be an indispensable tool for interpreting density maps in the era of cryo-EM structural biology.

Methods

Training the deep neural network of Emap2sec

We used two datasets, simulated EM maps and experimental EM maps for training and testing Emap2sec. We explain the training procedure on the simulated EM map dataset as the process for the experimental EM map dataset is essentially the same.

For the dataset of simulated maps, we downloaded one representative structure each from a superfamily level classification in the SCOPe database[29] (ver. 2.06), then structures were removed if they have over 25% sequence identity with another structure in the dataset. This remained 2000 structures. Next, we used the e2pdb2mrc program from EMAN2 package[32] (version 2.11) to generate simulated EM maps at 6.0 Å and 10.0 Å for each structure. Density values were normalized in each map to the range of 0.0 to 1.0 by subtracting the minimum density value in the map and then divide by the density difference between the maximum and the minimum density values. If there are negative density values, they were set to 0.0 before the normalization.

The input density data of Emap2sec is a voxel of a size 11 Å *11 Å *11 Å. We have also tried a smaller size, 5 Å *5 Å *5 Å, but it performed substantially worse than the current setting. The data were obtained from a map by traversing each map along the three dimensions with the voxels with a stride of 2.0 Å. Each voxel was assigned with a closest Cα atom that is within 3.0 Å to the center of the voxels, to which a secondary structure type was assigned using STRIDE[33]. Residues that have structure codes of H, G, or I by STRIDE were labeled as α helix, while those with codes of B/b or E were labeled as β strand.

The training dataset for the phase 1 network consisted of 31,951 voxels each for α helix, β strand, and other structures, thus in total of 95,853 voxels. To select these voxels, the 2000 maps were sorted by the abundance of voxels of each secondary structure, and voxels of the secondary structure were extracted from the top 8, 15, and 8 maps with the largest number of voxels of the secondary structure. Since the number of voxels was sufficient and helices and strands appear in various orientations in the voxels, we did not perform data augmentation by rotating maps. This training dataset was constructed at both 6.0 Å and 10.0 Å resolutions. The training was performed independently for 6.0 Å and 10.0 Å resolutions.

The phase 1 network consists of five layers of convolutional neural network (CNN) followed by a fully connected network (Fig. 1). With this training dataset, we performed a ten-fold cross-validation to determine hyper-parameters, which were regularization parameters for the CNN and the fully connected network, and the learning rate. L_2 and L_1 regularization were used for the CNN and the fully connected network, respectively. The regularization parameter values tested were [0.001, 0.005, 0.01, 0.05. 0.1, 0.5, 1, 10, 100] and the learning rate values tested were [0.0001, 0.001, 0.01, 0.1, 1, 10]. For the first fully connected layer (the connection between the layer with 1024 nodes and 256 nodes) we used the drop-out technique with a probability of 0.5. In the cross-validation, the dataset was split into ten subsets, nine among which were used for training under all the possible hyper-parameter combinations for 100–150 epochs using the Adam Optimizer to minimize the cross-entropy for the voxel-level accuracy in the softmax layer. Then, the best parameter combination was determined by the average performance on the ten testing subsets. The determined L_2 regularization parameter λ_1 of the CNN was 10, the L_1 regularization parameter λ_2 of the fully connected layer was 0.01, and the learning rate was determined to 0.001 by this procedure for both 6.0 Å and 10.0 Å resolutions.

For training the phase 2 network, which is a five-layer fully connected network, we used a different dataset of 32 simulated EM maps (each for 6.0 Å and 10.0 Å). After the first phase network was fully trained, we input the 32 maps to the phase 1 network and obtained probability values for α helix, β strand, and other structures, which are input for the phase 2 network. We trained the phase 2 network for the L_1 regularization parameter in the same way with a fixed learning rate of 0.001. The obtained parameter was 0.1 for both 6.0 Å and 10.0 Å resolutions.

We have also trained the phase 1 and 2 networks using all the maps except for the 34 test maps. They were 982 maps for training the phase 1 and other 981 maps for the phase 2 network. 3 maps were discarded because they caused errors when voxel data were generated. The results are shown in Supplementary Figure 1 and Supplementary Table 2.

The dataset of experimental EM maps

We also trained Emap2sec using actual EM maps retrieved from EMDB[3]. Density maps that were determined at a resolution between 5.0 Å to 10.0 Å and have an associated atomic structure in PDB were selected. Then, to ensure that a map and its associated structure have sufficient structural agreement, the cross-correlation between the experimental map and the simulated map density at the resolution of the experimental map computed from the structure was examined[34] and only maps with a cross-correlation of over 0.65 were kept. Finally, we computed the sequence identity between underlined proteins in pairs of EM maps, and a map was removed from the dataset if its underlined protein has over 25% identity to a protein of another map in the dataset. If a map has multiple protein chains, the map was removed if at least one of the chains has over 25% identity to any chains in another map. This procedure remained 43 experimental EM maps. The grid size of the maps was unified to 1.0 Å by applying tri-linear interpolation of the electron density in the maps. Secondary structure detection by Emap2sec was evaluated only for voxels that have associated protein structures (and thus correct secondary structure of that position was known).

