nature methods

ARTICLES

https://doi.org/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] and Daisuke Kihara[1,2*]

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

As a result of recent technological breakthroughs, cryo-EM has become an indispensable method for determining macromolecular structures[1,2]. Recent years have seen a steep increase in the number of biomolecular structures solved by cryo-EM, including many determined at a high, near-atomic resolution. On the other hand, there are still many structures being solved at intermediate resolutions (5–10 Å) or at even lower resolutions. Of those structures deposited to the Electron Microscopy Data Bank[3] (EMDB) between 2016 and 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 still be determined at intermediate resolutions over the coming decade, as the achievable map resolution is determined by various factors, including the structural flexibility of protein chains, the presence of multiple functional states and 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 more difficult[5]. This is especially true in cases of de novo modeling, which is employed when the structure of the protein has not been solved before or cannot be built by template-based modeling[6–10]. For maps of resolution around 2 Å, software originally designed for X-ray crystallography[11,12] can be used to build an atomic-resolution structure model. At resolutions of around 4 Å, a main-chain tracing method[13,14], such as MAINMAST[15,16], can be applied for modeling. For maps of resolution between 5 and 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 obtain. To aid structural interpretation of maps in this resolution range, several protein secondary structure detection methods have been developed. One class of 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 the Protein Data Bank (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 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. CNNs are very suitable for the identification of local protein structure in three-dimensional (3D) EM maps because the method 'convolves' local map density features to images of a larger region so that local structure detection is performed in the context of a large region of the map. The performance of Emap2sec was tested on two EM map datasets: a dataset of EM maps at resolutions of 6.0 and 10.0 Å that were simulated from 34 individual structures and a dataset of 43 experimental EM maps with resolution ranging from 5.0 to 9.5 Å. The overall residue-level Q3 accuracy, which is the fraction of residues with a correct secondary structure assignment, was 83.1% and 79.8% on average for the simulated maps at 6.0 and 10.0 Å, respectively. For the experimental map dataset, the accuracy was 64.4% on average with a highest recording of 91.6%.

Results

The network architecture of Emap2sec. Emap2sec scans a given EM map with a voxel of 11^3 Å^3 in volume. The voxel reads the density values, which are processed through the deep CNN, and outputs a detected secondary structure (that is, α-helices, β-sheets and what we term '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.

Figure 1b,c illustrates 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 (that is, 1,331 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. Thirty-two, 64 or 128 filters were used at each convolutional layer. Then, the output from the convolutional layers was processed through a maximum-pooling layer and fully connected layers. Finally, a softmax layer outputs a

[1]Department of Computer Science, Purdue University, West Lafayette, IN, USA. [2]Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. *e-mail: dkihara@purdue.edu

