2017

An Effective Computational Method Incorporating Multiple Secondary Structure Predictions in Topology Determination for Cryo-EM Images

Abhishek Biswas
Old Dominion University

Desh Ranjan
Old Dominion University

Mohammad Zubair
Old Dominion University

Stephanie Zeil
Old Dominion University

Kamal Al Nasr

See next page for additional authors

Original Publication Citation

Biswas, A., Ranjan, D., Zubair, M., Zeil, S., Al Nasr, K., & He, J. (2017). An effective computational method incorporating multiple secondary structure predictions in topology determination for cryo-em images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14*(3), 578-586. doi:10.1109/tcbb.2016.2543721

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2017 ; 14(3): 578–586. doi:10.1109/TCBB.2016.2543721.

An Effective Computational Method Incorporating Multiple Secondary Structure Predictions in Topology Determination for Cryo-EM Images

Abhishek Biswas[1], **Desh Ranjan**[1], **Mohammad Zubair**[1], **Stephanie Zeil**[1], **Kamal Al Nasr**[2], and **Jing He**[1,*]

[1]Dept. of Computer Science, Old Dominion University, Norfolk, VA 23529

[2]Dept. of Computer Science, Tennessee State University, Nashville, TN 37209

Abstract

A key idea in *de novo* modeling of a medium-resolution density image obtained from cryo-electron microscopy is to compute the optimal mapping between the secondary structure traces observed in the density image and those predicted on the protein sequence. When secondary structures are not determined precisely, either from the image or from the amino acid sequence of the protein, the computational problem becomes more complex. We present an efficient method that addresses the secondary structure placement problem in presence of multiple secondary structure predictions and computes the optimal mapping. We tested the method using 12 simulated images from α-proteins and two Cryo-EM images of α-β proteins. We observed that the rank of the true topologies is consistently improved by using multiple secondary structure predictions instead of a single prediction. The results show that the algorithm is robust and works well even when errors/misses in the predicted secondary structures are present in the image or the sequence. The results also show that the algorithm is efficient and is able to handle proteins with as many as 33 helices.

Keywords

Cryo-electron Microscopy; Dynamic Programming; Graph; Image; Protein; Secondary Structure

1 Introduction

THE field of cryo-electron microscopy (Cryo-EM) has undergone dramatic growth over the last several decades. Cryo-EM has become a major technique in the structure determination of large molecular complexes [4, 5]. Unlike X-ray crystallography and Nucleic Magnetic Resonance (NMR), Cryo-EM is particularly suitable for large molecular complexes, such as viruses, ribosomes, and membrane-bound ion channels [6–8]. For density maps (3D images) with high resolution (2–4 Å), the atomic structure can be derived directly, since the

*Corresponding author: jhe@cs.odu.edu.

Authors' contribution: AB, DR, MZ, and JH developed the algorithm, and AB implemented it. AB and SZ conducted the test. KA helped with software design.

backbone is mostly resolved. However, it is computationally challenging to derive atomic structures when the backbone of the protein is not resolved from the density maps, such as those with resolutions lower than 5 Å. In current approaches, a known atomic structure or a model built from known atomic structures is fit into the Cryo-EM density map [10–14]. However, those approaches are limited by the need for atomic structures that are either components of or homologous to the atypically-sized protein. When there is no template structure with sufficient similarity, *de novo* methods must be devised and used. These methods do not rely on templates and they aim to derive the structure from the intrinsic relationship among the secondary structures visible in the density map.

Although it is not possible to distinguish amino acids, most secondary structures, such as α-helices (red sticks in Fig. 1A) and (β-sheets, can be computationally identified from a density map with medium resolutions, such as 4–8Å [18–23]. Once a β-sheet density region is identified, β-strands may be predicted using StrandTwister by analyzing the twist of a β-sheet [25, 27]. A helix detected from a Cryo-EM image can be represented as a line–referred to here as an α-trace-that corresponds to the central axis of a helix (shown as red sticks in Fig. 1A). Similarly, a β-strand can be represented as a β-trace that corresponds to the central line of the β-strand (see Section 3.5 for more details). The term, secondary structure traces (SSTs), refers to the set of α-traces and β-traces detected from the 3-dimensional (3D) image (Fig. 1A).

In order to help determine the threading of the protein sequence through the SSTs, a computational method, such as JPred [9], is used to predict the subsequences (sequence segments) of the protein sequence that are likely to be the secondary structures. These subsequences are then mapped the SSTs. The *topology* of the SSTs refers to their order with respect to the protein sequence and the direction of each SST. For example, in Fig. 1, D_1 through D_{18} represent SSTs and S_1 through S_{18} represent the subsequences on the protein chain that correspond to the secondary structures. In this case, SSETracer was able to detect 18 of 20 helices (red sticks in Fig. 1A) from the 3D image. Each SST corresponds to a sequence segment in the true topology. For example, S_1 is mapped to D_{10} and S_2 is mapped to D_{15}. The order of SSTs in the true topology is $(D_{10},D_{15},D_{13},D_8,D_{12},D_{11},D_{14},D_{16},D_{17},D_{18},D_9,D_7,D_5,D_6,D_2,D_3,D_1,D_4)$. In other words, they are mapped to $(S_1,S_2,S_3,S_4,S_5,S_6,S_7,S_8,S_9,S_{10},S_{11},S_{12},S_{13},S_{14},S_{15},S_{16},S_{17},S_{18})$. Note that there are two possible directions when mapping a sequence segment to an SST (the arrows shown in Fig. 1A and the dots/crosses shown in Fig. 1B), since the sequence of a protein has a direction.

Previously, we have shown that finding the optimal mapping between SSTs and the sequence segments is an NP-hard problem [28]. A naïve approach used to find the optimal solution requires $\Omega(N!2^N)$ time, where N is the number of SSTs. A dynamic programming algorithm has been previously devised to find the optimal match in $O(N^2 2^N)$ computation time, reduced from $O(N!2^N)$ as in a naïve approach. In a general case where M sequence segments are mapped to N SSTs (assuming $M \geq N$, $\Delta = M - N$), we previously developed a constrained dynamic programming algorithm and a K shortest path algorithm, DP-TOSS, to find top K best mappings in $O(\Delta^2 N^2 2^N)$ time [24].