The training and testing were performed using experimental maps of 6.0 to 10.0 Å resolution using the four-fold cross-validation. The same hyper-parameter values established for the simulated maps were used for experimental maps. The trained model was further applied to experimental maps of 5.0 to 6.0 Å resolution.

Evaluation measures

Secondary structure detection by Emap2sec was evaluated in three levels, i.e. the voxel, the amino acid residue, and the secondary structure fragment levels. Emap2sec assigns a secondary structure to each voxel in an EM map. For each voxel in the map, the “correct” secondary structure defined by STRIDE was assigned, which was taken from the closest atom that is within 3.0 Å to the center of the voxel, as discussed in the previous section. In Table 1, we also showed results with a relaxed measure, where a voxel is assigned with the secondary structure of all Cα atoms within 3.0 Å, thus potentially have multiple correct secondary structures. For the voxel-level evaluation, we used the F1-score, which is the harmonic mean of the precision and the recall of the assignments given to the entire map or to each secondary structure class:

$$F1 - score = 2 * \frac{precision \times recall}{precision + recall}, \quad \text{(Eq. 1)}$$

where precision is the fraction of voxels with correct secondary structure detection over all the voxels with the secondary structure assignment by Emap2sec and recall is the fraction of the voxels with correct secondary structure detection over all the voxels that belong to the secondary structure class. The overall F1-score was computed as the weighted F1-score of the three secondary structure classes. For the residue-level evaluation, we reported the Q3 accuracy, which is the fraction of the number of residues with a correct secondary structure assignment for the entire protein and for each of the three secondary structure classes. The secondary structure of a Cα atom was considered as correctly predicted if a majority of neighboring voxels that are within 3.0 Å to the atom agreed to its secondary structure. We also reported the segment-level accuracy. A secondary structure segment is defined as a stretch of amino acids with the same secondary structure type, at least six amino acids for α helix and three residues for a β strand. A segment was considered as correctly detected if at least 50% of the voxels in that segment have the correctly assigned class label.

Software used

Software used in this work with the version and the download site information is provided in the Life Sciences Reporting Summary.

Data Availability

The raw data of accuracies are provided in Supplementary Information, Supplementary Table 1, 3, and 4. The experimental EM maps can be downloaded from EMDB. The data that support the findings of this study are available from the corresponding author upon request.

Code Availability

The Emap2sec program is freely available for academic use through http://kiharalab.org/emap2sec/index.html and https://github.com/kiharalab/Emap2sec

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge Charles Christoffer for his help in finalizing the manuscript. This work was partly supported by the National Institutes of Health (R01GM123055), the National Science Foundation (DMS1614777, CMMI1825941), and the Purdue Institute of Drug Discovery.

References

1. Kuhlbrandt W Cryo-EM enters a new era. Elife 3, e03678, (2014). [PubMed: 25122623]
2. Cheng Y Single-particle cryo-EM-How did it get here and where will it go. Science 361, 876–880, (2018). [PubMed: 30166484]
3. Patwardhan A Trends in the Electron Microscopy Data Bank (EMDB). Acta Crystallogr D Struct Biol 73, 503–508, (2017). [PubMed: 28580912]
4. Nogales E The development of cryo-EM into a mainstream structural biology technique. Nat Methods 13, 24–27, (2016). [PubMed: 27110629]
5. Esquivel-Rodriguez J & Kihara D Computational methods for constructing protein structure models from 3D electron microscopy maps. J Struct Biol 184, 93–102, (2013). [PubMed: 23796504]
6. Kirmizialtin S, Loerke J, Behrmann E, Spahn CM & Sanbonmatsu KY Using Molecular Simulation to Model High-Resolution Cryo-EM Reconstructions. Methods Enzymol 558, 497–514, (2015). [PubMed: 26068751]
7. Miyashita O, Kobayashi C, Mori T, Sugita Y & Tama F Flexible fitting to cryo-EM density map using ensemble molecular dynamics simulations. J Comput Chem 38, 1447–1461, (2017). [PubMed: 28370077]
8. Esquivel-Rodriguez J & Kihara D Fitting Multimeric Protein Complexes into Electron Microscopy Maps Using 3D Zernike Descriptors. J Phys Chem B 116, 6854–6861, (2012). [PubMed: 22417139]
9. Saha M & Morais MC FOLD-EM: automated fold recognition in medium- and low-resolution (4–15 A) electron density maps. Bioinformatics 28, 3265–3273, (2012). [PubMed: 23131460]
10. Zheng W Accurate flexible fitting of high-resolution protein structures into cryo-electron microscopy maps using coarse-grained pseudo-energy minimization. Biophys J 100, 478–488, (2011). [PubMed: 21244844]
11. Brown A et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. Acta Crystallogr D Biol Crystallogr 71, 136–153, (2015). [PubMed: 25615868]
12. Terwilliger TC et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr 64, 61–69, (2008). [PubMed: 18094468]
13. DiMaio F et al. Atomic-accuracy models from 4.5-A cryo-electron microscopy data with density-guided iterative local refinement. Nat Methods 12, 361–365, (2015). [PubMed: 25707030]
14. Chen M, Baldwin PR, Ludtke SJ & Baker ML De Novo modeling in cryo-EM density maps with Pathwalking. J Struct Biol 196, 289–298, (2016). [PubMed: 27436409]
15. Terashi G & Kihara D De novo main-chain modeling for EM maps using MAINMAST. Nat Commun 9, 1618, (2018). [PubMed: 29691408]
16. Terashi G & Kihara D De novo main-chain modeling with MAINMAST in 2015/2016 EM Model Challenge. J Struct Biol 204, 351–359, (2018). [PubMed: 30075190]
17. Jiang W, Baker ML, Ludtke SJ & Chiu W Bridging the information gap: computational tools for intermediate resolution structure interpretation. J Mol Biol 308, 1033–1044, (2001). [PubMed: 11352589]