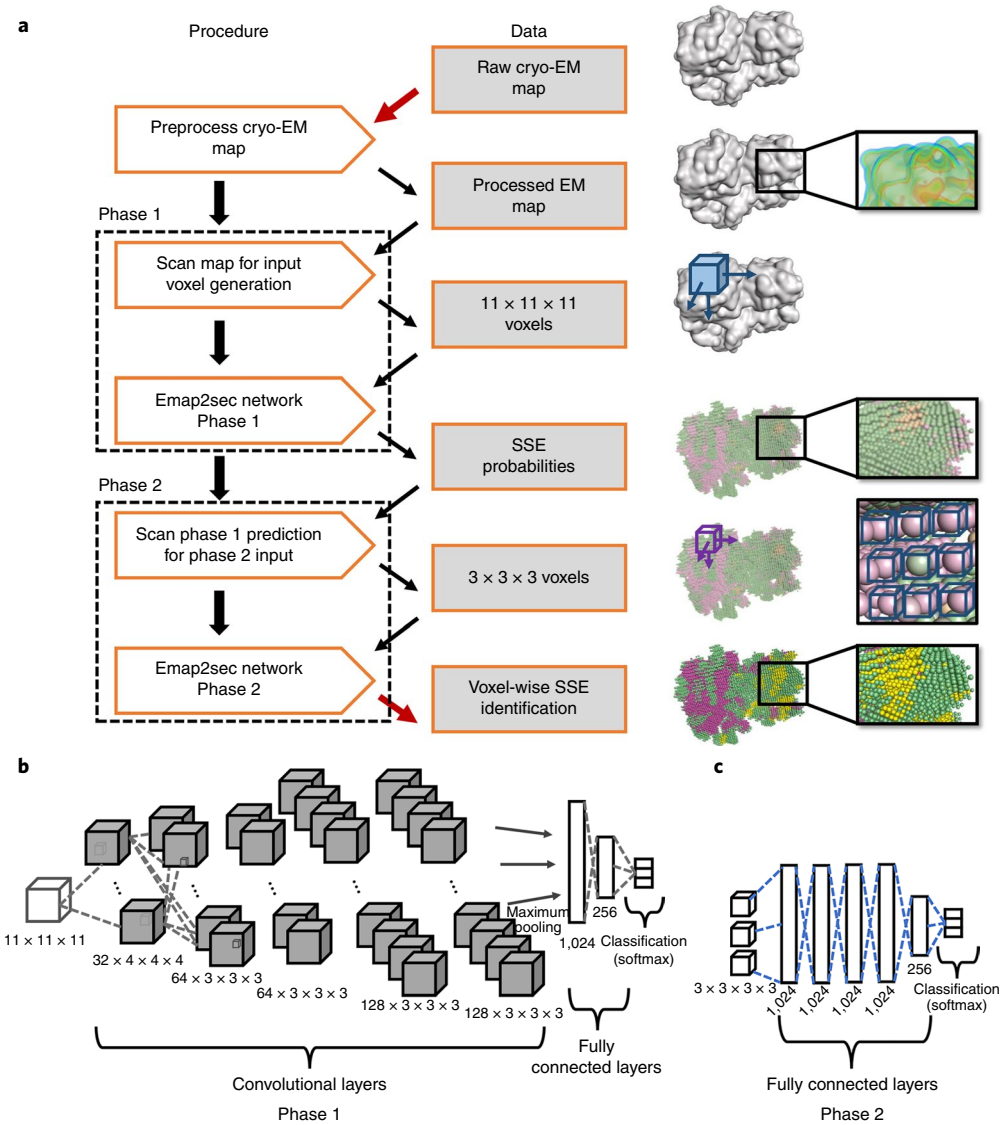

Fig. 1 | The architecture of Emap2sec. a, Emap2sec first takes a 3D EM density map as input, and scans it with a voxel of 11^3 Å^3 in size. There are then two phases in Emap2sec: the phase 1 network takes the normalized density values of a voxel and outputs the probability values of the three secondary structure classes (α-helix, β-sheet and other structure; defined as structure that is not α-helix nor β-sheet); and the phase 2 network takes the output from the phase 1 network and refines the assignment by considering assignments from neighboring voxels. The input to the phase 2 network is a voxel of 3^3 Å^3, where each prediction (shown as a sphere) originates from a voxel of 11^3 Å^3 in the phase 1 network. Finally, each voxel is assigned with a secondary structure class by selecting the class with the largest probability of the three structure types. **b**, The architecture of the phase 1 deep neural networks, consisting of five CNN layers followed by one maximum-pooling layer. The first CNN has 32 filters that are 4^3 Å^3 in size, the second and third CNNs have 64 filters that are 3^3 Å^3 in size, and the fourth and fifth CNNs have 128 filters of 3^3 Å^3 in size. The last layers of the network are two fully connected layers, which have 1,024 and 256 nodes each. The fully connected 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, consisting of five fully connected layers followed by an output layer.

probability value for α-helix, β-strand or other structures. A stride of one was used uniformly in all the filters, whereas a stride of two was used for the maximum-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 individual probability values for α-helices, β-strands and other structures in a voxel of 3^3 Å^3 in size. Thus, there are $3^3 \times 3 = 81$ 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 the target voxel. The purpose of the second phase network is to smooth predictions by considering predictions made for neighboring voxels. For experimental maps, we set the phase 2 network to only change between α-helix and β-strand or from other structure to α-helix or β-strand, because there were more of the other structures in the experimental maps than in the simulated maps. The two-phase network architecture is inspired by the sequence-based protein secondary structure prediction method 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 did not overlap with the data that were used for training (Methods).

Table 1 | Summary of the secondary structure identification for simulated maps

Measure[a]	Resolution	α-helices	β-strands	Others	All
Voxel based	6.0 Å	0.844 (0.852)	0.755 (0.771)	0.713 (0.730)	-
F_1 score	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]

Values for the relaxed measure are shown in parentheses for the F_1 score, the accuracy and the residue Q3. In the relaxed measure, multiple secondary structures were considered as correct for a voxel if there were multiple Cα atoms within 3.0 Å that had different secondary structure assignments. [a]The F_1 score and the accuracy are cube-level evaluations. The F_1 score is the harmonic mean of precision and recall, accuracy is the recall, residue Q3 is the residue-level accuracy (recall) and segments considers the fraction of segments that are correctly identified. See the Methods for further details of the evaluation measures. [b]Only α-helices and β-strands were considered.

