Combine Cryo-EM Density Map and Residue Contact for Protein Secondary Structure Topologies

Maytha Alshammari[1], Jing He[1]

[1]*Department of Computer Science, Old Dominion University, Norfolk, VA*

Abstract

Although atomic structures have been determined directly from cryo-EM density maps with high resolutions, current structure determination methods for medium resolution (5 to 10 Å) cryo-EM maps are limited by the availability of structure templates. Secondary structure traces are lines detected from a cryo-EM density map for a-helices and β-strands of a protein. A topology of secondary structures defines the mapping between a set of sequence segments in 1D and a set of traces of secondary structures in 3D. In order to enhance the accuracy in ranking secondary structure topologies, we propose a method that combines three sources of information – a set of sequence segments in 1D, a set of amino acid contact pairs in 2D, and a set of traces in 3D at the secondary structure level. A test of seven cases show that a small set of secondary structure topologies can be produced to include the true topology when the three sources of information are used, even when errors exist in one or more of the three sources of information. The use of amino acid contact information improves the ranking of the true topology in six of the seven cases in the test.

Introduction

As of March 2021, there are 5342 atomic structures in Protein Data Bank (PDB), for which electron density maps with better than 5Å resolution were obtained using cryo-EM technique. For density maps with better than 5Å resolution, the backbone of a protein chain is often distinguishable, and hence the atomic structure can be derived. For a density map with lower than 5Å resolution, it is challenging to derive the atomic structure from the density map, since molecular details are less resolved. Currently there are only 1056 structures in PDB, which are derived from density maps with medium resolution (5-10Å). Since molecular details are not sufficient to determine atomic structures for most medium-resolution density maps, template-based methods are mainly used to derive atomic structures from such maps. When no suitable template structures are available, such as for a new fold, matching secondary structures that are detected from the density map with those predicted from the sequence of the protein is a promising direction to derive the arrangement of secondary structures of the protein in 3-dimensional space (3D). The relative positioning of secondary structures in 3D provides a foundation to derive the tertiary structure of a protein. Protein secondary structures, four helices and one β-sheet region, were identified using a secondary structure detection method that uses convolutional neural networks (CNN)(Figure 1 (B)). A segmented helix region can be represented using the central line (also referred as a-trace) of the region. A segmented β-sheet region can be represented using a set of lines (also referred as β-traces) for β-strands using StrandTwister. In principle, it is possible to use a set of lines to represent the orientation and position of major helices and β-strands in the cryo-EM density map with medium resolution. As an example, seven secondary structure traces were detected and labeled from L0 to L6 (Figure 1). Four of them represent four helices (red), and three represent β-strands in the β-sheet (blue in Figure 1). The secondary structure traces show the relative geometric relationship among secondary structures, although such information needs to be linked with the sequence of amino acids to derive the tertiary structure of a protein. Mapping secondary structure traces to segments of amino acid sequence is referred as the process of finding the topology of secondary structures. Given N secondary structure traces detected from a cryo-EM density map, and M secondary structure segments from the protein sequence, a topology describes the order of the N traces and the direction of each trace with respect to the direction of the protein sequence. In this paper, we show the potential of combining three pieces of information: 3-dimensional location of secondary structures detected from the density map, sequence segments of secondary structures predicted from the protein sequence, and amino acid contact pairs predicted from the protein sequence in deriving protein structures for medium resolution cryo-EM density maps.

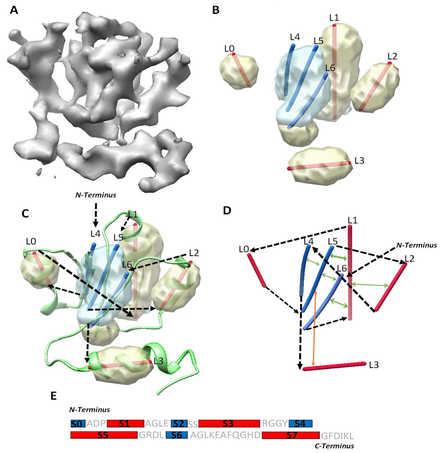

Figure 1. Secondary structures, topology, and contact. (A) The cryo-EM density map (gray, EMDB ID 6810). (B) The detected secondary structure of a-helices (yellow density) and β-sheet (blue density) using DeepSSETracer and the traces of a-helices (red lines) and β-strands (blue lines) predicted using StrandTwister. (C) An example of a correct topology. The black arrows indicate order of the true topology from N to C terminal. The Green ribbon is the atomic structure of 5y5x chain H. (D) An example of a wrong topology. The black arrows indicate order of the true topology. Green arrows indicate correctly predicted secondary structure contact pairs. Orange arrows indicate wrongly predicted secondary structure contact. (E) An illustration of the amino acid sequence of protein 5y5x chain H annotated with the location of helices (red rectangles) and β-strands (blue rectangles) predicted using JPred.