18. Dou H, Burrows DW, Baker ML & Ju T Flexible Fitting of Atomic Models into Cryo-EM Density Maps Guided by Helix Correspondences. Biophys J 112, 2479–2493, (2017). [PubMed: 28636906]
19. Kong Y, Zhang X, Baker TS & Ma J A Structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps. J Mol Biol 339, 117–130, (2004). [PubMed: 15123425]
20. Si D & He J Modeling Beta-Traces for Beta-Barrels from Cryo-EM Density Maps. Biomed Res Int 2017, 1793213, (2017). [PubMed: 28164115]
21. Si D & He J Tracing beta strands using StrandTwister from cryo-EM density maps at medium resolutions. Structure 22, 1665–1676, (2014). [PubMed: 25308866]
22. Lindert S et al. EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. Structure 20, 464–478, (2012). [PubMed: 22405005]
23. Berman HM et al. The Protein Data Bank. Nucleic Acids Res 28, 235–242, (2000). [PubMed: 10592235]
24. Biswas A et al. An Effective Computational Method Incorporating Multiple Secondary Structure Predictions in Topology Determination for Cryo-EM Images. IEEE/ACM Trans Comput Biol Bioinform 14, 578–586, (2017). [PubMed: 27008671]
25. Li RJ, Si D, Zeng T, Ji SW & He J Deep Convolutional Neural Networks for Detecting Secondary Structures in Protein Density Maps from Cryo-Electron Microscopy. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 41–46, (2016).
26. Russakovsky O et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115, 1–42, (2015).
27. Maturana D & Scherer S VoxNet: A 3D convolutional neural network for real-time object recognition. IEEE/RSJ International Conference on Intelligent Robots and Systems, 922–928, (2015).
28. Jones DT Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195, (1999). [PubMed: 10493868]
29. Fox NK, Brenner SE & Chandonia JM SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42, D304–309, (2014). [PubMed: 24304899]
30. Deneka D, Sawicka M, Lam AKM, Paulino C & Dutzler R Structure of a volume-regulated anion channel of the LRRC8 family. Nature 558, 254–259, (2018). [PubMed: 29769723]
31. Langer G, Cohen SX, Lamzin VS & Perrakis A Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat Protoc 3, 1171–1179, (2008). [PubMed: 18600222]
32. Tang G et al. EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol 157, 38–46, (2007). [PubMed: 16859925]
33. Frishman D & Argos P Knowledge-based protein secondary structure assignment. Proteins 23, 566–579, (1995). [PubMed: 8749853]
34. Monroe L, Terashi G & Kihara D Variability of Protein Structure Models from Electron Microscopy. Structure 25, 592–602 e592, (2017). [PubMed: 28262392]