Deriving the optimal secondary structure topology is more than a mapping problem. The accuracy of secondary structure prediction is about 80%, [17, 29–31], which is similar to the accuracy of the detection of SSTs from medium-resolution images [19, 23]. No single prediction method is superior to any other prediction method for all secondary structures. In topology determination, alternative positions for an individual secondary structure must often be considered (Fig 2.). However, that results in a significant computational cost. Let N be the number of secondary structures in a protein. Suppose there are a maximum of p alternative positions for each helix segment on the sequence and q alternative positions for each of the SSTs, then there are $p^N q^N$ possible pairs of secondary structure sets to be matched. The total number of possible matches will be $p^N q^N N! 2^N$, since there are $N! 2^N$ different ways (or different topologies) to map a set of SSTs to a set of sequence segments.

Previous algorithms, such as DP-TOSS [24] and Gorgon [32], predominantly address the mapping problem. A placement problem arises when alternative positions of the secondary structures need to be considered. One either has to submit the best estimated secondary structure positions to DP-TOSS or Gorgon or run either of these two programs multiple times using alternative positions that are produced from multiple secondary structure prediction servers. We previously attempted a dynamic graph approach in which the alternative positions are handled in the graph update process [33]. That approach yielded, on average, a running time that was about 34% lower than the naïve approach. To reduce the computational cost even more, we designed an effective two-step approach, the outline of which was presented at a conference [34]. In this paper, we demonstrate the effectiveness of this approach with new data and enhanced results. The two-step approach utilizes two dynamic programming algorithms, one in DP-TOSS to derive the top K topologies using a consensus secondary structure prediction, and another to derive optimal placement for each of the top K topologies. We have compared the two-step approach with the brute-force approach and have shown that, in principle, the approach is applicable to alternative positions of SSTs in the 3D image [35]. We also demonstrate the effectiveness of the two-step approach in handling alternative sequence segments predicted from multiple secondary structure prediction servers. Moreover, the results show that the ranking of the true topology improved when using multiple secondary structure predictions in comparison to any of the single prediction methods that were tested.

2 Methods

2.1 The Secondary Structure Mapping Problem

For a general protein, suppose there are N_α helices and N_β β-strands detected from a 3D image, and $N = N_\alpha + N_\beta$. Also assume that there are M_α helices and M_β β-strands predicted from the amino acid sequence of the protein, and $M = M_\alpha + M_\beta$. To simplify the description, we assume $M_\alpha = N_\alpha$ and $M_\beta = N_\beta$; consequently, $M = N$. Our actual algorithm and implementation handle the case where $M \neq N$. Let the sequence segments of the secondary structures be $\{S_1, S_2, \ldots, S_N\}$, where S_i denotes the i^{th} sequence segment from the N-terminal of the protein. Note that the direction of the protein sequence is from the N-terminal to the C-terminal. Let the SSTs of the 3D image be $\{D_1, D_2, \ldots, D_N\}$. For convenience, let $D_1, D_2, \ldots, D_{N_\alpha}$ be α-traces and $D_{N_\alpha+1}, D_{N_\alpha+2}, \ldots, D_{N_\alpha+N_\beta}$ be β-traces.

The secondary structure mapping problem is to find a mapping σ such that S_i is mapped to $D_{\sigma(i)}$, $i = 1,2,\ldots, N$ and the following two criteria are satisfied: (1) both S_i and $D_{\sigma(i)}$ correspond to either a helix or a β-strand and (2) the mapping score is optimal. An optimal SST topology corresponds to a mapping with the optimal score that often evaluates the overall differences between the two sets of secondary structures. The differences can be measured using various factors, such as the length of the secondary structures, the distance between two consecutive secondary structures, and the likelihood that the amino acids are on a loop [24] [36–38]. The scoring function used in this paper consists of a skeleton length between two secondary structure traces, the length of a secondary structure, and the loop length on the protein sequence.

Given a specific set of secondary structure traces and a specific set of predicted secondary structure sequence segments, K best mappings were determined using DP-TOSS. In the first step, a set of sequence segments predicted by a consensus secondary structure prediction server was used. The idea is to use the best estimation of the secondary structure positions in the first step in order to obtain a small number of possible topologies. For each possible topology, the best placement of the secondary structures will be determined in the second step.

2.2 Dynamic Programming for Finding Optimal Placement

Let us represent the alternative sequence segments for the secondary structure as the following. Let $\left(S_i, \alpha_i^l\right)$ be the l^{th} alternative for sequence segment S_i, where $l = 1,2,\ldots,p, i = 1,2,\ldots,N$. In other words, there are a maximum of p alternatives for each of the segments. For a given topology, mapping σ is known. The optimal placement problem is to find the placement of $\alpha_i^{l_i}, 1 \le l_1, l_2, \ldots, l_N \le p$ for each sequence segment $S_i, i = 1,2, \ldots, N$, such that the score of mapping $\left(S_1, \alpha_1^{l_1}\right), \left(S_2, \alpha_2^{l_2}\right), \ldots, \left(S_N, \alpha_N^{l_N}\right)$ to $(D_{\sigma(1)}, D_{\sigma(2)}, \ldots, D_{\sigma(N)})$ is minimized.

