Towards De Novo Folding of Protein Structures from Cryo-EM 3D Images at Medium Resolutions

Jing He[*] and Dong Si

Department of Computer Science
Old Dominion University
Norfolk, VA
[*]jhe@cs.odu.edu

***Abstract* - Electron Cryo-microscopy (cryo-EM) is an important biophysical technique that produces 3-dimensional (3D) images at different resolutions. De novo folding is becoming a promising approach to derive the backbone of proteins from the cryo-EM 3D images at medium resolutions. We summarize our findings in major steps of de novo folding and the challenges from the inaccurate data.**

Secondary structure, protein, graph, electron microscopy, algorithm, Sampling, image

I. Introduction

Deriving atomic protein structures from 3-dimensional (3D) images, also called density maps, obtained using electron cryo-microscopy (cryo-EM) technique is a challenging problem. When the resolution of a 3D image is better than 4Å, the backbone of protein can be resolved and the structure can be derived [1]. When the resolution is lower than 5Å, amino acid features are not resolved and it is more challenging to derive structures from such images. The major approach available is to use an existing structure that is homologous to the target protein [2]. The homologous structure will be modified during fitting of the image. Although this approach has been successful in deriving many structures, it relies on a template structure. It is still challenging to find a suitable template for many proteins. *De novo* folding aims to derive atomic protein structures from 3D images directly, applicable when a template structure is not available. It has been demonstrated as a promising method to derive the backbone from 3D images at about 4Å resolution [5]. At medium resolutions (5-10Å), major secondary structure elements such as α-helices and β-sheets can be detected in such 3D images [6-11]. Figure 1A shows an example of a 3D image, extracted from an experimentally-obtained cryo-EM density map. The location of helices (red lines) and β-sheet (purple density) was detected using *SSETracer* [3].

Although it is possible to detect the location of major β-sheets, it has been a challenging problem to detect β-strands from the β-sheet density. The spacing of β-strands is between 4.5Å and 5Å, which makes them almost impossible to be resolved in a 3D image with 5-10Å resolution. We will summarize our recent work tracing β-strands from β-sheet density in this paper.

α-helices and β-strands can be detected as lines/traces (as in Figure 1B), but their order along the protein sequence needs to be determined. Topology of SSTs refers to the order of the SSTs and the direction of each SST (Figure 1C). The topology of SSTs can be inferred by combining two sources of information about secondary structures, one (Figure 1B) from the 3D image and the other (Figure 1D) from the protein sequence [12-16]. In addition to the position of secondary structures, skeleton that represents possible connection among the secondary structures can be extracted from the 3D image [17-19] . It provides constraints about how secondary structures are possibly connected. The length along the skeleton between two SSTs is an important constraint and is often compared with the corresponding loop length on the protein sequence to evaluate a topology. Additional constraints can also be included in the evaluation.

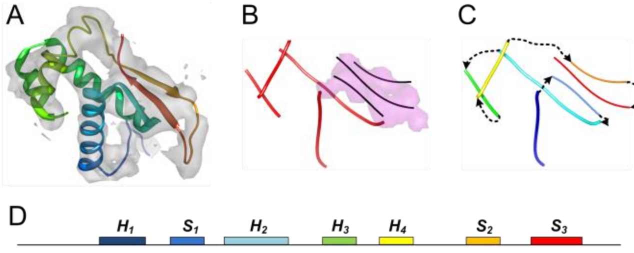

Figure 1: Secondary structure traces (SST) detected from a 3D image and the topology of secondary structures. (A) The 3D image extracted from cryo-EM map (EMD-5030, 6.4Å resolution) is superimposed with the corresponding PDB (3FIN_R) structure. (B) The traces of α-helices (red lines) and the region of β-sheet (purple) were detected using *SSETracer* [3]. The traces of β-strands (black lines) were detected using *StrandTwister* [4]. (C) The 2nd ranked topology using *DP-TOSS* is the true topology. The order and the direction are indicated using arrows, dashed lines and rainbow color. (D) Illustration of the sequence segments of secondary structures obtained from the sequence of 3FIN_R (PDB ID).