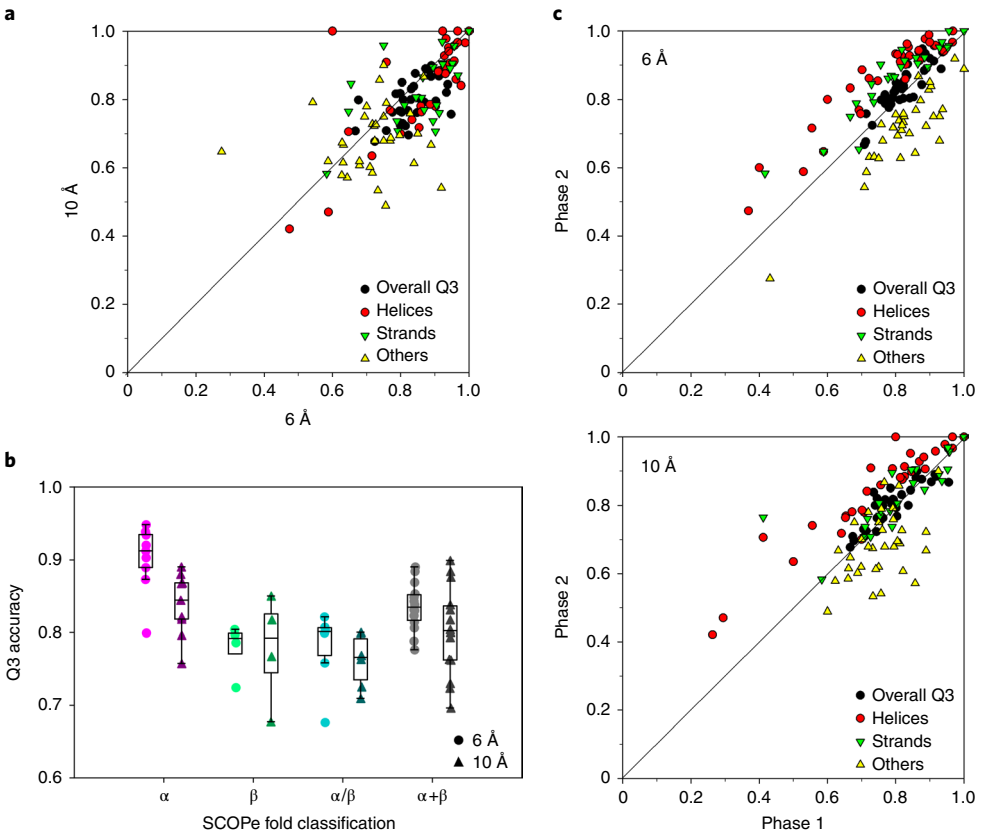

Fig. 2 | Emap2sec performance on the simulated map dataset. The dataset consists of maps of 34 protein structures computed at resolutions of 6.0 Å and 10.0 Å. **a**, Q3-residue-based accuracy by map resolution. **b**, Q3 accuracy values for fourfold classes defined in the SCOPe database are shown in box plots. The number of structures in each class is as follows: α = 9; β = 4; α/β = 6; 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 third (3Q) quartile, and the minimum (Min) and maximum (Max) values. For the α-class (6.0 Å), M = 0.912, 1Q = 0.889, 3Q = 0.934, Min = 0.873, Max = 0.948; for the α-class (10.0 Å), M = 0.844, 1Q = 0.818, 3Q = 0.864, Min = 0.757, Max = 0.890; for the β-class (6.0 Å), M = 0.792, 1Q = 0.771, 3Q = 0.799, Min = 0.771, Max = 0.804; for β-class (10.0 Å), M = 0.792, 1Q = 0.745, 3Q = 0.825, Min = 0.677, Max = 0.850; for the α/β-class (6.0 Å), M = 0.801, 1Q = 0.768, 3Q = 0.806, Min = 0.758, Max = 0.821; for the α/β-class (10.0 Å), M = 0.766, 1Q = 0.735, 3Q = 0.791, Min = 0.709, Max = 0.800; for α+β-class (6.0 Å), M = 0.835, 1Q = 0.816, 3Q = 0.852, Min = 0.776, Max = 0.890; and for the α+β-class (10.0 Å), M = 0.802, 1Q = 0.762, 3Q = 0.836, Min = 0.696, Max = 0.899. **c**, Q3 accuracy before and after applying the phase 2 network. The results for the 6.0-Å maps are shown at the top, and the results for the 10.0-Å maps are shown at the bottom. Raw data for all the maps are provided in Supplementary Table 1.

Table 1 summarizes the accuracy of the secondary structure identification at the voxel, residue and segment levels. For each voxel for which Emap2sec made an individual structure identification, the correct secondary structure was defined as the one from the closest Cα atom within 3.0 Å of the center of the voxel. Emap2sec made fairly accurate structure detections overall. For the 6.0-Å-resolution maps, Emap2sec achieved an average overall residue-level Q3 accuracy of 0.798. Among the three secondary