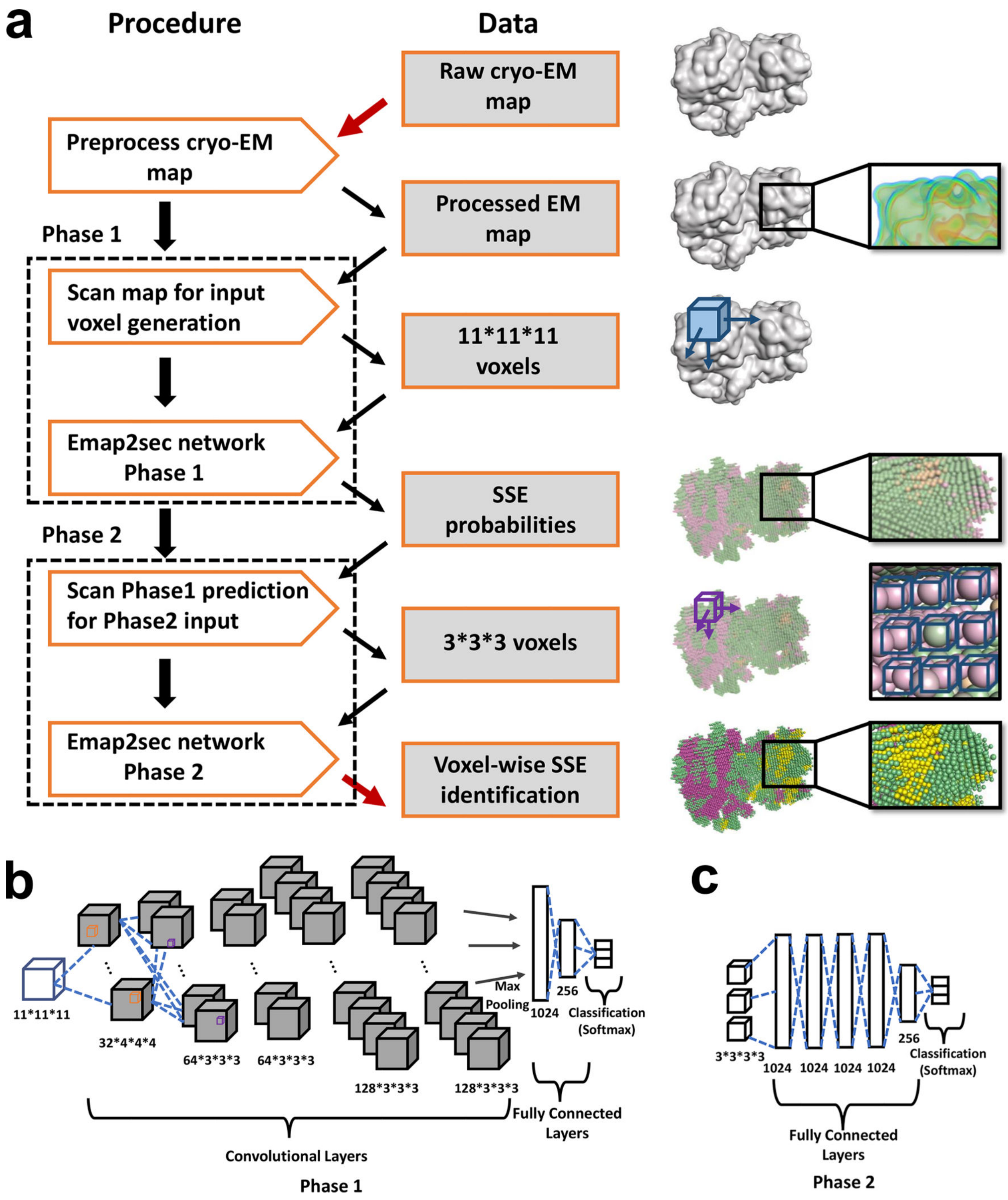

Figure 1.

The architecture of Emap2sec. **a**, the flowchart. Emap2sec takes a 3D EM density map as input, and scans it with a voxel of a size of 11^3Å^3. There are two phases in Emap2sec. The phase 1 network takes the normalized density values of a voxel and outputs the probability values of the three secondary structure classes. The phase 2 network takes the output from the phase 1 network and refines the assignment by considering assignments made to neighboring voxels. The input is a voxel of 3^3 Å^3 where each prediction (shown in a sphere) originates from a voxel of 11^3Å^3 in the phase 1 network. Finally, each voxel is assigned with a secondary structure class, which is the one with the largest probability among the three structure types. **b**, the architecture of the phase 1 deep neural networks. It has five CNN layers followed by one max pooling layer. The first CNN has 32 filters of a 4^3 Å^3 size, the second and the third ones are with 64 filters of a 3^3 Å^3 size, and the fourth and the fifth ones with 128 filters of a 3^3 Å^3 size. The last layers of the network are two fully connected (FC) layers, which have 1024 and 256 nodes each. The FC layers are connected to the output layer, which uses the softmax function to compute the probabilities for the three secondary structure classes. **c**, the phase 2 network. It consists of five FC layers followed by an output layer.