Even in an ideal situation in which SSTs are accurately detected from the 3D image and secondary structures are accurately predicted from the protein sequence, the determination of the topology is NP-hard [15]. The nature of the problem makes earlier enumeration approaches [12, 20, 21] difficult to apply towards large proteins. A dynamic programming approach was developed to find the optimal topology as the constrained shortest path in $O(N^2 2^N)$ time [22]. We later demonstrated that DP-TOSS, a constrained *K*-shortest path method, is effective finding the topology from large proteins such as those with 33 helices [14]. In reality, secondary structures are often not accurately identified, neither from the 3D image, nor from the protein sequence. The accuracy of secondary structure prediction from a protein sequence is often between 70% and 80% [23]. Current graph approaches face challenges in developing effective algorithms to take consideration of potential errors in the data. Monte Carlo approach [24] creates a new topology using operation add/delete, swap, flip, shrink/grow to sample the solution space. It is based on Monte Carlo Metropolis algorithm that randomly samples the solution space, but it is not clear if it is effective for this problem. The topology determination problem has two characters. (1) It is a highly combinatorial problem. All combinations of orders and directions may need to be sampled. (2). Different operations have different significance to the problem. For example, a shift to a helix will affect the topology less than a flip of the direction to the helix. More effective algorithms are needed to sample the possible topologies of the secondary structures. Once limited possible topologies are predicted, atomic models can be built for each topology [21, 24-26]. The models will be further selected using energy functions.

Table 1: The Identified Secondary Structures from the Experimental Cryo-EM Density Maps

EMD_PDB ID, Resolution (Å)	H< =8[a]	H> 8[b]	AA. H[c]	Sht[d]	Str[e]	AA. S[f]
1237_2GSY_A, 7.2	4/8	3/3	51/69	6/6	22/24	180/187
1733_3C91_H, 6.8	0/0	5/5	70/86	2/3	10/12	51/62
1740_3C92_A, 6.8	0/1	5/6	92/101	2/2	10/10	53/58
1780_3IZ6_K, 5.5	0/1	2/2	27/37	1/1	4/5	25/29
5030_3FIN_R, 6.4	0/0	4/4	57/59	1/1	3/3	12/14
Totals	4/10	19/20	297/352	12/13	49/54	321/350

a. The number of detected helices / the number of observed helices that have less than 8 amino acids on each;

b. The number of detected helices / the number of observed helices that have more than 8 amino acids on each;

c. The total number of detected amino acids / the total number of observed amino acids in helices;

d. The number of detected sheets / the number of observed β-sheets;

e. The number of detected strands in the best of top ten sets/ the number of observed strands;

f. The total number of detected amino acids in the best of the top ten sets / the total number of observed amino acids in sheets.

II. Traces of beta-strands

The location of secondary structures such as helices and β-strands are critical in de novo folding of protein structures from 3D images at medium resolutions. However, β-strands are often not visible in cryo-EM density maps at medium resolutions, and therefore it is extremely hard to detect them even after β-sheets are detected. We recently developed a method, *StrandTwister* [4], that predicts the location of β-strands from cryo-EM density maps at medium resolutions. The idea of *StrandTwister* is to analyze the twist of the β-sheet. The foundation of the algorithm is the discovery of the relationship between the orientation of β-strands and the twist in the β-sheet. As an example, the locations of four helices in the image were detected using *SSETracer* (red lines in Figure 1B) as α-traces, representing the central lines of the helices. Three β-strands were detected using *StrandTwister* as β-traces that represent the central lines of β-strands (black lines in Figure 1B). Table 1 shows the accuracy of the detected locations of helices, β-sheets and β-traces using five experimentally-obtained cryo-EM density maps at 5.5Å-7.2Å resolutions. *SSETracer* was applied to detect the location of helices and β-sheets. It was able to detect most of the helices longer than three turns, but it missed most of the short helices. Twelve of the thirteen β-sheets were detected using *SSETracer*, among which eleven are β-sheets with more than two strands each. 2-stranded β-sheets are often hard to detect since they resemble helices in the images. *StrandTwister* uses the detected β-sheet density region as input and predicts possible sets of β-traces. The best of top ten predicted sets of β-traces are evaluated in Table 1 column 6 and 7. When the Cα atom of an amino acid on the secondary structure is within 2.5Å from the detected line, we consider it as a detected amino acid. It appears that most β-strands can be traced for the major β-sheets detected from the image.