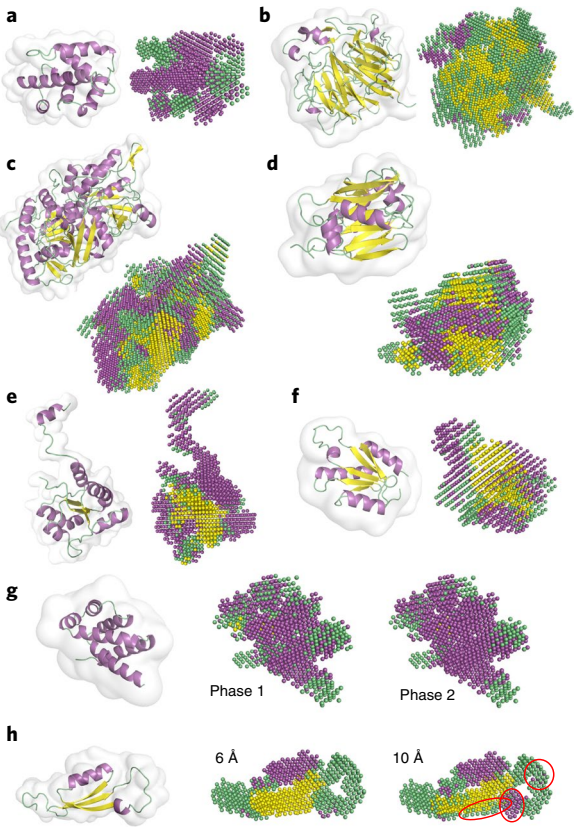

Fig. 3 | Examples of the secondary structure assignment by Emap2sec for simulated maps at resolutions of 6.0 and 10.0 Å. The proteins are taken from SCOPe. For each panel, the main-chain structure of the protein in the simulated EM map is shown on the left while the final structure assignments from the phase 2 network are shown on the right with spheres. Spheres in magenta are assignments of α-helices, yellow spheres are β-strands and green spheres are other structures. See Supplementary Table 1 for all accuracy values of the examples. **a**, A 114-residue α-class protein (transducin α-subunit insertion domain; SCOPe code d1azta1). The map was simulated at 6.0 Å. The overall F_1 score, voxel-based accuracy (Acc), Q3 accuracy and segment-based accuracy of α-helices and β-strands (Seg_{ab}) were 0.908, 0.908, 0.948 and 1.0, respectively. **b**, A 401-residue β-class protein (regulator of chromosome condensation 1; SCOPe code d1a12a_). The map was simulated at 10.0 Å. $F_1 = 0.642$; Acc = 0.641; Q3 = 0.677; $Seg_{ab} = 0.862$. The segment-based accuracy for β-strands was 0.862. **c**, A 527-residue α/β-class protein (aconitase; SCOPe code d1acoa2). The map was simulated at 6.0 Å. $F_1 = 0.750$; Acc = 0.748; Q3 = 0.799; $Seg_{ab} = 0.865$. **d**, A 210-residue α+β-class protein (formaldehyde ferredoxin oxidoreductase; SCOPe code d1b25a2). The map was simulated at 10.0 Å. $F_1 = 0.661$; Acc = 0.665; Q3 = 0.730; $Seg_{ab} = 0.941$. **e**, A 138-residue α/β-class protein (C-terminal domain of carbamoyl phosphate synthetase; SCOPe code d1a9xa2). The map was simulated at 6.0 Å. $F_1 = 0.633$; Acc = 0.670; Q3 = 0.676; $Seg_{ab} = 1.0$. Accuracies for other structures were as follows: $F_1 = 0.396$; Acc = 0.283; and Q3 = 0.275. **f**, A 111-residue α/β-class protein (glycyl-tRNA synthetase; SCOPe code d1atia1). The map was simulated at 10.0 Å. $F_1 = 0.698$; Acc = 0.711; Q3 = 0.768; $Seg_{ab} = 1.0$. Accuracies for other structures were as follows: $F_1 = 0.581$; Acc = 0.48; and Q3 = 0.533. **g**, An example of assignment changes from the phase 1 network to the phase 2 network. A 105-residue α-class protein (N-terminal domain of the δ-subunit of F1F0-ATP synthase; SCOPe code d1abva_). The map was simulated at 10.0 Å. Phase 1: $F_1 = 0.752$ (overall), 0.807 (α-helices) and 0.469 (other structures). Phase 2: $F_1 = 0.834$ (overall), 0.889 (α-helices) and 0.546 (other structures). **h**, An example of structure detection for 6.0- and 10.0-Å maps of a 67-residue protein (allophycocyanin linker chain; SCOPe code d1b33n_). F_1 scores for the 6.0- and 10.0-Å maps: 0.848 and 0.779 (overall), 0.788 and 0.756 (α-helices), 0.893 and 0.820 (β-sheets), and 0.851 and 0.764 (other structures).

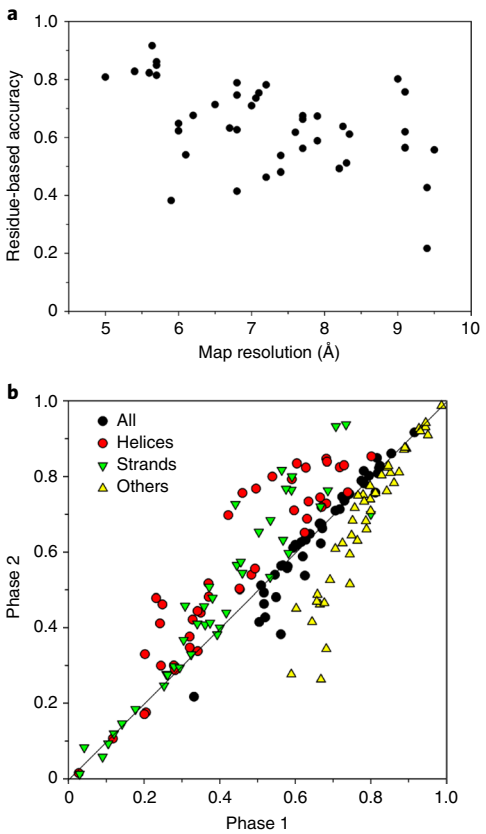

Fig. 4 | Emap2sec secondary structure detection accuracy on 43 experimental maps. See Supplementary Table 3 for details of the accuracy of each map. **a**, Q3 accuracy relative to the map resolution. **b**, Q3 accuracy before and after applying the phase 2 network. Supplementary Table 3 provides raw data for all the maps.

structure classes, α-helices were detected with the highest accuracy (0.848), followed by β-strands with an accuracy of 0.828. At Q3, accuracy higher than the voxel level was recorded. At the segment level, the accuracy values were even higher (0.872 overall). As the residue- and segment-level assignments were made by the majority vote from the level that was one step more detailed (that is, the voxel and residue levels, respectively), the results indicate that our method was very successful in capturing overall main-chain-level protein structures in the maps.

We compared the accuracy between maps simulated at resolutions of 6.0 Å and 10.0 Å (Table 1 and Fig. 2a). The Q3 accuracy dropped for maps with resolution of 10.0 Å; however, the margin of the drop was surprisingly small. The average Q3 values for 6.0- 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 that helices are the most tolerant to noise and visible even in lower-resolution maps.

In Fig. 2b, we show Q3 accuracy for individual fold classes for both 6.0- and 10.0-Å maps. We used the fold class classification by SCOPe. 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 the α/β-class proteins. This is partly because proteins in the α/β-class had more residues in β-strands than α+β-class proteins in this dataset. The average number of residues in β-strands was 47.0 and 40.2 for the