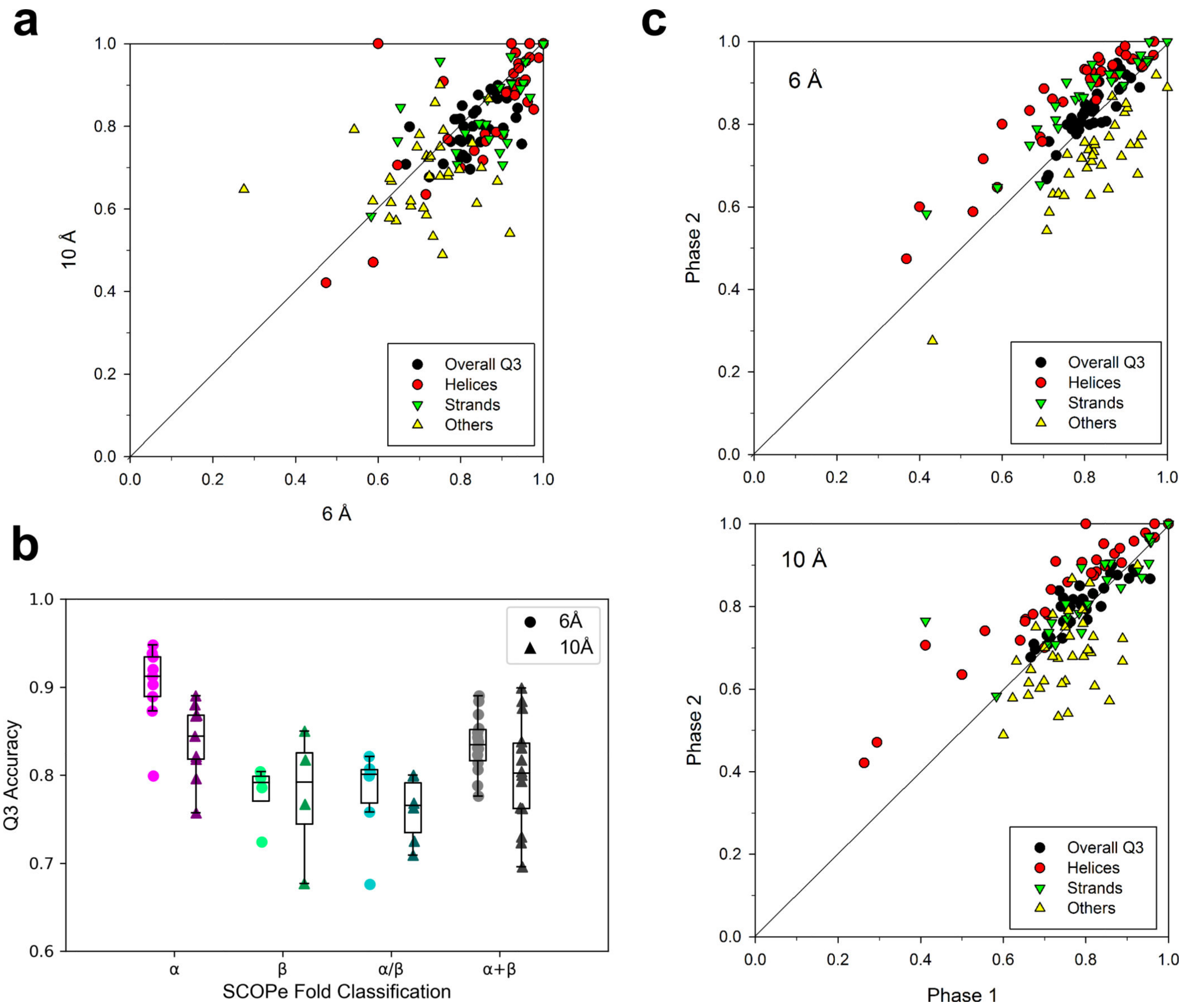

Figure 2.

The secondary structure assignment performance on the simulated map dataset. The dataset consists of maps of 34 protein structures computed at 6.0 Å and at 10.0 Å. **a**, Q3 residue-based accuracy by map resolution. **b**, the Q3 accuracy for four-fold classes defined in the SCOPe database shown in box plots. The number of structures in each class is α, 9; β, 4; α/β, 6; and α+β, 14. One protein from the small protein class in the dataset is not included in this plot. The box plots show the median (M), the first (1Q) and the third quartile (3Q), and the minimum (Mi) and the maximum (Mx) values. For α class, 6.0 Å: M: 0.912, 1Q: 0.889, 3Q: 0.934, Mi: 0.873, Mx: 0.948. 10.0 Å: M: 0.844, 1Q: 0.818, 3Q: 0.864, Mi: 0.757, Mx: 0.890. For β class, 6.0 Å: M: 0.792, 1Q: 0.771, 3Q: 0.799, Mi: 0.771, Mx: 0.804. 10.0 Å: M: 0.792, 1Q: 0.745, 3Q: 0.825, Mi: 0.677, Mx: 0.850. For α/β class, 6.0 Å: M: 0.801, 1Q: 0.768, 3Q: 0.806, Mi: 0.758, Mx: 0.821. 10.0 Å: M: 0.766, 1Q: 0.735, 3Q: 0.791, Mi: 0.709, Mx: 0.800. For α+β class, M: 0.835, 1Q: 0.816, 3Q: 0.852, Mi: 0.776, Mx: 0.890. 10.0 Å: M: 0.802, 1Q: 0.762, 3Q: 0.836, Mi: 0.696, Mx: 0.899. **c**, The Q3 accuracy before after applying the phase 2 network. The results for the 6.0 Å maps are shown on the top, and the bottom panel is for the 10.0 Å maps. Raw data of all the maps are provided in Supplementary Table 1.