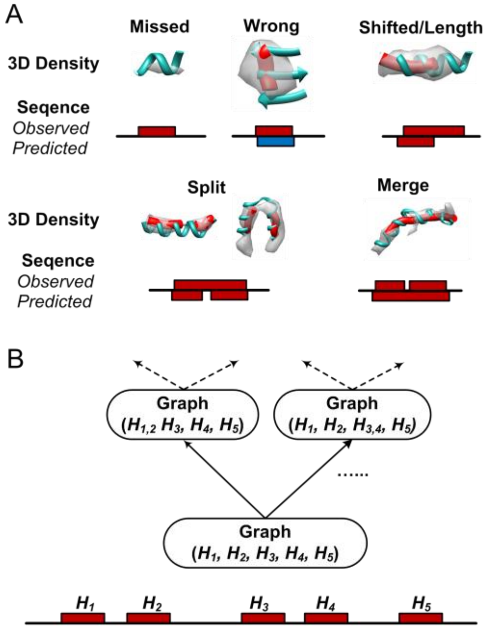

Figure 2: Secondary structure errors. (A) The positions of secondary structures detected from 3D images are shown as red sticks, and the true structures are shown as blue ribbons. An observed helix on protein sequence is shown as a rectangle (red) above the line, and the predicted sequence fragment is shown as a red (helix) or blue (β-strand) rectangle below the line. (B) A possible graph hierarchy in topology determination for potential split errors.

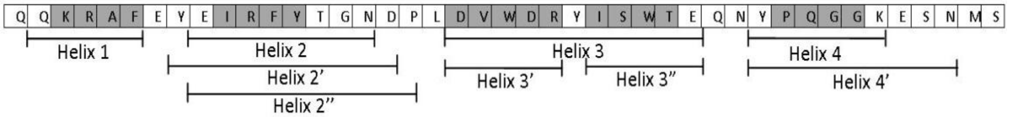
Q Q K R A F E Y E I R F Y T G N D P L D V W D R Y I S W T E Q N Y P Q G G K E S N M S
Helix 1
Helix 2
Helix 2’
Helix 2”
Helix 3
Helix 3’
Helix 3”
Helix 4
Helix 4’

Figure 3: Prediction errors for helices. The observed position of the helices (Helix 1, …, Helix 4), the predicted helices (Helix 2’, Helix 3’, Helix 3”, Helix 4’, Helix 2”) from multiple prediction methods, and the consensus prediction of the helices (shaded) are illustrated.

III. Topology of Secondary Structure Traces using Computationally Detected Beta-Traces

A protein sequence has a direction from N-terminal to C-terminal. When the protein sequence is threaded in the 3D image, it visits the detected α-traces and β-traces in a unique order. For medium resolution cryo-EM density maps, current approaches in topology determination either restrict in α-proteins that do not have β-sheets or use the true positions of β-strands from the PDB structure, due to the lack of computational methods to detect β-strands. We report preliminary data here to derive the topology using the α-traces and β-traces detected the 3D image extracted from cryo-EM map (EMD-5030). The α-traces and β-traces are fairly accurately detected (Figure 1B and Table 1 row 5). A topology graph was built by matching the detected traces with the sequence segments of the secondary structures. The correct topology (Figure 1C) was ranked as the 2nd on the list using DP-TOSS [14], a dynamic programming approach searching for the shortest K paths in the graph. The true sequence fragments of the secondary structures were used in this test. This result suggests that it is possible to derive the correct topology from an α/β protein when secondary structures are traced fairly accurately.