Method

1-Protein secondary structure contact: Amino acid contact prediction was performed using DNCON2, which is a tool of MULTICOM software for the six cases involving cryo-EM density maps and RaptorX for the two targets of CASP. In order to extract significant long-range contacts, screening was conducted to 1) remove all pairs with near zero p-values; 2) remove short-range pairs with less than or equal to 3 amino acids separating them; 3) extract those pairs that have p-values larger than three standard deviation of the p-values of the protein. The predicted contact pairs of amino acids are mapped to the predicted secondary structures that were obtained from online server Jpred.

2-Secondary structure traces from Cryo-EM density maps: The cryo-EM density maps were downloaded from the Electron Microscopy Data Bank (EMDB). The corresponding atomic structures were downloaded from PDB. Since there are no cryo-EM density maps corresponding to the two CASP targets (T1029, T1033), density maps were simulated in Chimera to 8Å resolution. The region of a-helices and β-sheets were detected from the density map using DeepSSETracer. For each segmented helix region, Principle Component Analysis (PCA) was used to derive a line (a-trace) for the central axis of an a-helix. For each segmented β sheet region, STrandTwister was used to predict traces of β-strands.

3-Deriving Topologies without Amino Acid Contacts: Secondary structure traces (SSTs), refer to the set of a-traces and β-traces detected from the Cryo-EM density map. The secondary structure sequence segments refer to a-helices or β-strands predicted using existing software such as JPred or SYMPRED. MultiTopoDP is a graph-based dynamic programming method to match between the secondary structure traces with secondary structure sequence segments. MultiTopoDP produces a list of top-ranked topologies and indicates the rank of the true topology.

4-Re-rank Topologies using Secondary Structure Contact Pairs: After amino acid contact pairs are mapped to secondary structure sequence segments, the secondary structure contact pairs were used to evaluate each possible topology and those topologies that satisfy the contact constraints were ranked higher(Figure 2). In each possible topology generated from MultiTopoDP, the set of secondary structure traces are mapped to the set of sequence segments. For a pair of secondary structure sequence segments that were predicted in contact, their corresponding traces indicated in each topology were evaluated for the shortest distance between the two traces. The shortest distance between the pair of traces is defined as the shortest distance between any two points, one from each line. A threshold of 12Å was used, and those pairs of traces with shortest distance more than the threshold were not considered as in contact. Finally, the percentage of pairs of secondary structure traces that satisfies the distance constraints ((Number of satisfied pairs/total number of pairs in contact) * 100) was calculated for each possible topology to re-rank the topologies.

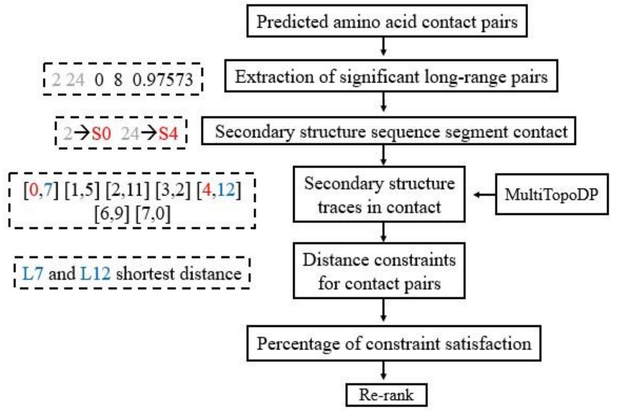

Figure 2. Evaluation of possible topologies using amino acid contact pairs. A list of possible topologies was produced using MultiTopoDP.

Results

The proposed approach to rank possible topologies of secondary structures combines three pieces of information: 3-dimensional location of secondary structures detected from the density map, sequence segments of secondary structures predicted from the protein sequence, and amino acid contact pairs predicted from the protein sequence. The approach was tested in seven cases including five cryo-EM density maps and two simulated density maps. The rank of the true secondary structure topology was improved after using amino acid contact information in six of the seven cases(Table 1). Results show a small set of possible topologies that includes the true topology can be produced even when errors exist in one or more of the three sources of information. The case of 6810-5y5x-H has 104 amino acid(Table 1 column 2) and its atomic structure contains four helices and one β-sheet (Table 1 column 3). Four a-traces and three β-traces (Figure 1) were used to match with four helix segments and four β-strand segments predicted using JPred to produce a list of possible topologies (Table 1 column 4 and 5). To evaluate the effect of using secondary structure contact pairs, we compared the rank of the true topology of secondary structures in two settings with/without using contact pairs. The rank of the true topology ideally is to be top 1, although it is often challenging to do so. When no secondary structure contact pair was incorporated, the true topology was ranked the 5th on the list (Table 1 column 7). The rank of the true topology is improved to top 1 (Table 1 column 8) when the six contact pairs of secondary structures were included (Table 1 column 6).