A naïve approach to finding the best placement of a topology is to exhaustively score the p^N different placements. Below, we show a dynamic programming algorithm in which we store and reuse information. Let $g(i, k)$ denote the best cost that can be obtained when $(S_1, S_2, \ldots, S_i)$ is mapped to $(D_{\sigma(1)}, D_{\sigma(2)}, \ldots, D_{\sigma(i)})$ with the k^{th} placement α_i^k used for S_i. Then, for any position $\alpha_{\sigma(i+1)}^{k'}$ of S_{i+1}, $g(i + 1, k')$ is only affected by the values $g(i, k)$, where $k = 1,2, \ldots,p$, and the score for positioning the i^{th} mapped segment and the $(i + 1)^{th}$ mapped segment. More precisely, for $k' \in \{1,2, \ldots,p\}$,

$$
\begin{aligned}
g\left(i+1, k'\right) = \min_{k \in \{1,2,\ldots,p\}} \big(& g(i,k) \\
& + \left| l\left(S_{(i+1)}, \alpha_{(i+1)}^{k'}\right) - l\left(D_{\sigma(i+1)}\right) \right| \\
& + \left| d\left(\left(S_i, \alpha_i^k\right), \left(S_{(i+1)}, \alpha_{(i+1)}^{k'}\right)\right) \right. \\
& \left. - \delta\left(D_{\sigma(i)}, D_{\sigma(i+1)}\right) \right| \big)
\end{aligned}
$$

Note that $l(D_{\sigma(i+1)})$ measures the length of SST ($D_{\sigma(i+1)}$) and $\delta(D_{\sigma(i)},D_{\sigma(i+1)})$ measures the length along the skeleton between $D_{\sigma(i)}$ and $D_{\sigma(i+1)}$. Ideally, $\delta(D_{\sigma(i)}, D_{\sigma(i+1)})$ corresponds to the length of the loop connecting the two secondary structures $D_{\sigma(i)}$ and $D_{\sigma(i+1)}$ and $d(a,b)$ measures the loop length between two consecutive secondary structures, a and b, on the sequence.

2.3 Secondary Structure Predictions from Multiple Servers

Secondary structure prediction was performed using five online servers (SYMPRED [3], JPred [9], PSIPRED [17], PREDATOR [26], and Sable [39]). SYMPRED and JPred are consensus servers. The initial positions include the predicted positions using either SYMPRED or JPred, whichever predicted a greater number of helices. These initial positions were used to obtain the initial topologies using DP-TOSS. Alternative positions of each secondary structure were generated based on the results from the multiple secondary structure predictions.

3 Results

The accuracy and efficiency of the two-step approach were tested using 12 α-proteins and two Cryo-EM proteins that contain both α-helices and β-sheets. While α-proteins do not contain β-sheets, they provide test cases for large proteins. The length of the α-proteins ranged from 142 amino acids (1FLP) to 585 amino acids (2XVV). Therefore, the α-protein dataset is suitable for testing the efficiency of the method and its capability of handling large complicated cases in topology determination. For the α-protein dataset, the atomic structures were downloaded from the Protein Data Bank (PDB), and they were used to simulate density maps at 10Å resolution using EMAN software. The two Cryo-EM test cases use experimentally derived Cryo-EM density maps (EMD-5030-4V68_BR and EMD-1780-3IZ6_K) downloaded from the Electron Microscopy Data Bank (EMDB) [40]. The atomic structures of chain BR of 4V68 (PDB ID) and chain K of 3IZ6 (PDB ID) were used to extract the density regions that correspond to the chains.

3.1 The Accuracy of the Helix Detection from the Density Images

The helices and β-sheets were detected from the density maps using SSETracer [2], and the β-strands were detected using StrandTwister [27]. Since the accuracy of topology determination is affected by the accuracy of the detected secondary structures, we discuss the detection accuracy in detail. The accuracy was evaluated at two different levels: the number of detected helices and the number of detected C_α atoms on the helices [23]. We observed that short helices tend to be missed in the detection, particularly those that are shorter than three turns. SSETracer detected all of the helices for three of the 12 cases (Rows 1, 2, and 4 in Table 1). Short helices were missed in the detection for the other nine cases. However, our previous experience and the results in this paper have shown that short helices play a minor role in the detection of the correct topology. A helix may be detected longer or shorter than it is; thus, fine measurement is needed to evaluate the accuracy. For example, although SSETracer detected all seven helices in 1FLP (Row 1, Table 1), some of the helices were detected slightly shorter, since the sensitivity is 93.94%. Some of the helices might be detected longer than or shifted from the actual helix, since the specificity is 72.09% (Row 1,

Table 1). It is often more accurate to measure the number of detected C_α atoms on the helices. A C_α atom on a helix is considered to be a detected atom if it is within 2.5Å from an identified helix voxel. For the simulated dataset shown in Table 1, SSETracer was able to detect most of the helices with an average specificity of 82.24% and an average sensitivity of 84.83%. This suggests that the ability of SSETracer to detect the helices in this dataset was fairly accurate. We noticed that secondary structure prediction servers, such as SYM-PRED and JPred, predicted a greater number of helices than SSETracer (Columns 4 and 5, Table 1). DP-TOSS was designed to incorporate non-identical numbers of the secondary structures predicted from the sequence and from the 3D image.

3.2 Topology Ranking in the Two-Step Approach

The input into the two-step approach includes three components: the α-traces detected using SSETracer (blue sticks in Fig. 3B), the skeleton derived using SkelEM [15](yellow density in Fig. 3B), and five secondary structure predictions obtained from different online servers (Fig. 2). In the first step, the top 1000 ranked topologies were derived using the consensus prediction obtained from either SYMPRED or JPred. For each of the possible topologies, optimal placement was searched using the newly devised dynamic programming algorithm and secondary structure predictions from the five online servers. We use the largest test case, 2XVV, as an example to illustrate the data and process used in the two-step approach. In this case, about 74% of the helices were detected from the 3D image, and 14 short helices were missed. Two helices are immediate neighbors on the sequence and they were detected as one long helix. One question arises: Is it still possible to distinguish the true topology if only about 74% of the helices are detected? Using the two-step approach, the correct topology for the 19 detected SSTs was ranked 8th, near the top of the list, considering that there are $p^N \begin{pmatrix} M \\ N \end{pmatrix} N! 2^N$ possible ways to match, where $p = 5, M = 28, N = 19$ for 2XVV. We observed that, to some extent, the skeleton may compensate for the mistake in the secondary structure detection. For example, even though some of the short helices were not detected, the skeleton still passes through the region of the missed helix. Since the missed helices are generally short, the effect of a missed helix is reduced due to the existence of the skeleton. In fact, if the true sequence segments are used in topology determination, in this present case, the true topology was ranked 4th (Column 3, Table 2). Note that although the skeleton in this case is fairly clear, ambiguity is often observed in the skeleton. An ambiguity point is where multiple skeleton braches meet at the same point, leading to multiple ways to connect the secondary structures. This ambiguity is resolved in the dynamic programming graph and the search for the constrained shortest path [24].