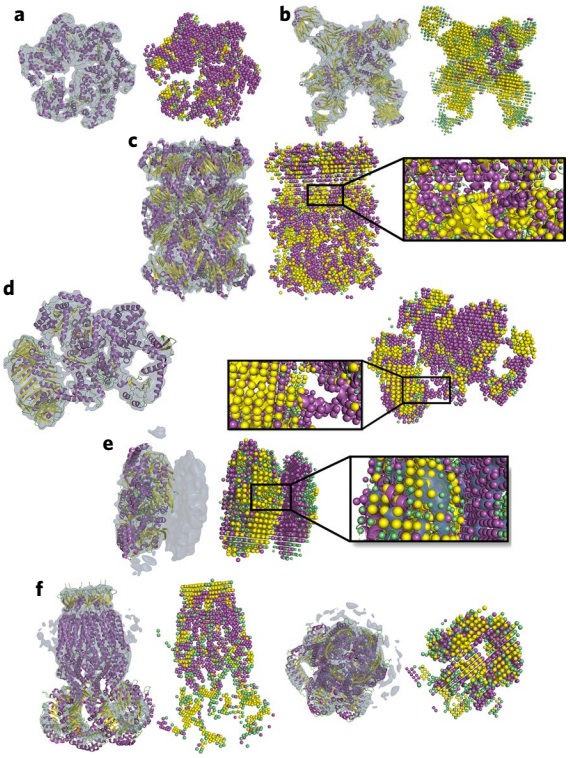

Fig. 5 | Examples of Emap2sec application to experimental maps. The density map and protein structures (associated PDB structure for the map listed in the EMDB) are shown on the left and the secondary structure detection by Emap2sec is shown on the right. Spheres in magenta, yellow and green show detected α-helices, β-strands and other structures, respectively. 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 for these maps. **a**, Katanin hexamer (EMD-8796; PDB 5WCB). The map resolution was 6.0 Å. This protein complex has six chains and 1,662 residues in total, of which 47.7% are in α-helices. Overall accuracies were as follows (accuracies for α-helices are shown in parentheses): $F_1 = 0.434$ (0.675); Acc = 0.495 (0.825); Q3 = 0.622 (0.839); and $Seg_{ab} = 0.957$ (0.941). **b**, BG505 SOSIP.664 in complex with antibodies BG1 and 8ANC195 (EMD-8693; PDB 5VIY). The map resolution was 6.2 Å. This protein complex has 16 chains and 3,940 residues, of which 42.0% are in β-strands. Overall accuracies were as follows (accuracies for β-strands are shown in parentheses): $F_1 = 0.505$ (0.662); Acc = 0.529 (0.728); Q3 = 0.676 (0.767); and $Seg_{ab} = 0.797$ (0.788). **c**, Archaeal 20S proteasome (EMD-1733; PDB 3C91). The map resolution was 6.8 Å. This protein complex has 28 chains and 6,020 residues. Overall accuracies were as follows (accuracies for α-helices and β-strands are shown in parentheses): $F_1 = 0.520$ (0.677 and 0.593); Acc = 0.554 (0.797 and 0.619); Q3 = 0.746 (0.8 and 0.632); and $Seg_{ab} = 0.757$ (0.923 and 0.740). **d**, *E. coli* replicative DNA polymerase complex (EMD-3201; PDB 5FKU). The map resolution was 8.34 Å. This protein complex has 5 chains with 2,219 residues. Overall accuracies were as follows (accuracies for α-helices and β-strands are shown in parentheses): $F_1 = 0.406$ (0.587 and 0.381); Acc = 0.447 (0.735 and 0.398); Q3 = 0.611 (0.756 and 0.413); and $Seg_{ab} = 0.560$ (0.781 and 0.441). **e**, Bacteriophage phi6 packaging hexamer P4 (EMD-3572; PDB 5MUV). The map resolution was 9.1 Å. Overall accuracies were as follows: $F_1 = 0.398$; Q3 = 0.565. The domain on the right does not have a structure assignment from the authors. **f**, mLRRC8A/C volume-gated anion channel (EMD-4361). The map resolution was 7.94 Å. The PDB structure shown is from another EM map (EMD-4366) of the same protein (PDB 6G9L). Overall accuracies were as follows: $F_1 = 0.571$; Q3 = 0.862. A side view is shown on the left and a top view is shown on the right. This map was not included in the dataset of 43 experimental maps because it does not have an associated PDB structure.

α/β- and α+β-class, respectively. Also, by definition, α+β-class proteins have helices and strands in spatially distinct regions, which probably made identification easier for Emap2sec. The Q3 accuracy of α-helices was similar for the α/β- and α+β-classes (0.854 and 0.848, respectively), while the β-strand accuracy was less similar (0.788 and 0.840 for the α/β- and α+β-classes, respectively).

Figure 2c illustrates the effect of the two-phase network architecture of Emap2sec. The results for maps at the two resolutions showed the same trend: the phase 2 network improved the accuracy for α-helices and β-strands, as well as the overall accuracy. On average, the overall Q3 accuracy and the Q3 accuracies of α-helices and β-strands were improved by the phase 2 network by 0.011 and 0.011, 0.093 and 0.089, and 0.062 and 0.021, for the 6.0- and 10.0-Å maps, respectively. In contrast, the accuracy of other structures was 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 were computed using networks trained on 63 EM maps. We also trained the networks on a larger dataset of 1,963 maps (Supplementary Fig. 1 and Supplementary Table 2) but only observed marginal improvement. Changes in the accuracy of individual maps are shown in Supplementary Fig. 1.

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

On the other hand, Fig. 3e,f shows examples where a part of the identification was apparently wrong. In these two cases, extended loop regions that had a somewhat wriggly conformation were misidentified as α-helices. Figure 3g illustrates how the phase 2 network modifies the output from phase 1. The phase 1 output had more other structure assignments, even in helix regions, which were correctly changed to helix assignments in phase 2. Owing to this modification in phase 2, the overall voxel-based accuracy improved from 0.705 to 0.810. Figure 3h shows an example of results for 6.0-Å and 10.0-Å maps of the same protein. The overall F_1 score for the 6.0-Å map was 0.848, which deteriorated to 0.779 for the 10.0-Å map. Noticeable differences are circled on the 10.0-Å map results; some parts of β-sheets were not correctly recognized, and some over-detection of α-helices was 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 (Methods). The number of residues of the proteins in these maps ranged from 664 to 11,475, and the number of chains ranged from 1 to 62. Figure 4a,b shows the residue-based and 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. Notably, high accuracy was also observed for maps at resolutions of around 9.0 Å for example, 0.802 for EMD-1655 (determined at a map resolution of 9.0 Å) and 0.757 for EMD-1263 (determined at a map resolution of 9.1 Å)).

In comparison with the simulated map cases, detecting structures in experimental maps was more difficult. For the simulated map dataset, the average residue-based accuracy was 0.831 and 0.798 for map resolutions of 6.0 Å and 10.0 Å, respectively (Table 1). The average accuracy for experimental maps at resolutions between 5.5 and 6.5 Å was 0.723 and the accuracy for maps at resolutions of

over 9.0 Å was 0.563 (Supplementary Table 3), which is a decrease of 13% and 29% from the simulated map counterparts.