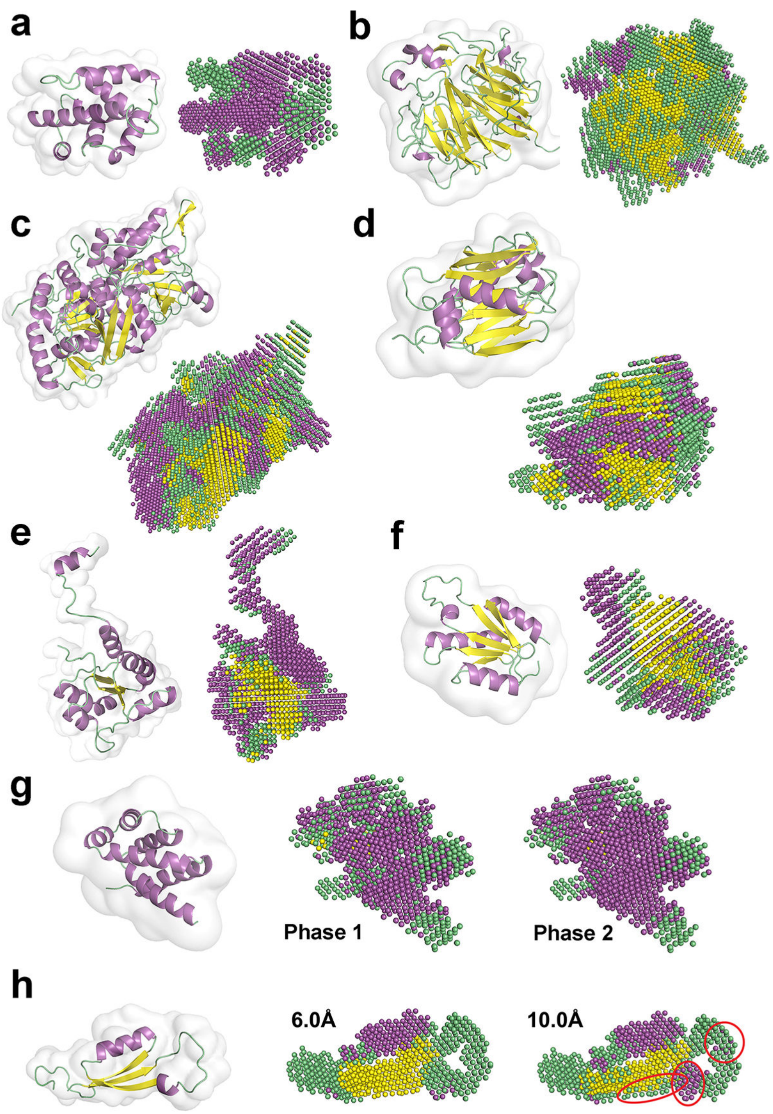

Figure 3.

Examples of the secondary structure assignment for simulated maps at 6.0 and 10.0 Å. For each panel, the main-chain structure of the protein in the simulated EM map is shown on the left while the right figure with small spheres shows the final structure assignments from the phase 2 network. Spheres in magenta are assignments of α helices; yellow, β strands; and green, other structures. See Supplementary Table 1 for all accuracy values of the examples. **a**, an α class protein of 114 residue-long (SCOPe code: d1azta1). The map was simulated at 6.0 Å. The overall F1 score, voxel-based accuracy (Acc), Q3 accuracy, and the segment-based accuracy of α helices and β strands (Seg_{ab}) were 0.908, 0.908, 0.948, and 1.0, respectively. **b**, a β class protein (d1a12a_), 401 residues. Resolution: 10.0 Å. F1: 0.642; Acc: 0.641; Q3: 0.677; Seg_{ab}: 0.862. The segment-based accuracy for β strands was 0.862. **c**, an α/β class protein (d1acoa2), 527 residues. Resolution: 6.0 Å. F1: 0.750; Acc: 0.748; Q3: 0.799; Seg_{ab}: 0.865. **d**, an α+β class protein (d1b25a2), 210 residues. Resolution: 10.0 Å. F1: 0.661; Acc: 0.665; Q3: 0.730; Seg_{ab}: 0.941. **e**, a structure with a region of a low accuracy for other structures. A 138 residue-long α/β class protein (d1a9×a2). Resolution:

6.0 Å. F1: 0.633; Acc: 0.670; Q3: 0.676; Seg_{ab}: 1.0. Accuracies for other structures: F1: 0.396; Acc: 0.283; Q3, 0.275. **f**, another structure with a low accuracy for other structures. An α/β class protein (d1atia1), 111 residues. Resolution: 10.0 Å. F1: 0.698; Acc: 0.711; Q3: 0.768; Seg_{ab}: 1.0. Accuracies for other structures: F1: 0.581; Acc: 0.48; Q3, 0.533. **g**, an example of assignment changes from the phase 1 to phase 2 network. An α class protein (d1abva_), 105 residues. Resolution: 10.0 Å. Phase 1, F1: 0.752 (overall), 0.807 (α), and 0.469 (other structures). Phase 2, F1: 0.834 (overall), 0.889 (α), and 0.546 (other structures). **h**, an example of structure detection for 6.0 Å and 10.0 Å maps. db133n_, 67 residues. The map on the left is simulated at 6.0 Å. F1 scores of 6.0 Å /10.0 Å maps: 0.848/0.779 (overall), 0.788/0.756 (α), 0.893/0.820 (β), 0.851/0.764 (other structures).

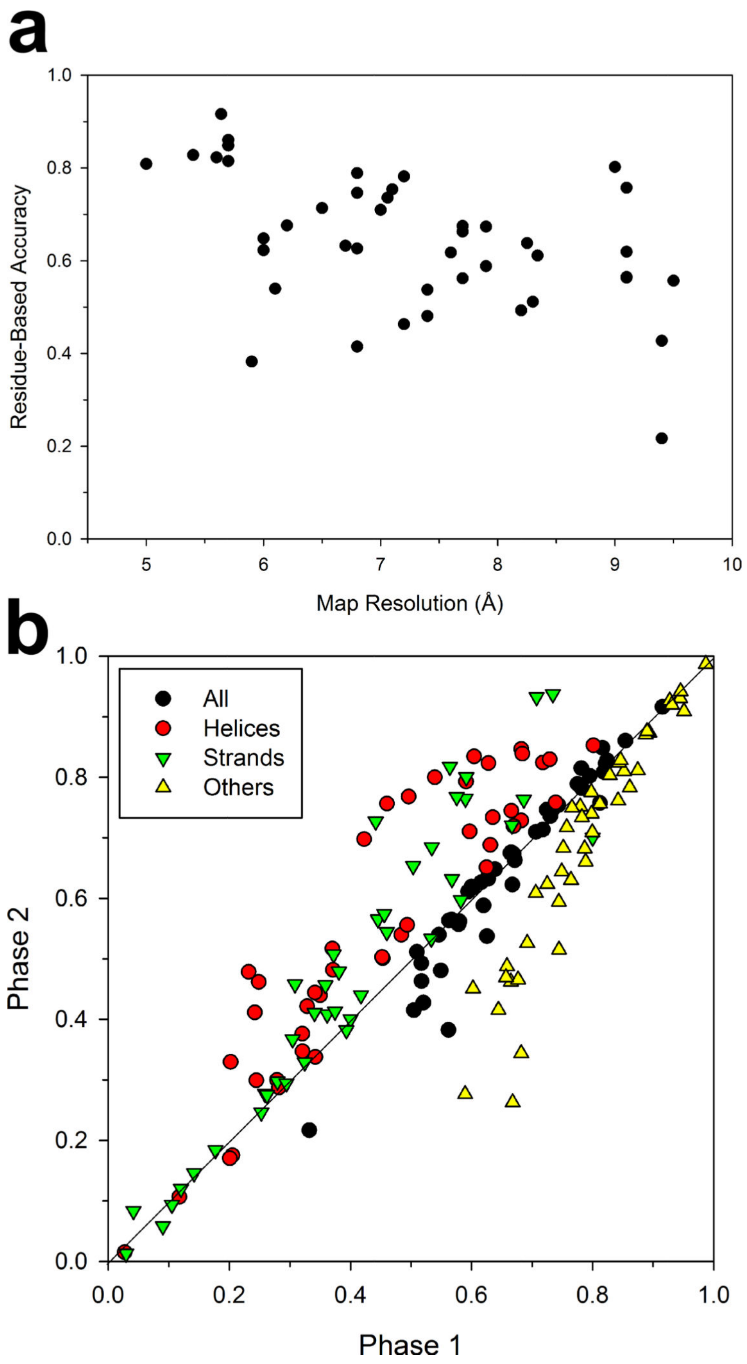

Figure 4.

The secondary structure detection accuracy on the 43 experimental maps. See Supplementary Table 3 for details of the accuracy of each map. **a**, the Q3 accuracy relative to the map resolution. **b**, Q3 accuracy before after applying the phase 2 network. Supplementary Table 3 provides raw data of all the maps.

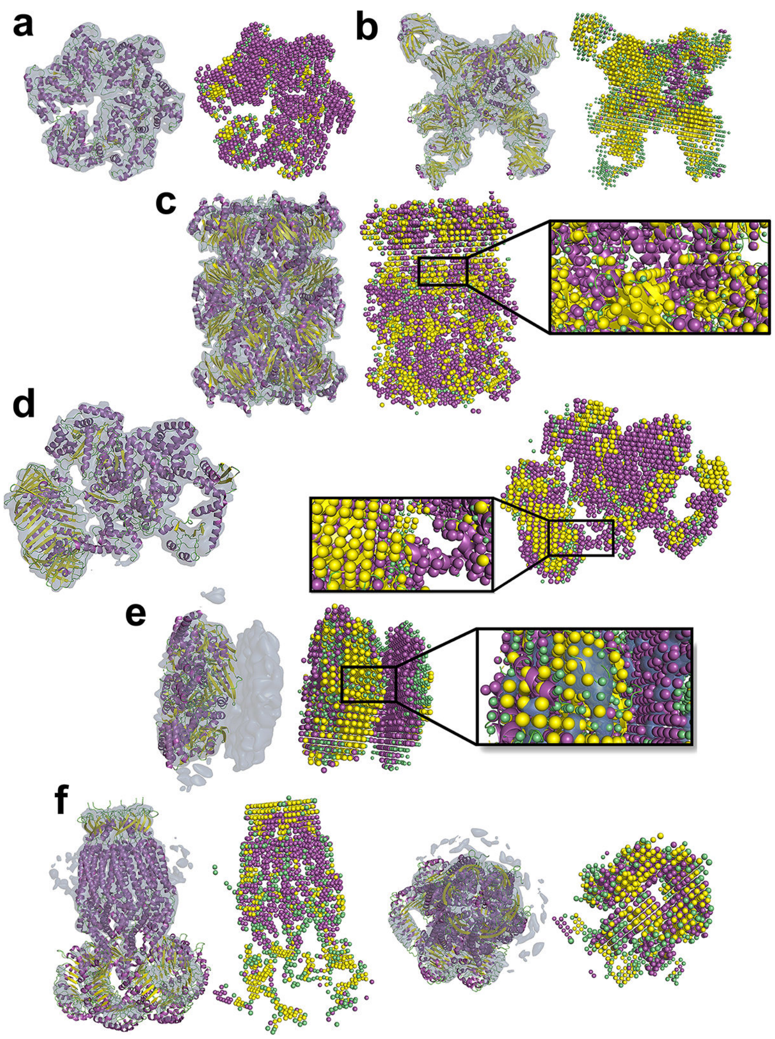

Figure 5.

Examples for experimental maps. The density map and the protein structures (associated PDB structure for the map listed in EMDB) are shown on the left and the secondary structure detection by Emap2sec is shown on the right. Spheres in magenta, yellow, and green show detected α helices, β strands, and other structures, respectively, by Emap2sec. The author-recommended contour level was used to visualize EM maps with a darker color than in Fig. 3. Supplementary Table 3 provides all evaluation values of these maps. **a**, Katanin hexamer, (EMD-8796). Resolution: 6.0 Å. This protein complex has six chains, 1662 residues in total, among which 47.7% are in α helices. The overall (α helices) F1: 0.434 (0.675); Acc: 0.495 (0.825); Q3: 0.622 (0.839); and Seg_{ab} 0.957 (0.941). In the parentheses show are values for α helices. **b**, BG505 SOSIP.664 in complex with antibodies BG1 and 8ANC195 (EMD-8693). Resolution: 6.2 Å. 16 chains and 3940 residues, among which 42.0% are in β strands. The overall (β strands) F1: 0.505 (0.662); Acc: 0.529 (0.728); Q3: 0.676 (0.767); Seg_{ab}: 0.797 (0.788). **c**, Archaeal 20S proteasome (EMD-1733). Resolution: 6.8 Å. 28 chains with 6020 residues. The overall (α helices/β strands) F1: 0.520 (0.677/0.593); Acc: 0.554 (0.797/0.619); Q3: 0.746 (0.8/0.632); Seg_{ab}: 0.757 (0.923/0.740). In the parentheses shown are values for α helices and β strands. **d**, *E. coli* replicative DNA polymerase complex (EMD-3201). Resolution: 8.34 Å. 5 chains with 2219 residues. The overall (α helices/β strands) F1: 0.406 (0.587/0.381); Acc: 0.447 (0.735/0.398); Q3: 0.611 (0.756/0.413); Seg_{ab} 0.560 (0.781/0.441). **e**, bacteriophage phi6 packaging hexamer P4 (EMD-3572). Resolution: 9.1 Å. F1: 0.398; Q3: 0.565. The domain on the right does not have structure assignment by the authors. **f**, mLRRC8A/C volume-gated anion channel (EMD-4361). Resolution: 7.94 Å. The PDB structure shown is from another EM map (EMD-4366) of the same protein (6g9l). F1: 0.571; Q3: 0.862. Two left panels, a side-view; right panels, a top-view. This map was not included in the dataset of 43 experimental maps because it does not have associated PDB structure.

Table 1.

Summary of the secondary structure identification for simulated maps.

Measure[a]	Resolution	α helices	β strands	Others	All
Voxel-based F1 score	6.0 Å	0.844 (0.852)	0.755 (0.771)	0.713 (0.730)	-
	10.0 Å	0.791 (0.799)	0.729 (0.743)	0.664 (0.680)	-
Voxel-based Accuracy	6.0 Å	0.848 (0.853)	0.828 (0.839)	0.672 (0.693)	0.798 (0.811)
	10.0 Å	0.824 (0.828)	0.753 (0.763)	0.637 (0.657)	0.756 (0.769)
Residue Q3	6.0 Å	0.866 (0.866)	0.866 (0.866)	0.718 (0.718)	0.831 (0.831)
	10.0 Å	0.843 (0.843)	0.839 (0.839)	0.681 (0.681)	0.798 (0.798)
Segments	6.0 Å	0.993	0.942	-	0.971[b]
	10.0 Å	0.960	0.905	-	0.938[b]

In parentheses for the F1 score, the accuracy and the residue Q3, values are shown for the relaxed measure, where multiple secondary structures may be considered as correct for a voxel if there are multiple Cα atoms within 3.0 Å that have different secondary structure assignments.

[a] the F1 score and the accuracy are cube-level evaluations. The F1 score is the harmonic mean of precision and recall while the accuracy is the recall. The residue Q3 is the accuracy is the residue-level accuracy (thus recall); and the segment considers the fraction of segments that are correctly identified. See Method for further details of the evaluation measures.

[b] only α helices and β strands were considered.