3.3 Using Multiple Secondary Structure Predictions Versus Using Single Prediction

Many secondary structure prediction methods are available and some of them provide online services. Although certain methods are more accurate than others, overall, we observed that no single method is superior to any other for all the secondary structures. For example, certain helices are predicted more accurately by SYMPRED, but others are predicted more accurately by different methods (Fig. 2). To utilize the advantage of all the prediction methods, it is always important to use all of the predictions. However, doing so results in

significant computational overhead. We present a two-step approach in which alternative positions are generated from all predictions in the second step. A dynamic programming placement method is devised to quickly find the best alternatives that fit the constraints from all of the SSTs. We tested the two-step approach on 12 large proteins. The rank of the true topology was used to evaluate the accuracy for each of the seven methods. The difference among the seven methods resides in the input of the secondary structure positions on the sequence obtained by (1) PDB, (2) SYMPRED, (3) JPred, (4) PSIPRED, (5) PREDATOR, (6) Sable, and (7) all five of the online predictions. Intuitively, the best accuracy comes from the use of the true secondary structure positions on the sequence. Amazingly, this is true for only four of the 12 cases (Columns 3, 9, Table 2). Since the SSTs detected from the image are not 100% accurate, using the true sequence position for the helices in the matching is not always the best approach. For example, the correct topology was ranked 4th for 3LTJ when all five predictions were used, but it was ranked 9th when the true sequence segments of the helices were used (Row 4, Table 2). Since small helices are generally harder to detect from both the image and the sequence, missing them from both sources appears to be more favorable than having them in only one of the two sets.

Our results clearly indicate that the two-step approach that utilizes all five secondary structure predictions is the most accurate approach from among the seven different methods. The true topology rank is the highest among the other five methods (use of SYMPRED, JPred, PSIPRED, PREDATOR, and Sable) when all of them were used for all 12 proteins in the test (Column 9, Table 2). For example, true topology was ranked 15th for 2XB5 when all five of the online secondary structure prediction methods were used to generate alternatives. The true topology of 2XB5 was ranked 40th, 44th, 40th, 75th, and 53rd, respectively, when SYMPRED, JPred, PSIPRED, PREDATOR, and Sable were used individually. We observed substantial enhancement in ranking of the true topology when multiple secondary structure predictions were used for 10 of the 12 cases. We noticed that these 10 cases are the largest 10 of the 12 cases, with their lengths ranging from 201 to 585.

3.4 The Run-Time of the Two-Step Approach

The major time in the two-step approach occurs at the mapping step in which the initial 1000 top-ranked topologies are generated. The second step is a placement step, and it can be quickly done using the dynamic programming algorithm given in this paper. The time it takes to compute the initial top-ranked topologies and the placement of those topologies are shown in Table 2, Column 10. Apparently the time is quite little. For example, to produce the top 1000 topologies and to derive the optimal placement for those topologies only takes 14.89 seconds for 3ODS, in which 16 of the 23 helices were detected from the image. The experiments in this paper were executed on a 2x Intel Xenon E5-2660 v2, 2.2GHz server machine. The factors affecting the run-time include the number of secondary structures and the quality of the skeleton. We noticed that the skeletons produced from the simulated density images are often much better than those produced from experimentally-derived images.

3.5 Two Cryo-EM Cases Involving α-β Proteins

It is generally more challenging to determine the topology for proteins with β-sheets than α-proteins. First, the detection of β-sheets is generally more challenging than the detection of helices. Second, the close spacing of β-strands makes it more challenging to identify the correct topology. We applied the two-step approach to two experimentally-derived Cryo-EM density maps, (EMDB_5030 and EMDB_1780) that were downloaded from the Electron Microscopy Data Bank (EMDB). Each density map corresponds to an atomic structure; thus, they can be used to test the accuracy of our approach. In the case of EMDB_5030, all three helices and three β-strands were detected using SSETracer and StrandTwister (Fig. 4). The true topology was ranked 47th when multiple secondary structure predictions and dynamic programming placement were used. Amazingly, the rank (47th) is even better than the rank (55^{th}) derived using true secondary structure positions on the protein sequence. In the case of EMDB_1780, the rank of the true topology is the 2^{nd} when either multiple secondary structure prediction methods or the true sequence segments of secondary structures are used. Although the two Cryo-EM proteins are smaller than most of the other proteins in the test, they are the first two cases that successfully demonstrated topology determination directly using computationally-obtained β-traces and multiple secondary structure predictions.

4 Conclusions

Due to inaccuracy in the estimation of secondary structures, the determination of topology for SSTs requires the exploration of alternatives. Effective methods are needed to explore the large solution space that results from those alternatives. We propose a dynamic programming algorithm to find the optimal placement when a topology is given. This algorithm is combined with our previous mapping algorithm and the shortest K paths algorithm to form a two-step approach. A test using 12 proteins showed that the two-step approach improves the ranking of the true topology in comparison to using single consensus prediction. We demonstrate for the first time that computationally-detected helices and β-strands from an experimentally-derived Cryo-EM density image can be combined with multiple secondary structure predictions to rank the true topology near the top of the list. Our previous methods were mostly tested using the true positions of secondary structures. In this present study, we have taken a significant step by establishing an efficient algorithm to address the increased computational cost due to the alternatives.