Figure 4b shows the accuracy change by the phase 2 network over the phase 1 network. Similarly to results for the simulated map dataset (Fig. 2c), the accuracy for α-helices and β-strands was consistently improved 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 (Figs. 4 and 5, and Supplementary Table 3) were obtained by using Emap2sec trained on experimental maps. Adding simulated maps in the training set to increase the amount of training data did not yield consistent improvement of detection results. The voxel-based F_1 score improved for 14 maps (32.6%) but deteriorated for 22 maps (51.2%) (with no change for 7 maps). As a consequence, the overall average voxel-based F_1 score slightly deteriorated (Supplementary Fig. 2). Thus, probably because of the different nature of the two types of maps, adding simulated maps in training did not substantially improve structure detection in experimental maps.

Figure 5a,d illustrates cases where Emap2sec performed well. Figure 5a (katanin hexamer) and Fig. 5b (BG505 SOSIP.664 in complex with antibodies BG1 and 8ANC195) represent 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 captured, which yielded high accuracy values. β-sheets in Fig. 5a and α-helices in Fig. 5b, which shared a small portion in the maps, were accurately detected. Figure 5c, showing the 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 residues, respectively). Emap2sec was able to detect the secondary structures distinctively and successfully captured the overall architecture of the structure, which has layers of α-helices and β-sheet domains. Figure 5d shows the *Escherichia coli* replicative DNA polymerase complex, another example of a structure with a similar amount of α-helices (39.8%) and β-strands (22.7%); however, the map resolution is lower, at 8.34 Å. Although the overall accuracy was lower than in Fig. 5c, the β-strand-rich domain on the left in the figure and the α-helix-rich domain and structural details were well captured, including the locations of β-sheets in the α-helix-rich domain and a single helix that bridges the two large domains. Figure 5e shows an example where the detection did not work very well. For this map of bacteriophage phi6 packaging hexamer P4, which was determined at a resolution of 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 this was because many such helices have lower density than β-strands, often even almost outside the contour level, which is the opposite of the usual case where α-helices have higher density than β-strands.

In Fig. 5f, we applied Emap2sec to an EM map determined at a resolution of 7.94 Å, which had no structural assignment (mLRRC8A/C volume-gated anion channel[30]; EMD-4361). Although this map does not have associated PDB files, the authors also deposited two additional higher-resolution EM maps (EMD-4366, determined at 5.01 Å; and EMD-4367, determined at 4.25 Å) with associated PDB entries. When compared with these PDB structures, detection by Emap2sec was quite precise, including β-propeller structures at the top and β-sheets at the bottom (LRRC8A subunits) (the α-helices surrounding the β-sheets were outside the author-recommended contour level; hence, they were not analyzed by Emap2sec). Thus, Emap2sec was able to perform structural assignment even with a map at a resolution of 7.94 Å.

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 could help in assigning β-sandwich structures of Fab proteins as well as three other α-helical chains that were 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, were clearly visible in the detected structure.

Comparison with related works. We found Emap2sec to have substantially better performance than two widely used protein structure modeling methods, Phenix[12] and ARP/wARP (v.8.0)[31], when tested on the dataset of simulated maps at resolutions of 6.0 Å and 10.0 Å (Supplementary Table 4 and Supplementary Fig. 4). As these two modeling methods were not designed for intermediate-resolution maps, the purpose of this comparison is only to illustrate how they differ in nature from Emap2sec.

We also compared Emap2sec with two similar methods, HelixHunter[17], which uses a probe helix to identify helices in a map, and a more recent method published by Li et al.[25], which uses a CNN. Overall Emap2sec showed better performance than these two existing methods on the simulated maps used in the published papers (Supplementary Tables 5 and 6).

Discussion

We developed Emap2sec, a method to detect protein secondary structures in EM density maps of intermediate resolution. This work shows that structure information of proteins can be obtained from intermediate-resolution EM maps more accurately than with conventional methods partly owing to a recent advanced image processing algorithm. A current limitation of Emap2sec is that it detects density regions of secondary structure in a map but does not actually place α-helices and β-strands to the detected local density regions. Expanding the approach to handle other molecules, such as DNA, RNA and lipids as well as large post-translational modifications, is left for future work. Emap2sec should push the limit of extracting structure information and aid structural modeling, and it will be an indispensable tool for interpreting density maps in the era of cryo-EM structural biology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0500-1.

Received: 3 December 2018; Accepted: 24 June 2019;
Published online: 29 July 2019

References

1. Kuhlbrandt, W. Cryo-EM enters a new era. *eLife* **3**, e03678 (2014).
2. Cheng, Y. Single-particle cryo-EM—how did it get here and where will it go. *Science* **361**, 876–880 (2018).
3. Patwardhan, A. Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr. D Struct. Biol.* **73**, 503–508 (2017).
4. Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* **13**, 24–27 (2016).
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).
6. Kirmizialtin, S., Loerke, J., Behrmann, E., Spahn, C. M. & Sanbonmatsu, K. Y. Using molecular simulation to model high-resolution cryo-EM reconstructions. *Methods Enzym.* **558**, 497–514 (2015).
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).
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).