IV. Dynamic graph for inaccuracy in Secondary Structures

One of the major challenges in topology determination arises from the errors in secondary structure detection from the image and in secondary structure prediction from the protein sequence (Figure 2A). Short helices and small β-sheets are often missed in the detection obtained from the 3D image. The detected positions may be shifted from the true positions and the detected length may be shorter or longer than that of the true secondary structure. Similar types of errors exist in the secondary structure prediction from the protein sequence.

The nature of each error is different, and the frequency of each kind of errors is different too. Ideally different kinds of error may be handled differently. For example, split-error refers to the situation in which one helix is detected/predicted as two shorter helices. We observed that a split error is not as popular as a shift error. A potential solution is to consider both situations, one containing two separate secondary structures and the other containing one longer secondary structure. For example, three graphs may be considered if helix H_1 and H_2 are close enough and H_3 and H_4 are also close enough (Figure 2B). However, effective algorithms are needed to reduce the repetitive computation as much as possible.

For simplicity in discussion of a dynamic graph, we assume there are no β-sheets in the protein. Ideally, the secondary structures are predicted accurately from the sequence, and we let $H = (H_1, H_2, \dots, H_N)$ be an ordered list of sequence segments that form helices (Figure 1D). Let $S = \{S_1, S_2, \dots, S_M\}$ be a set of helix traces accurately detected from the density map. Then the topology of SSTs can be inferred from graph $G =< H, S >$. In reality however, shift error may exist in the predicted sequence segments, and therefore a few alternative positions may be used for each helix. Let a tuple $H_i = \left(H^1_{i_1}, H^2_{i_2}, \dots, H^N_{i_N}\right)$ where $1 \le i_1 \le B, \dots, 1 \le i_N \le B$ represent an ordered list of sequence segments. B is the maximum number of alternatives allowed. An example of a tuple (Figure 3) is (Helix 1, Helix 2’, Helix3’, Helix 3”, Helix 4’). Similarly for 3D image, let $S_j = \left(S^1_{j_1}, S^2_{j_2}, \dots, H^M_{j_M}\right) where\ 1 \le j_1 \le BB, \dots, 1 \le j_M \le BB.$ BB is the maximum number of alternatives allowed for each secondary structures. Each tuple of sequence segments will be used to match with each tuple of helix sticks/traces detected from the image using a graph $G =< H_{i,}S_j >$. Realizing that most of the segments may remain the same between two tuples, we previously developed a dynamic graph method to reduce the computation in transforming one graph to another [27]. The approach only updates the tables associated with nodes as necessary during dynamic programming process.

Seven α-proteins were randomly chosen from the PDB and their 3D images were simulated to 10Å resolution using EMAN [28]. The only requirement imposed is to choose proteins with medium to large sizes. The proteins contain 164 to 501 amino acids with 5 to 26 helices. The positions of helices were detected using *SSELearner* [11]. As an example, it detected twenty-one of twenty-six helices from the 3D image of protein 2X79. The sequence of the protein was submitted to five prediction servers including SYMPRED [29], JPRED [30], PSIPRED [31], PREDATOR [32] and SABLE [33]. Fourteen of the twenty-six helices were detected at approximately correct locations of helices. After pre-processing the predicted sequence segments, thirty-two tuples were generated (for details see [27]), each of which was matched to the twenty-one α-traces in the 3D image. Instead of computing the dynamic programming tables for thirty-two graphs from scratch, our dynamic graph method generates a new graph with reduced update from a previous graph. In this case, finding the optimal solution took 72.8% of the time if we were to compute all the thirty-two graphs naively from scratch. It appears that the dynamic graph method cut down about 35% of time from the naïve method.

Table 2: Improved run-time using dynamic graph update in topology determination

PDB ID	#AA[a]	#P./O. Hlcs[b]	#St.[c]	Hlx Tup.[d]	All Comb.[e]	Naive Time[f]	%[g]
3FYQ	199	6/5	5	12	0.51	0.84	60.7
2WVI	164	8/9	7	2	0.76	0.92	82.6
1NG6	148	7/9	8	8	2.3	4.8	47.9
3HFW	357	15/23	16	2	350.84	400.0	87.5
2X79	501	14/26	21	32	20.5k	28.2k	72.8
1TBF	347	16/22	16	24	6.8k	15.0k	45.6
3L6A	364	18/25	19	16	13.3k	21.8k	60.8
Average							65.45

a: The number of amino acids in the structure (PDB file);
b: The number of helix regions in secondary structure predictions/The number of observed helices in PDB file;
c: The number of sticks (α-traces) detected from the 3D image;
d: The number of tuples of helix sequence segments generated after pre-processing;
e: The time (in seconds) to update the graph for all helix tuples and to find the shortest path;
f: Brute force time to re-compute the entire graph for all the tuples and to search for the shortest path;
g: Percentage of the total time for dynamic update g=e/f.