Case	#a.a.[a]	True Struct[b]	Seq Pred[c]	Image Detect[d]	Contact pairs[e]	Rank of True Topology	
						No_C[f]	With_C[g]
6810-5y5x-H	104	4/2	4/4	4/3	6/1	5	1
9534-5gpn-Ae	116	4/0	5/1	4/0	4/0	10	4
8518-5u8s-A	209	6/2	5/5	5/2	10/0	39	27
3948-6esg-B	102	3/0	3/1	3/0	3/0	13	7
2620-4uje-BH	194	6/3+3	5/3+3	4/3+3	7/2	15	4
T1029	125	6/4	3/5	3/4	4/0	462	200
T1033	100	3/0	6/0	4/0	3/0	85	85

Table 1. Secondary structure topology ranks produced using secondary structure sequence segments, amino acid contact pairs, and secondary structure traces. [a]The number of amino acids in the protein. [b]The number of a-helices/β-Strands in the protein' s true structure. (+) indicates number of β-Strands in each sheet. [c]The number of a-helices/β-Strands predicted using JPred. [d]The number of a-traces/ β-traces detected from the 3D density map. [e]The number of correct/wrong contact pairs predicted using MULTICOM or RaptorX. [f]Rank of True Topology without using contacts pairs. [g]Rank of True Topology using contacts pairs.

In the case of 6810-5y5x-H (Table 2), 58 pairs of long-range significant residue contacts were extracted, and 46 pairs involve two secondary structures. The 46 pairs were mapped to seven pairs of secondary structures that were predicted using JPred. Among the seven pairs of secondary structures, a pair of β-strands (S4, S6) has 20 pairs of significant long-range pairs of amino acids in contact. This suggests the existence of contact between secondary structures S4 and S6. Among the seven contact pairs of secondary structures, six are correctly predicted after a cross-check with the atomic structure. One pair (S4, S7) is not correct, with two significant long-range pair of amino acid predicted. We noticed that three of the seven pairs have 20, 11, and 8 predicted amino acid contact pairs respectively, many more than the other four pairs have. The analysis of the amino acid contact prediction suggests that those three pairs of secondary structures are most likely to be in contact.

Case	Contact Secondary Structure Pairs	AA pairs	Case	Contact Secondary Structure Pairs	AA pairs
6810-5y5x-H	(S0, S4) (β, β)	11	2620-4uje-BH	(S2, S4) (β, α)	12
	(S2, S3) (β, α)	2		(S2, S5) (β, β)	12
	(S3, S5) (α, α)	8		(S2, S3) (β, β)	4
	(S3, S4) (α, β)	2		(S3, S5) (β, β)	1
	(S4, S6) (β, β)	20		(S4, S5) (α, β)	2
	(S4, S7) (β, α)	2		(S5, S9) (β, α)	1
	(S5, S6) (α, β)	1		(S7, S8) (β, β)	4
9534-5gpn-Ae	(S2, S3) (α, α)	7		(S7, S9) (β, α)	1
	(S2, S5) (α, α)	8		(S9, S10) (α, β)	1
	(S3, S4) (α, α)	1	3948-6esg-B	(S1, S3) (α, α)	3
	(S4, S5) (α, α)	4		(S1, S2) (α, β)	2
8518-5u8s-A	(S1, S2) (α, α)	18		(S2, S3) (β, α)	4
	(S2, S4) (α, α)	11	T1029	(S2, S3) (β, β)	9
	(S5, S8) (β, β)	18		(S3, S4) (β, β)	11
	(S5, S9) (β, α)	4		(S4, S5) (β, β)	12
	(S6, S7) (β, β)	11		(S0, S3) (α, α)	2
	(S5, S10) (β, β)	8	T1033	(S2, S3) (α, α)	9
	(S5, S7) (β, β)	1		(S3, S4) (α, α)	2

Table 2. Secondary structure contact pairs derived from MULTICOM amino acid contact prediction.

Conclusion

Our results show the potential of combining the cryo-EM density maps with well analyzed contact information in deriving protein structures for cryo-EM density maps at medium resolution.

Acknowledgements

The work in this poster was supported by NIH R01-GM062968.