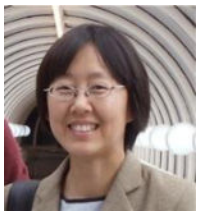
Acknowledgments

This work was partially supported by NSF DBI-1356621, NIH R01-GM062968. AB was partially supported by an Modeling & Simulation grant of Old Dominion University and JH was partially supported by FP3 fund of Old Dominion University.

Biographies

Abhishek Biswas received his Ph.D. in Computer Science from Old Dominion University in 2015. His research interests include parallel genome assembly, population genomics and modelling protein structures. He has published in peer reviewed journals like Bioinformatics, Journal of Computational Biology and Journal of Bioinformatics and Computational Biology.

Desh Ranjan is a Professor of the Computer Science Department at Old Dominion University. His research interests are in the areas of Efficient Algorithm Design, Parallel Computing, Bioinformatics and Computational Complexity. He received his PhD in computer science from Cornell University in 1992, has authored or co-authored more than 80 refereed research papers and has served PI or Co-PI for numerous peer-reviewed grants.



Jing He is an Associate Professor at the Department of Computer Science Old Dominion University. She has been particularly interested in the computational problem of deriving protein structures from the 3-dimensional image of Cryo-EM technique. She obtained PhD in Structural and Computational Biology and Molecular Biophysics from Baylor College of Medicine in 2001.

Mohammad Zubair is a Professor at the Department of Computer Science, Old Dominion University. He received his PhD in Electrical Engineering at Indian Institute of Technology Delhi in 1987. His primary area of interest is in the area of high performance computing and management of large information.

Kamal Al Nasr is an assistant professor of Computer Science at Tennessee State University, Nashville, TN. He received his Bachelor's and Master's degree in Computer Science from Yarmouk University, Jordan and another Master's degree in Computer Science from New Mexico State University. He received his Ph.D. in Computer Science from Old Dominion University in 2012. He Joined the Department of Systems and Computer Science at Howard University, Washington, D.C. as a postdoctoral research scientist in 2012. His research interest is centered on Structural Bioinformatics, 3D image analysis, Graph theory and high performance computing

Stephanie Zeil is a junior at Department of Computer Science of Old Dominion University. She has been working in research with Dr. Jing He since June of 2015.

References

1. Ludtke SJ, Baldwin PR, Chiu W. EMAN: Semi-automated software for high resolution single particle reconstructions. J Struct Biol. 1999; 128(1):82–97. [PubMed: 10600563]
2. Si, D., He, J. Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps. BCB'13: Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics; Washington, D.C. September 22–25; 2013. p. 764-70.
3. Simossis VA, Heringa J. The influence of gapped positions in multiple sequence alignments on secondary structure prediction methods. Computational Biology and Chemistry. 2004; 28(5–6):351–366. [PubMed: 15556476]
4. Chiu W, Baker ML, Jiang W, Dougherty M, Schmid MF. Electron cryomicroscopy of biological machines at subnanometer resolution. Structure. Mar; 2005 13(3):363–72. [PubMed: 15766537]
5. Hryc CF, Chen DH, Chiu W. Near-Atomic-Resolution Cryo-EM for Molecular Virology. Curr Opin Virol. Aug 1; 2011 1(2):110–117. [PubMed: 21845206]
6. Anger AM, Armache JP, Berninghausen O, Habeck M, Subklewe M, Wilson DN, Beckmann R. Structures of the human and Drosophila 80S ribosome. Nature. May 2; 2013 497(7447):80–5. [PubMed: 23636399]

7. Jiang W, Baker ML, Jakana J, Weigele PR, King J, Chiu W. Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. Nature. Feb 28; 2008 451(7182):1130–4. [PubMed: 18305544]
8. Zhang XK, Ge P, Yu XK, Brannan JM, Bi GQ, Zhang QF, Schein S, Zhou ZH. Cryo-EM structure of the mature dengue virus at 3.5-angstrom resolution. Nature Structural & Molecular Biology. 2013; 20(1):105–10.
9. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics. Jan 1; 1998 14(10):892–893. [PubMed: 9927721]
10. Wriggers W, Birmanns S. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. J Struct Biol. Feb-Mar;2001 133(2–3):193–202. [PubMed: 11472090]
11. Chan KY, Trabuco LG, Schreiner E, Schulten K. Cryo-Electron Microscopy Modeling by the Molecular Dynamics Flexible Fitting Method. Biopolymers. 2012; 97(9):678–686. [PubMed: 22696404]
12. Schröder GF, Brunger AT, Levitt M. Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. Structure (London, England: 1993). 2007; 15(12):1630–1641.
13. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc Natl Acad Sci U S A. Jan 31; 2012 109(5):1380–7. [PubMed: 22307589]
14. Zhang J, Baker ML, Schroder GF, Douglas NR, Reissmann S, Jakana J, Dougherty M, Fu CJ, Levitt M, Ludtke SJ, Frydman J, Chiu W. Mechanism of folding chamber closure in a group II chaperonin. Nature. Jan 21; 2010 463(7279):379–83. [PubMed: 20090755]
15. Al Nasr K, Liu C, Rwebangira M, Burge L, He J. Intensity-based skeletonization of CryoEM gray-scale images using a true segmentation-free algorithm. IEEE/ACM Trans Comput Biol Bioinform. Sep-Oct;2013 10(5):1289–98. [PubMed: 24384713]
16. Pettersen E, Goddard T, Huang C, Couch G, Greenblatt D, Meng E, Ferrin T. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25(13):1605–12. [PubMed: 15264254]
17. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. Apr; 2000 16(4):404–5. [PubMed: 10869041]
18. Baker ML, Ju T, Chiu W. Identification of secondary structure elements in intermediate-resolution density maps. Structure. Jan; 2007 15(1):7–19. [PubMed: 17223528]
19. Dal Palu A, He J, Pontelli E, Lu Y. Identification of Alpha-Helices from Low Resolution Protein Density Maps. Proceeding of Computational Systems Bioinformatics Conference(CSB). 2006:89–98.
20. Jiang W, Baker ML, Ludtke SJ, Chiu W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. J Mol Biol. May; 2001 308(5):1033–44. [PubMed: 11352589]
21. Kong Y, Ma J. A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. J Mol Biol. Sep 12; 2003 332(2):399–413. [PubMed: 12948490]
22. Rusu M, Wriggers W. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. J Struct Biol. Feb; 2012 177(2):410–9. [PubMed: 22155667]
23. Si D, Ji S, Nasr KA, He J. A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. Biopolymers. Sep; 2012 97(9): 698–708. [PubMed: 22696406]
24. Al Nasr K, Ranjan D, Zubair M, Chen L, He J. Sovling the secondary structure matching problem in cryo-EM de novo modeling using a constrained K-shortest path graph algorithm. IEEE/ACM Trans Comput Biol Bioinform. 2014; 11(2):419–29. [PubMed: 26355788]
25. Si, D., He, J. Orientations of beta-strand traces and near maximum twist. Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; Newport Beach, California. 2014. p. 690-694.

26. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics. 1997; 27(3):329–335.
27. Si D, He J. Tracing beta-strands using strandtwister from cryo-EM density maps at medium resolutions. Structure. 2014; 22(11):1665–76. pp. 22(11). [PubMed: 25308866]
28. Al Nasr K, Ranjan D, Zubair M, He J. Ranking valid topologies of the secondary structure elements using a constraint graph. J Bioinform Comput Biol. Jun; 2011 9(3):415–30. [PubMed: 21714133]
29. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics. Apr 15; 2005 21(8):1719–20. [PubMed: 15585524]
30. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins. Feb; 2002 46(2):197–205. [PubMed: 11807948]
31. Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics. Sep 1; 2003 19(13):1650–5. [PubMed: 12967961]
32. Baker ML, Abeysinghe SS, Schuh S, Coleman RA, Abrams A, Marsh MP, Hryc CF, Ruths T, Chiu W, Ju T. Modeling protein structure at near atomic resolutions with Gorgon. Journal of Structural Biology. 2011; 174(2):360–373. [PubMed: 21296162]
33. Biswas A, Si D, Al Nasr K, Ranjan D, Zubair M, He J. Improved efficiency in cryo-EM secondary structure topology determination from inaccurate data. J Bioinform Comput Biol. 2012; 10(3): 1242006. [PubMed: 22809382]
34. Biswas, A., Ranjan, D., Zubair, M., He, J. A Novel Computational Method for Deriving Protein Secondary Structure Topologies Using Cryo-EM Density Maps and Multiple Secondary Structure Predictions. In: Harrison, R.Li, Y., Măndoiu, I., editors. Bioinformatics Research and Applications. Springer International Publishing; 2015. p. 60-71.Lecture Notes in Computer Science
35. Biswas A, Ranjan D, Zubair M, He J. A Dynamic Programming Algorithm for Finding the Optimal Placement of a Secondary Structure Topology in Cryo-EM Data. Journal of Computational Biology. 2015; 22(9):837–843. 2015/09/01. [PubMed: 26244416]
36. McKnight A, Si D, Al Nasr K, Chernikov A, Chrisochoides N, He J. Estimating loop length from CryoEM images at medium resolutions. BMC Structural Biology. 2013; 13(suppl 1):S5. [PubMed: 24565041]
37. Abeysinghe S, Ju T, Baker ML, Chiu W. Shape modeling and matching in identifying 3D protein structures. Computer Aided-design. 2008; 40:708–20.
38. Lindert S, Staritzbichler R, Wötzel N, Karakas M, Stewart PL, Meiler J. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure. Jul 15; 2009 17(7):990–1003. [PubMed: 19604479]
39. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins: Structure, Function, and Bioinformatics. 2005; 59(3):467–475.
40. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, van Ginkel G, Devkota B, Lagerstedt I, Ludtke SJ, Newman RH, Oldfield TJ, Rees I, Sahni G, Sala R, Velankar S, Warren J, Westbrook JD, Henrick K, Kleywegt GJ, Berman HM, Chiu W. EMDataBank.org: unified data resource for CryoEM. Nucleic Acids Res. Jan.2011 39:D456–64. no. Database issue. [PubMed: 20935055]

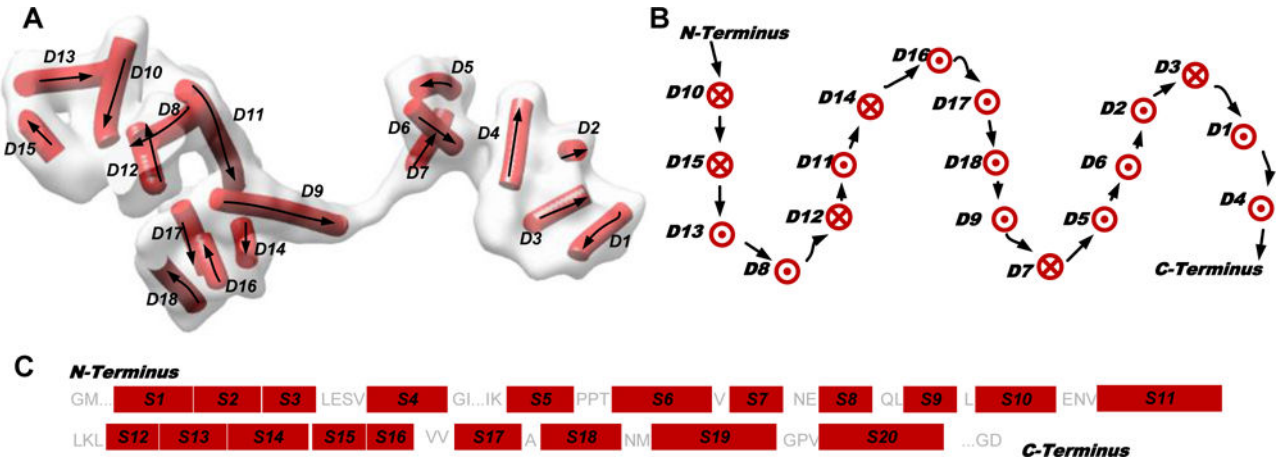


Fig. 1.

Secondary structures and topology. (A) The density map (gray) was simulated to 10 Å resolution using the atomic structure of protein 3HJL (Protein Data Bank (PDB) ID) and EMAN software [1]. The secondary structure traces of helices (red sticks) were detected using SSETracer [2] and viewed using Chimera [16]. Arrows: the direction of the protein sequence; (B) The true topology of SSTs (arrows, crosses, and dots indicate the direction of the protein sequence); (C) An illustration of the amino acid sequence of protein 3HJL annotated with the location of α-helices (red rectangles) based on the structure. Loops longer than four amino acids are indicated using "…".

Fig. 2.

Secondary structure predictions from multiple servers. The amino acid sequence of protein 2XVV (PDB ID) is labeled at the outermost circle. The positions of helices are shown as red rectangles from outer to inner circles as the true position of the secondary structures obtained from PDB, using SYMPRED [3], JPred [9], PSIPRED [17], and PREDATOR [26] prediction methods, respectively. The α-traces (blue lines) detected from the density map of 2XVV using SSETracer are shown in the center.

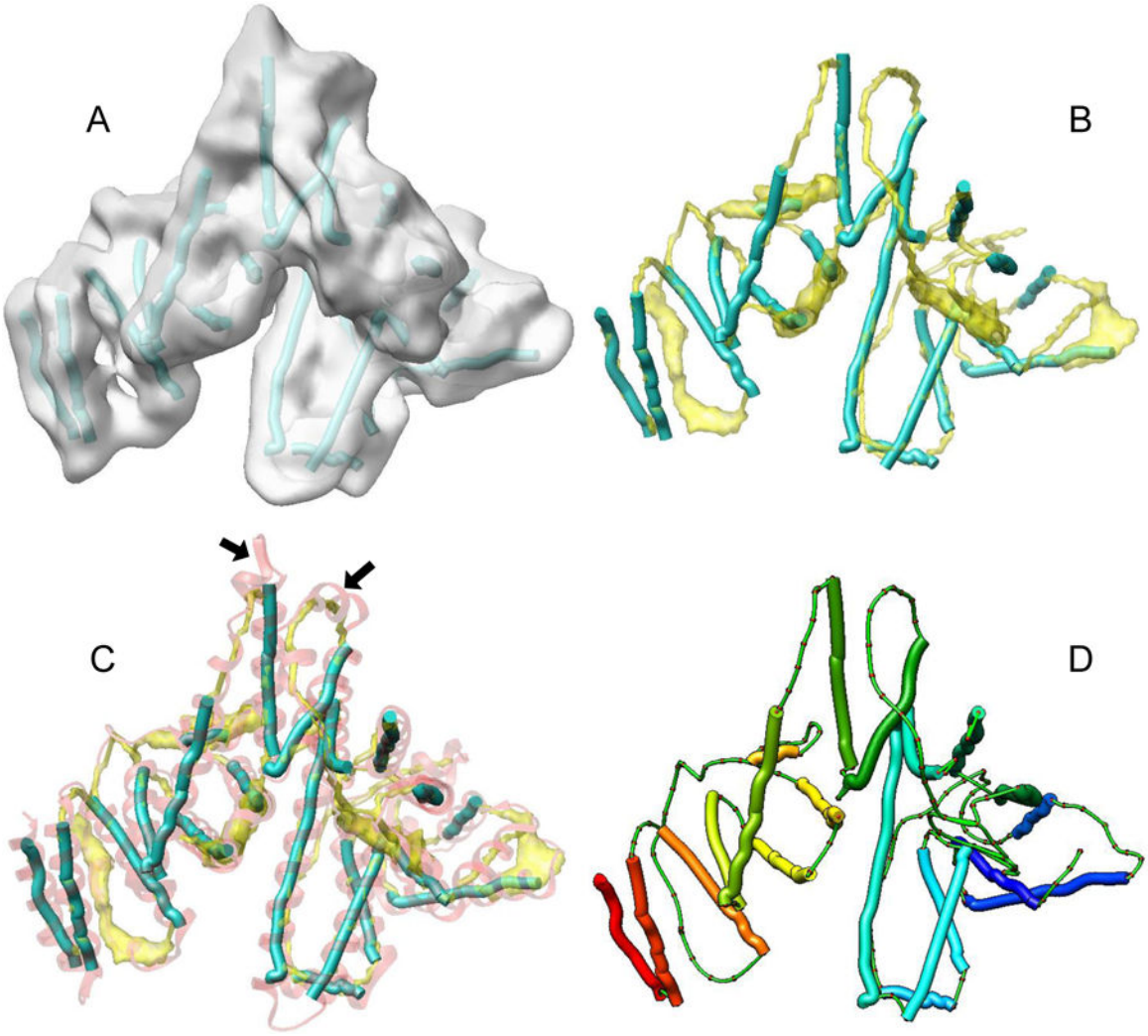


Fig. 3.

The true topology derived from the two-step approach for 2XVV (PDB ID). (A) The 3D image (gray) and the SSTs (blue sticks) detected using SSETracer [2]; (B) Skeletons (yellow density) derived from the Cryo-EM density map using SkelEM [15] and the SSTs; (C) The atomic structure (pink ribbon) superimposed on the SST elements and the skeleton. Examples of missed helices in the detection are shown (arrows); (D) The true topology computed by the two-step approach (shown in multiple colors from the blue/N-terminal to the red/C-terminal) is ranked 8th. The connecting traces were identified from the skeleton using DP-TOSS [24]

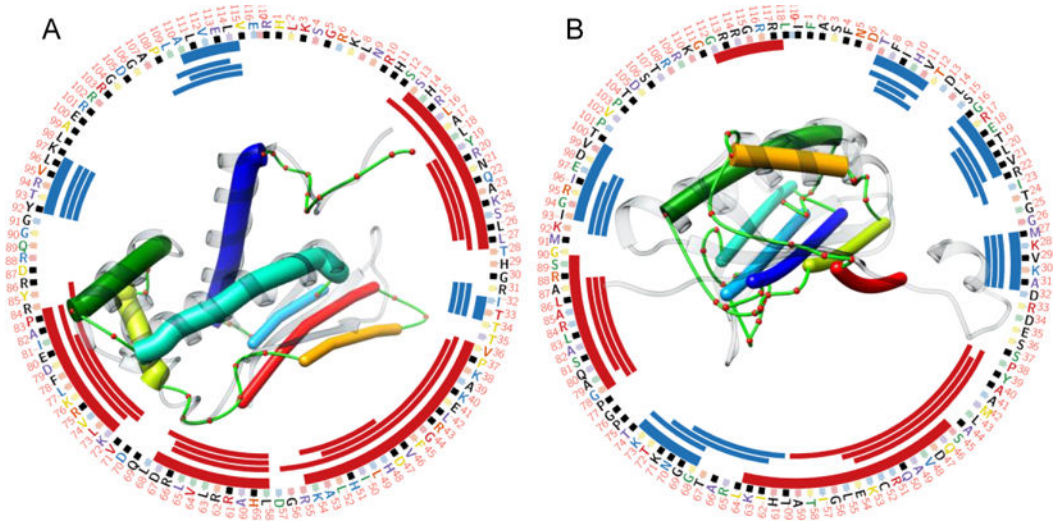


Fig. 4.

Topologies derived from the Cryo-EM density maps. Secondary structure positions derived from the true structure and those predicted using SYMPRED [3], JPred [9], and PSIPRED [17] are shown from the outer circles to the inner circles for protein 4V68_BR(PDB ID) in (A) and 3IZ6_K(PDB ID) in (B). The α-traces (thicker sticks) and the β-traces (thinner sticks) were detected from the experimentally-derived Cryo-EM map EMDB_5030 in (A) and EMDB_1780 in (B) using SSETracer [2] and StrandTwister [27]. The true topology (shown in rainbow colors from blue/N-terminal to red/C-terminal) is ranked 47th for EMDB_5030 and 2nd for EMDB_1780. The connecting trace was identified from the skeleton using DP-TOSS. The true structure (ribbon) is superimposed for each.

Table 1

The detection of SSTs for α-helix proteins.

PDB ID[a]	#a.a.[b]	#α-helices[c]	#SS(α-hlx)[d]	#α-stk[e]	% specificity[f]	% sensitivity[g]
1FLP	142	7	8	7	72.09	93.94
1NG6	148	7	7	7	100.00	86.92
2OVJ	201	12	9	9	84.91	83.11
3LTJ	201	12	13	12	90.48	85.29
2XB5	207	13	14	11	66.67	91.07
1HG5	289	11	11	9	91.55	85.94
3ACW	293	17	17	16	67.39	92.44
3HJL	329	20	20	18	94.44	79.30
1Z1L	345	24	17	15	60.53	80.92
1HZ4	373	19	18	16	94.59	77.81
3ODS	415	23	27	16	81.63	87.08
2XVV	585	34	28	19	82.65	74.17
				Average	82.24	84.83

[a] The PDB ID of the protein.

[b] The number of amino acids in the protein.

[c] The number of helices in the protein structure.

[d] The maximum number of helices predicted using SYMPRED or JPred.

[e] The number of α-traces detected from the 3D image.

[f] Specificity of helix detection from the 3D image; specificity=1−fp/(Total−TotalHlx); fp: the number of wrongly detected helix C_α atoms; Total: the number of C_α atoms; TotalHlx: total number of helix C_α atoms.

[g] Sensitivity of helix detection from the 3D image. Sensitivity=TpHlx/TotalHlx. TpHlx: the number of true positive helix C_α atoms.

Table 2

True topology rankings for different secondary structure prediction methods and the two-step approach.

		True Topology Ranking							
PID[a]	**#a.a.**[b]	**True**[c]	**Pred0**[d]	**Pred1**[e]	**Pred2**[f]	**Pred3**[g]	**Pred4**[h]	**DynP**[i]	**Time**[j]
1FLP	142	2	2	2	2	2	2	1	2.09
1NG6	148	3	1	3	3	2	1	1	4.26
2OVJ	201	2	46	3	14	57	14	1	8.22
3LTJ	201	9	9	9	9	9	6	4	10.56
2XB5	207	28	40	44	40	75	53	15	7.65
1HG5	289	244	189	106	106	251	174	81	7.08
3ACW	293	1	382	97	103	68	72	68	8.19
3HJL	329	4	15	18	3	66	7	3	529.3
1Z1L	345	5	92	64	824	96	54	12	8.88
1HZ4	373	311	3	2	39	1	1	1	897.2
3ODS	415	12	198	488	168	266	716	22	14.89
2XVV	585	4	88	124	196	174	118	8	1056.6

[a] The PDB ID with chain.

[b] The number of amino acids in the protein.

[c] The the true topology ranking using the true sequence position of SSEs.

[d] The true topology ranking using predicted sequence positions from SYMPRED [3].

[e] The true topology ranking using predicted sequence positions from JPred [9].

[f] The true topology ranking using predicted sequence positions from PSIPRED [17].

[g] The true topology ranking using predicted SSE sequence from PREDATOR [26].

[h] The true topology ranking using predicted SSE sequence from Sable [39].

[i] The true topology ranking using multiple secondary structure predictions with dynamic programming algorithm for optimal placement.

[j] The run-time (in seconds) of the two-step approach. This includes the time to generate the top 1000 topologies and the total time to find the optimal placement for the top 1000 topologies.

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2017 June 10.