V. SUMMARY

Topology determination is a critical step in de novo folding for cryo-EM density maps at medium resolutions. This paper summarizes our recent advances in major steps of this problem (1) the detection of β-strands using *StrandTwister* about which details are under review in a separate paper; (2) the prediction of topologies using DP-TOSS. Many methods have been developed to detect the location of helices from cryo-EM density maps. However, there has not been any tool to detect β-strands from cryo-EM density maps at medium resolutions. We find that *StrandTwister* is able to detect most of β-strands from major β-sheets in such density maps. Using computationally detected α-traces and β-traces from the image and the exact sequence segments of the secondary structures, *DP-TOSS* was able to rank the true topology as the 2nd in a test using the 3D image extracted from cryo-EM density map (EMD-5030). This result shows, for the first time, that it is possible to rank the true topology high in the list using computationally detected α-traces and β-traces from experimentally-obtained density maps at a medium resolution. In spite of recent advances in de novo folding, errors in the secondary structure detection present a challenging problem. Effective algorithms are needed to derive the topology in presence such errors. Our exploration of the shift error shows that dynamic graph is an option cutting down the computation about 35%. However, more effective methods are needed.

ACKNOWLEDGMENT

This work is supported in part by the startup fund, MSF fund, and the Modeling & Simulation Fellowship at ODU and NSF-1356621. We thank Lin Chen for the assistance in using DP-TOSS.

REFERENCES

1. Zhang, X.K., et al., *Cryo-EM structure of the mature dengue virus at 3.5-angstrom resolution.* Nature Structural & Molecular Biology, 2013. **20**(1): p. 105-10.
2. Beck, F., et al., *Near-atomic resolution structural model of the yeast 26S proteasome.* Proc Natl Acad Sci U S A, 2012. **109**(37): p. 14870-5.
3. Si, D. and J. He, *Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps.* Proceeding of ACM Conference on Bioinformatics, Computational Biology and Biomedicine Workshop, 2013: p. 765-771.
4. Si, D. and J. He, *Tracing Beta-strands using StrandTwister from Cryo-EM Density Maps at Medium Resolutions.* Structure, (under revision).
5. Ludtke, S.J., et al., *De novo backbone trace of GroEL from single particle electron cryomicroscopy.* Structure, 2008. **16**(3): p. 441-8.
6. Jiang, W., et al., *Bridging the information gap: computational tools for intermediate resolution structure interpretation.* J Mol Biol, 2001. **308**(5): p. 1033-44.
7. Del Palu, A., et al., *Identification of Alpha-Helices from Low Resolution Protein Density Maps.* Proceeding of Computational Systems Bioinformatics Conference(CSB), 2006: p. 89-98.
8. Baker, M.L., T. Ju, and W. Chiu, *Identification of secondary structure elements in intermediate-resolution density maps.* Structure, 2007. **15**(1): p. 7-19.
9. Kong, Y., et al., *A Structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps.* J Mol Biol, 2004. **339**(1): p. 117-30.
10. Zeyun, Y. and C. Bajaj, *Computational Approaches for Automatic Structural Analysis of Large Biomolecular Complexes.* Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2008. **5**(4): p. 568-582.
11. Si, D., et al., *A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps.* Biopolymers, 2012. **97**(9): p. 698-708.
12. He, J., Y. Lu, and E. Pontelli, *A Parallel Algorithm for Helix Mapping between 3-D and 1-D Protein Structure using the Length Constraints.* Lecture Notes in Computer Science, 2004. **3358**: p. 746-756.
13. Abeysinghe, S. and T. Ju. *Shape modeling and matching in identifying protein structure from low resolution images*. in *Proceedings of the 2007 ACM symposium on Solid and physical modeling* 2007. Beijing, China.

14. Al Nasr, K., et al., *Solving the secondary structure matching problem in de novo modeling using a constrained K-shortest path graph algorithm.* IEEE Transaction of Computational Biology and Bioinformatics, 2014. **11**(2): p. 419-30.
15. Al Nasr, K., et al., *Ranking Valid Topologies of the Secondary Structure Elements Using a Constraint Graph.* Journal of Bioinformatics and Computational Biology, 2011. **09**(03): p. 415-430.
16. Lu, Y., J. He, and C.E. Strauss, *Deriving topology and sequence alignment for the helix skeleton in low-resolution protein density maps.* J Bioinform Comput Biol, 2008. **6**(1): p. 183-201.
17. Ju, T., M.L. Baker, and W. Chiu, *Computing a family of skeletons of volumetric models for shape description.* Comput Aided Des, 2007. **39**(5): p. 352-360.
18. Baker, M.L., et al., *Modeling protein structure at near atomic resolutions with Gorgon.* Journal of Structural Biology, 2011. **174**(2): p. 360-373.
19. Al Nasr, K., et al., *Intensity-Based Skeletonization of CryoEM Grayscale Images Using a True Segmentation-Free Algorithm.* IEEE/ACM Trans Comput Biol Bioinform, 2013. **10**(5): p. 1289-98.
20. Wu, Y., et al., *Determining protein topology from skeletons of secondary structures.* J Mol Biol, 2005. **350**(3): p. 571-86.
21. Sun, W. and J. He, *Native secondary structure topology has near minimum contact energy among all possible geometrically constrained topologies.* Proteins, 2009. **77**(1): p. 159-73.
22. Al Nasr, K., et al., *Ranking valid topologies of the secondary structure elements using a constraint graph.* J Bioinform Comput Biol. **9**(3): p. 415-30.
23. Dor, O. and Y. Zhou, *Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training.* Proteins: Structure, Function, and Bioinformatics, 2007. **66**: p. 838-845.
24. Lindert, S., et al., *EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps.* Structure, 2009. **17**(7): p. 990-1003.
25. Al Nasr, K., W. Sun, and J. He, *Structure prediction for the helical skeletons detected from the low resolution protein density map.* BMC Bioinformatics, 2010. **11 Suppl 1**: p. S44.
26. Al Nasr, K., et al., *Building the initial chain of the proteins through de novo modeling of the cryo-electron microscopy volume data at the medium resolutions*, in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*2012, ACM: Orlando, Florida. p. 490-497.
27. Biswas, A., et al., *Improved efficiency in cryo-EM secondary structure topology determination from inaccurate data.* J Bioinform Comput Biol, 2012. **10**(3): p. 1242006.
28. Ludtke, S.J., P.R. Baldwin, and W. Chiu, *EMAN: semiautomated software for high-resolution single-particle reconstructions.* J Struct Biol, 1999. **128**(1): p. 82-97.
29. Simossis, V.A. and J. Heringa, *The influence of gapped positions in multiple sequence alignments on secondary structure prediction methods.* Computational Biology and Chemistry, 2004. **28**(5-6): p. 351-366.
30. Cuff, J.A., et al., *JPred: a consensus secondary structure prediction server.* Bioinformatics, 1998. **14**(10): p. 892-893.
31. Bryson, K., et al., *Protein structure prediction servers at University College London.* Nucleic Acids Res, 2005. **33(**Web Server issue): p. W36-8.
32. Frishman, D. and P. Argos, *Seventy-five percent accuracy in protein secondary structure prediction.* Proteins: Structure, Function, and Bioinformatics, 1997. **27**(3): p. 329-335.
33. Adamczak, R., A. Porollo, and J. Meller, *Combining prediction of secondary structure and solvent accessibility in proteins.* Proteins: Structure, Function, and Bioinformatics, 2005. **59**(3): p. 467-475.