9. Saha, M. & Morais, M. C. FOLD-EM: automated fold recognition in medium- and low-resolution (4–15 Å) electron density maps. *Bioinformatics* **28**, 3265–3273 (2012).
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).
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).
12. Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the PHENIX auto build wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
13. DiMaio, F. et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **12**, 361–365 (2015).
14. Chen, M., Baldwin, P. R., Ludtke, S. J. & Baker, M. L. De novo modeling in cryo-EM density maps with pathwalking. *J. Struct. Biol.* **196**, 289–298 (2016).
15. Terashi, G. & Kihara, D. De novo main-chain modeling for EM maps using MAINMAST. *Nat. Commun.* **9**, 1618 (2018).
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).
17. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308**, 1033–1044 (2001).
18. Dou, H., Burrows, D. W., Baker, M. L. & Ju, T. Flexible fitting of atomic models into cryo-EM density maps guided by helix correspondences. *Biophys. J.* **112**, 2479–2493 (2017).
19. Kong, Y., Zhang, X., Baker, T. S. & Ma, J. A structural-informatics approach for tracing β-sheets: building pseudo-Cα traces for β-strands in intermediate-resolution density maps. *J. Mol. Biol.* **339**, 117–130 (2004).
20. Si, D. & He, J. Modeling β-traces for β-barrels from cryo-EM density maps. *Biomed. Res. Int.* **2017**, 1793213 (2017).
21. Si, D. & He, J. Tracing β strands using StrandTwister from cryo-EM density maps at medium resolutions. *Structure* **22**, 1665–1676 (2014).
22. Lindert, S. et al. EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure* **20**, 464–478 (2012).
23. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
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).
25. Li, R. J., Si, D., Zeng, T., Ji, S. W. & He, J. Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (eds Tian, T. et al.) 41–46 (IEEE, 2016).
26. Russakovsky, O. et al. Image net large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 1–42 (2015).
27. Maturana, D. & Scherer, S. VoxNet: a 3D convolutional neural network for real-time object recognition. in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 922–928 (IEEE, 2015).
28. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195 (1999).
29. Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
30. Deneka, D., Sawicka, M., Lam, A. K. M., Paulino, C. & Dutzler, R. Structure of a volume-regulated anion channel of the LRRC8 family. *Nature* **558**, 254–259 (2018).
31. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171–1179 (2008).

Acknowledgements

The authors acknowledge C. 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 and CMMI1825941) and the Purdue Institute of Drug Discovery.

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 interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0500-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.K.

Peer review information: Allison Doerr was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 for 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] (v.2.06) and removed structures if they had over 25% sequence identity with another structure in the dataset. This left 2,000 structures (Supplementary Table 7). Next, we used the e2pdb2mrc program from the EMAN2 package[32] (v.2.11) to generate simulated EM maps for each structure at resolutions of 6.0 Å and 10.0 Å. The simulated maps have been made available at Code Ocean[33] in two directories, 'data/simulated_maps/6' and 'data/simulated_maps/10'. Density values were normalized in each map to the range 0–1.0 by subtracting the minimum density value in the map and then dividing by the density difference between the maximum and the minimum density values. If there were negative density values, they were set to 0 before the normalization.

The input density data of Emap2sec were voxels of size 11 Å × 11 Å × 11 Å; we also tried a smaller size, 5 Å × 5 Å × 5 Å, but this performed substantially worse. The data were obtained from a map by traversing along the three dimensions of the voxels with a stride of 2.0 Å. Each voxel was assigned with the closest Cα atom that was within 3.0 Å of the center of the voxel, and each Cα atom was assigned a secondary structure type using STRIDE[34]. Residues that had structure codes of H, G or I, as assigned by STRIDE, were labeled as α-helices, while those with codes of B/b or E were labeled as β-strands.

The training dataset for the phase 1 network consisted of 31,951 voxels each for α-helix, β-strand and other structures (95,853 voxels in total). To select these voxels, the 2,000 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 for α-helices, β-strands and other structures, respectively. As the number of voxels was sufficient and helices and strands appeared in various orientations in the voxels, we did not perform data augmentation by rotating maps. This training dataset was constructed at resolutions of 6.0 Å and 10.0 Å. The training was performed independently for resolutions of 6.0 Å and 10.0 Å.

The phase 1 network consisted of five layers of CNN followed by a fully connected network (Fig. 1). This is a standard architecture for a CNN. With this training dataset, we performed a tenfold cross-validation to determine hyperparameters, 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 and 100) and the learning rate values tested were (0.0001, 0.001, 0.01, 0.1, 1 and 10). For the first fully connected layer (the connection between the layer with 1,024 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 of which were used for training under all possible hyperparameter 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, and the L_1 regularization parameter λ_2 of the fully connected layer was 0.01. Using this procedure, the learning rate was determined to be 0.001 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 (32 each for 6.0 Å and 10.0 Å). After the phase 1 network was fully trained, we input the 32 maps to the phase 1 network and obtained probability values for α-helices, β-strands and other structures, which were input for the phase 2 network. We trained the phase 2 network for the L_1 regularization parameter in the same way using a fixed learning rate of 0.001. The obtained parameter was 0.1 for both 6.0-Å and 10.0-Å resolutions. For the phase 2 network, we did not use a CNN because the purpose is to locally smooth outputs or detected secondary structures.

After training of the phase 1 and 2 networks was completed, we applied them to the 34 test maps, which were not included in the maps used for the training.

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

Experimental EM map dataset. We also trained Emap2sec using actual EM maps retrieved from EMDB[3]. Density maps that were determined at a resolution of between 5.0 Å and 10.0 Å and had an associated atomic structure in PDB were selected. Then, to ensure that a map and its associated structure had 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[35], and only maps with a cross-correlation of over 0.65 were kept. Finally, we computed the sequence identity between proteins in pairs of EM maps, and a map was removed from the dataset if its protein had over 25% identity to a protein of another map in the dataset. If a map had multiple protein chains, the map was removed if at least one of the chains had over 25% identity to any chains in another map. After this procedure, 43 experimental EM maps were retained (Supplementary Table 3). The grid size of the maps was unified to 1.0 Å by applying trilinear interpolation of the electron density in the maps. Secondary structure detection by Emap2sec was evaluated only for voxels that had associated protein structures (that is, voxels for which correct secondary structure at that position was known).

The training and testing were performed using experimental maps at resolutions between 6.0 and 10.0 Å using fourfold cross-validation. The same hyperparameter values established for the simulated maps were used for experimental maps. The trained model was further applied to experimental maps at resolutions between 5.0 and 6.0 Å. To perform fourfold cross-validation, 43 maps were split into four groups of 11, 11, 11 and 10 maps. Then, three out of the four groups were used for training, and the resulting network was tested on the remaining group. This process was repeated four times by changing the group for testing.

Evaluation measures. Secondary structure detection by Emap2sec was evaluated at the voxel, amino acid residue and 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 was within 3.0 Å of the center of the voxel, as discussed in the previous section. In Table 1, we also show results with a relaxed measure, where a voxel was assigned with the secondary structure of all Cα atoms within 3.0 Å, thus potentially allowing multiple correct secondary structures. For the voxel-level evaluation, we used the F_1 score, which is the harmonic mean of the precision and recall of the assignments given to the entire map or to each secondary structure class

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where precision is the fraction of voxels with correct secondary structure detection over all the voxels of 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 F_1 score was computed as the weighted F_1 score of the three secondary structure classes. For the residue-level evaluation, we report the Q3 accuracy, which is the fraction 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 were within 3.0 Å of the atom had the same secondary structure assignment. We also report the segment-level accuracy. A secondary structure segment was defined as a stretch of amino acids with the same secondary structure type—at least six amino acids for an 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 had the correctly assigned class label.

Comparison with other published tools. Emap2sec was compared with Phenix (Supplementary Table 4), ARP/wARP (Supplementary Table 4), HelixHunter[29] (Supplementary Table 5) and the method published by Li et al.[25] (Supplementary Table 6). The comparison with Phenix and ARP/wARP was performed by running these two methods on the simulated maps at resolutions of 6.0 Å and 10.0 Å. In the comparison with HelixHunter and the method published by Li et al., we trained Emap2sec on the same training set of simulated maps at resolutions of 6.0 Å and 10.0 Å. The maps that were used for training did not have more than 25% identity to the test maps of HelixHunter and the maps in the method published by Li et al.

Software. The following software packages were used: Pymol (v.2.3; https://www.pymol.org/2/) and Chimera (v.1.13.1; https://www.cgl.ucsf.edu/chimera/download.html) were used for visualization of protein structure and maps; STRIDE (http://webclu.bio.wzw.tum.de/stride/install.html) was used for for secondary structure definition; EMAN2 (v.2.11; https://blake.bcm.edu/emanwiki/EMAN2/Install) was used for computing simulated EM maps; Tensorflow (v.1.x; https://www.tensorflow.org/install/pip) and Python (v.2.7 and v.3.6; https://www.python.org/downloads/) were used for the development of Emap2sec; and Phenix (v.1.14; https://www.phenix-online.org/download/nightly_builds.cgi) and ARP/wARP (v.8.0; http://www.embl-hamburg.de/ARP/) were used for comparison with Emap2sec.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw data of accuracies are provided in Supplementary Tables 1, 3 and 4. The experimental EM maps can be downloaded from EMDB (Supplementary Table 3). Output files from Emap2sec for the simulated and experimental maps 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 Code Ocean[33] and via http://www.kiharalab.org/emap2sec/index.html and https://www.github.com/kiharalab/Emap2sec. Simulated maps are available in the Code Ocean code capsule.

References

32. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
33. Subramaniya, S. R. M. V., Terashi, G. & Kihara, D. *Protein secondary structure detection in intermediate resolution cryo-electron microscopy maps using deep learning* v.2.0 (Code Ocean, 2019); https://doi.org/10.24433/CO.3068754.v2
34. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).
35. Monroe, L., Terashi, G. & Kihara, D. Variability of protein structure models from electron microscopy. *Structure* **25**, 592–602 (2017).

nature research

Corresponding author(s): Daisuke Kihara

Last updated by author(s): 2019-5-1

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Pymol (ver2.3) (https://pymol.org/2/); Chimera (ver. 1.13.1) (https://www.cgl.ucsf.edu/chimera/download.html) both for data visualization
Data analysis	STRIDE (http://webclu.bio.wzw.tum.de/stride/install.html) for defining secondary structure; EMAN2 (ver. 2.11) (https://blake.bcm.edu/emanwiki/EMAN2/Install) for computing simulated maps; Tensorflow (ver. 1.x) (https://www.tensorflow.org/install/pip) and Python (ver. 2.7 and 3.6) (https://www.python.org/downloads/) for developing the software Emap2sec; Phenix (ver. 1.14) (https://www.phenix-online.org/download/nightly_builds.cgi) and ARP/wARP (ver. 8.0) (http://www.embl-hamburg.de/ARP/) for detecting secondary structure from EM maps and compare the performance with Emap2sec.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data of accuracies are provided in Supplementary Information, Supp. Table S1, S3, and S4. 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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	PDB and EMDB entries used to train and test the machine learning method were shown in Supplemental data. Regarding the training data, in addition to the final data size we used we also used a larger datasets and showed that the larger training data did not make meaningful improvement, which shows that the current training data size is sufficient. The number of voxels used in training for simulated maps was 95,853. For training the network for real EM maps, we did a four-fold crossvalidation using 43 em maps.
Data exclusions	For selecting experimental EM maps, 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 examined34 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.
Replication	This work is on the computational method development and no biological experimental data involved. Detection of the secondary structure with Emap2Sec was not repeated for each map because the algorithm gives the same results for the same map.
Randomization	This work is on the computational method development and no biological experimental data involved, thus the randomization of data is not applicable.
Blinding	This work is on the computational method development and no medical trials involved, thus blinding is not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging