

Detection of Secondary Structures from 3D Protein Images of Medium Resolutions and its Challenges

Jing He^(✉), Dong Si, and Maryam Arab

Department of Computer Science, Old Dominion University,
Norfolk, VA 23529, USA
{jhe,dsi,marab}@cs.odu.edu

Abstract. Protein secondary structures such as α -helices and β -strands are major structural components in most proteins. The position of secondary structures provides important constraints in computing the tertiary structure of a protein. Electron cryomicroscopy is a biophysical technique that produces 3-dimensional images of large molecular complexes. For images at medium resolutions, such as 5–10 Å, major secondary structures may be computationally detected. This paper summarizes our recent work in detection of secondary structures using *SSETracer*, *SSELearner*, *StrandTwister* and *StrandRoller*. The detection of helices and β -strands is illustrated using *SSETracer* and *StrandTwister* with a small dataset.

Keywords: Image · Protein · Secondary structure · Electron cryo-microscopy · Pattern recognition · Geometrical modeling

1 Introduction

Proteins are essential in all biological processes. A protein is a polymer of amino acids folded in 3-dimensional space. The sequence of a protein refers to the linear order of amino acids. In nature, there are only twenty different kinds of amino acids, and therefore a protein can be considered as a string of twenty alphabets. Certain segments of the protein sequence tend to fold into helices and other segments tend to fold into β -strands (Fig. 1A). Multiple β -strands are stabilized by hydrogen bonds between them to form a β -sheet (Fig. 1). For example, the protein in Fig. 1 has five helices and one β -sheet containing four β -strands. Helices and β -sheets are major secondary structures of a protein. Almost all proteins contain helices and/or β -sheets. Therefore, secondary structures provide essential information about the tertiary structure of the protein.

Although proteins are small objects with typical sizes in the nanometer scale, experimental techniques are available to obtain 3D images of proteins. One of such techniques is called electron cryo-microscopy (cryo-EM), a biophysical technique to produce 3D images of large assembly of proteins [1–4]. Depending on the nature of biological specimen and accuracy of experimental procedures, the resulting images may reveal different level of details and have different resolutions. With a high-resolution image, structural details are resolved well enough to derive the atomic structure. At medium

resolutions, such as 5–10 Å, structural details are not well resolved. However, major secondary structures such as helices and β -sheets are visible. In principle, a helix with more than two turns appears as a cylinder, and a β -sheet appear as a thin layer of density. However, due to noise or incomplete data in the image, automatic detection of secondary structures is still challenging in certain situations. Various methods and tools have been developed to detect secondary structures from such images [5–12]. In this paper, we summarize the methods and tools we have developed recently and the challenges encountered.

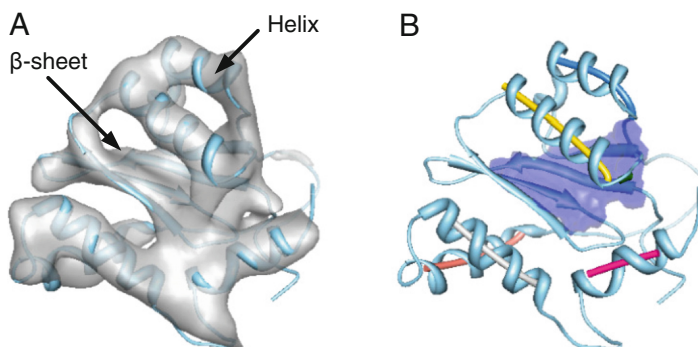


Fig. 1. Detection of helices and β -sheets from protein density images using *SSETracer*. (A) The backbone of a protein structure (PDB ID 2ITG) is shown as a ribbon (cyan). The surface view of its corresponding density image (gray) is superimposed. A helix and a β -sheet of the backbone are labeled. (B) Helices (colored lines) and a β -sheet (blue voxels) detected from density image using *SSETracer* are overlaid with the true structure of the protein (ribbon) (Color figure online).

2 Protein Secondary Structure Detection from 3D Images

2.1 Protein Density Images and the Pattern of Secondary Structures

The protein density map obtained using cryo-EM experimental technique is a 3D volumetric image, in principle, representing electron density of the protein. Those voxels with high values of density generally correspond to the locations where more atoms are located. Due to the nature of a helix, many atoms are positioned along the protein chain in a helical manner. At medium resolution such as 5–10 Å resolution, a helix generally appears as a cylinder, and a β -sheet may appear as a thin layer of density (Fig. 1A). Although such patterns are generally observed, they are affected by their closely located neighbors. The problem of secondary structure detection is to detect the location of helices, β -sheets and β -strands from a 3D image of the protein. In order to detect β -strands, β -sheets need to be detected first for images of medium resolutions. We have developed four methods for the detection of secondary structures *SSETracer*, *SSELearner*, *StrandTwister* and *StrandRoller*. Their relationship is illustrated in Fig. 2.

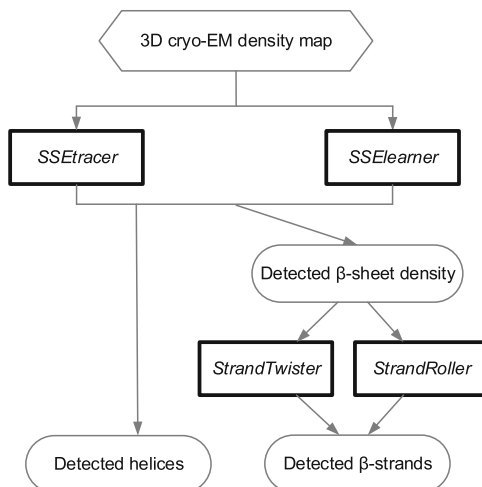


Fig. 2. Tools we developed for detection of protein secondary structures.

2.2 *SSETracer* and *SSELearner* for Detection of Helices and β -Sheets

SSETracer is a method for automatic identification of helices and β -sheets from 3D cryo-EM images at medium resolutions [9]. The methodology of the method was published in [9]. We here provide additional test cases using the most updated version. *SSETracer* characterizes three main local features: local structure tensor, distribution of skeleton voxels and local thickness. A simple voting procedure was used to determine if a particular point belongs to a helix or a β -sheet. We show the results of *SSETracer* using eight cases randomly selected from the Protein Data Bank (PDB). For each protein, the true structure downloaded from PDB was used to simulate a density image to 10 Å resolution using Chimera [13]. *SSETracer* was applied to such density maps and the results are summarized in Table 1. As an example for 2ITG, *SSETracer* correctly detected five of the six helices in this protein. The detected position of a helix is represented as a line that corresponds to the central axis of the helix. It missed a small 3_{10} helix with five amino acids in length (row 2 column 2 of Table 1). It wrongly detected a small helix (at the back of the protein, not clearly visible in Fig. 1B) that is supposed to be a turn. A turn of a protein chain may appear as a short cylinder in the image and therefore it can be confused with a short helix. Alternatively, we may estimate the accuracy in terms of the number of $C\alpha$ atoms that are located along the helices. There are sixty-six $C\alpha$ atoms in the six helices of the protein, and fifty-four were correctly detected. Seven $C\alpha$ atoms were wrongly detected due to the wrongly detected short helix and the longer detected length than expected for some helices. There is one β -sheet in this protein and eighteen $C\alpha$ atoms on the β -sheet were correctly detected. A correctly detected $C\alpha$ atom refers to a $C\alpha$ atom that has a detected helix/ β -sheet voxel within 2.5 Å radius of the atom.

Table 1. The accuracy of identified helices and β -Sheets using SSETracer.

PDB ID	#Hlx ^a	#C _{α} Hlx ^b	#Sht ^c	#C _{α} Sht ^d	Time ^e
1WAB	6/8/1	67/96/10	1/1	22/24/14	0:32
2ITG	5/6/1	54/66/7	1/1	18/21/17	0:23
4CSV	8/10/0	80/121/0	2/2	30/37/25	0:40
1CV1	8/9/0	108/123/4	1/1	11/14/0	0:23
4P1T	19/22/2	290/400/15	0/1	0/4/0	1:26
4D44	7/11/1	90/150/2	1/1	33/37/20	0:40
4XDA	8/12/0	97/143/6	1/2	50/78/38	0:39
4OZW	13/18/1	198/257/4	0/1	0/4/0	0:32

^aThe number of correctly detected/total number/wrongly detected helices.

^bThe number of correctly detected/total number/wrongly detected C α atoms of helices.

^cThe number of correctly detected/total number of β -sheets.

^dThe number of correctly detected/total number/wrongly detected C α atoms in β -sheets.

^eTime (in minutes) of the detection.

Table 2. The accuracy of identified helices and β -sheets using SSELearner.

PDB ID	#Hlx ^a	#C _{α} Hlx ^b	# Sht ^c	#C _{α} Sht ^d	Time ^e
1WAB	7/8/0	88/96/7	1/1	23/24/9	1:29
2ITG	6/6/1	62/66/6	1/1	19/21/7	1:23
4CSV	9/10/2	103/121/15	1/2	28/37/14	1:48
1CV1	8/9/0	107/123/9	1/1	10/14/13	1:01
4P1T	19/22/0	369/400/3	1/1	2/4/4	2:27
4D44	8/11/1	105/150/6	1/1	23/37/23	1:52
4XDA	11/12/2	112/143/36	2/2	59/78/30	2:01
4OZW	16/18/0	224/257/5	1/1	2/4/2	1:22

See Table 1 for caption a–d.

^eTime (in minutes) of the detection without the time of model generation.

We previously developed a machine learning approach, *SSELearner*, to automatically identify helices and β -sheets by using the knowledge from existing volumetric images [10]. The first component of the method develops features using local structure tensor and local thickness. The second component performs multi-task classification using Support Vector Machine (SVM). The post-processing step performs additional voxel clustering and filtering. *SSELearner* shows that it is possible to use one cryo-EM

map for learning in order to detect helices or β -sheet in another cryo-EM map of similar quality. With careful training, it is possible to improve detection accuracy using machine learning in the secondary structure detection problem [10].

We applied *SSETracer* and *SSELearner* on the same set of data containing eight protein images. The accuracy appears to be generally comparable between the two methods (Tables 1 and 2), although *SSELearner* is slightly more sensitive detecting more $C\alpha$ atoms. In certain cases such as 4XDA, it is less specific than *SSETracer* by detecting more wrong $C\alpha$ atoms. The main difference between the two methods is two-fold. *SSETracer* uses skeleton and *SSELearner* does not. *SSELearner* is a machine learning method that needs to generate models in the training process. *SSETracer* runs faster than *SSELearner* (Column 6 of Tables 1 and 2). For example, it takes about 1 min 26 s to run *SSETracer* and 2 min 27 s to run *SSELearner*. Note that the time does not include the time to generate a model for *SSELearner*. It may take many hours to produce a model. Libsvm library that was used in *SSELearner* needs to be tuned according to the actual problem for optimal performance. Our current version uses the default parameters of Libsvm in model generation, and that is a bottle neck of the method.

Table 3. Accuracy of β -strands detected using *StrandTwister*.

PDB ID	#Strands ^a	#C α ^b	Time ^c
1WAB	5/5	22/24	2:33
2ITG	4/4	16/21	2:38
4CSV	4/5	18/31	2:31
4D44	6/7	12/37	10:55
4XDA	8/9	29/45	17:25

^aThe correctly detected/total number of β -strands in the β -sheet.

^bThe correctly detected/total number of $C\alpha$ atoms of the β -sheet.

^cExecution time (in minute) to derive top ten possible sets of β -strands using *StrandTwister*.

2.3 *StrandTwister* to Predict β -Strands for Single Sheets

SSETracer is a tool to detect helices and β -sheets from 3D images, but it does not detect β -strands. Each β -sheet is composed of multiple β -strands. The spacing between two β -strands is 4.5–5 Å that makes it almost impossible to be visualized in a density image with 5–10 Å resolution. We proposed a new method, *StrandTwister* [8], to predict the traces of β -strands from a chunk of isolated β -sheet density. It does not rely on the existence of separation of β -strands and can be applied to images at lower resolutions. We showed that it is possible to predict the orientation of β -strands through the analysis of twist of a β -sheet. *StrandTwister* has two major components. The first one simplifies the voxels of a β -sheet into a polynomial surface. The second, also the major step, identifies right-handed β -twist from the polynomial surface model. *StrandTwister* appears to detect the traces of β -strands on major β -sheets quite accurately, particularly

at the central area of a β -sheet. The requirement of the program is to provide an image of an isolated single sheet. The current version of the program does not handle an image with multiple β -sheets. We report the result of β -strand prediction for five new cases in Table 3. As an example in 2ITG, the isolated density of β -sheet (gray in Fig. 3) was generated from *SSETracer*. *StrandTwister* produces ten possible sets of β -strands. The best detection (red lines in Fig. 3) refers to the set of detected lines that are closest to the true β -strands. In this case, the best detection contains all four β -strands of the β -sheet or sixteen of the twenty-one C α atoms of the β -sheet (Table 3 row 2). We observe in the five test cases and also previously that if the estimation of a β -sheet has good accuracy, *StrandTwister* often produces accurate results [8].

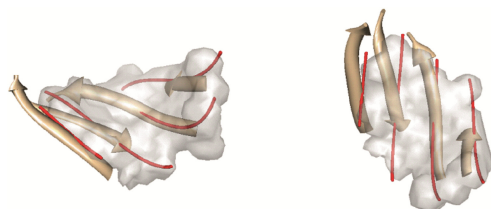


Fig. 3. Detection of β -strands from isolated β -sheet density using *StrandTwister*. β -sheet density (gray, same as the β -sheet in Fig. 1) is detected using *SSETracer* for sheet A of protein 2ITG (PDB ID). It is superimposed with the β -traces (red lines) best predicted using *StrandTwister* and the true structure (ribbon). The side view (left) and the top view (right) are shown (Color figure online).

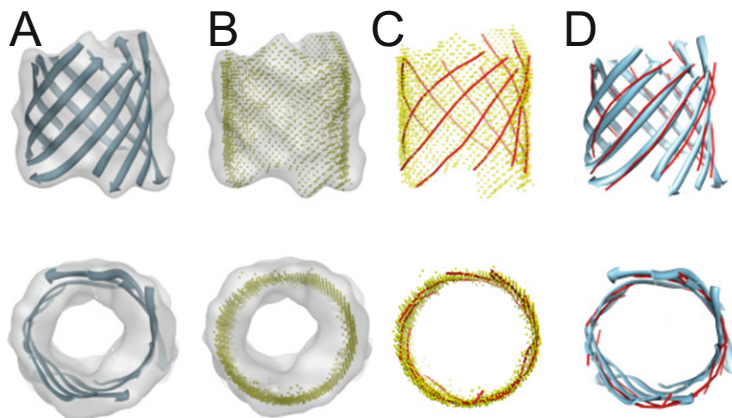


Fig. 4. Detection of β -strands from isolated β -barrel image using *StrandRoller*. (A) β -barrel image (gray) was simulated to 10 Å resolution using β -barrel structure (ribbon, PDB ID: 1RRX sheet A12). (B) The model surface (yellow) fit in the image. (C) The best predicted β -traces (red lines) were generated using *StrandRoller*. (D) The predicted β -traces superimposed with the true structure (ribbon) (Color figure online).

2.4 *StrandRoller* to Predict β -Strands for a β -Barrel Image

Various shapes of β -sheets have been observed in nature, particularly for β -sheets with long β -strands and/or more number of β -strands. A β -barrel is a β -sheet with an overall shape of a barrel (Fig. 4A). Many β -barrels are large β -sheets forming channels for trans-membrane activities. We proposed a method called *StrandRoller* to predict possible sets of β -traces from a β -barrel image at medium resolutions [14]. A β -barrel surface model (yellow in Fig. 4B) is first fit in the image and then possible sets of β -strands are modeled by strand-walking. The results of *StrandRoller* suggest that it is possible to derive a small set of possible β -traces (red lines in Fig. 4D) from a β -barrel image even when it is not possible to visualize the separation of β -strands [8].

3 Challenges in Secondary Structure Detection

The first method in automatic detection of secondary structures from 3D cryo-EM images is *Helixhunter*, developed in 2001 [5]. It was able to detect helices but not β -sheets. Although many methods have been developed to detect secondary structures, it is still challenging to detect them accurately. In this paper we showed four methods developed in our group to address this problem. We showed that the combination of image processing and geometrical modeling is capable of deriving new information that is not possible to derive using image processing alone. Deriving the position of β -strands from 3D images that do not, in principle, resolve β -strands has been a challenging problem in the last ten years. We showed that using image processing techniques, β -sheets can be first identified using *SSETracer* or other methods. The isolated β -sheet density image can be modeled geometrically using *StrandTwister* through the analysis of twist to derive the position of β -strands. Alternatively, the isolated β -barrel can be modeled using tilt angle knowledge to derive β -strands for β -barrels.

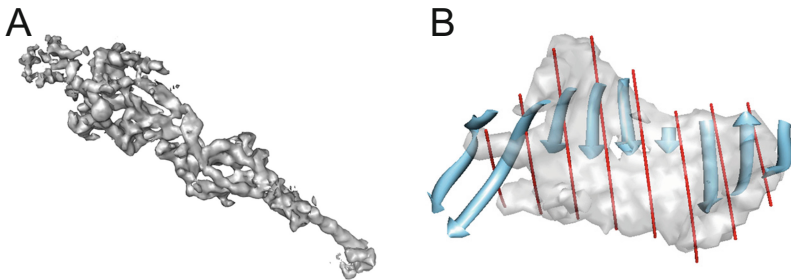


Fig. 5. Challenges in Secondary Structure Detection. (A) E2 monomer image (gray) from Encephalitis Virus (EMD_5276) at 4.4 Å resolution. The upper left region has much weaker density than the main part of the density map. (B) β -sheet image (gray) detected from simulated image of protein 4XDA (PDB ID) using *SSETracer*. The best predicted β -traces (red lines) are superimposed with the true structure (ribbon) of β -sheet AA1 (Color figure online).

In spite of improvement in methodology, cryo-EM data present great challenges for understanding 3D images. The 3D images obtained from cryo-EM technique are noisy and incomplete in many places. Even in the same image, it is often observed that certain regions have much better quality than others. As an example, the upper left region of the 3D image in Fig. 5A appears weaker than other regions. In this case, the resolution of the map is 4.4 Å. It is harder to interpret this region than others. A good secondary structure detection method needs to be aware of the local quality of the image. Although *StrandTwister* is able to predict β -strands from many β -sheet images, it is not very accurate for large and complicated β -sheets. For example, the β -sheet in Fig. 5B contains eight β -strands. Unlike a typical β -sheet, the β -strand near the middle of the sheet is very short in this case. As a result, there is over-estimation of the β -sheet at the middle and the β -strands predicted are not accurate for some strands. The nature of cryo-EM data and the biological diversity of molecules present interesting and challenging future for the understanding of 3D images.

Acknowledgements. The work in this paper is partially supported by NSF DBI-1356621, M&S fellowship and FP3 fund of the Old Dominion University. Authors' contribution: MA conducted the tests with the help of DS. DS prepared the figures. JH guided the project and wrote the manuscript.

References

1. Hryc, C.F., Chen, D.H., Chiu, W.: Near-atomic-resolution cryo-EM for molecular virology. *Curr. Opin. Virol.* **1**(2), 110–117 (2011)
2. Zhou, Z.H.: Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **18**(2), 218–228 (2008)
3. Zhou, Z.H., Dougherty, M., Jakana, J., He, J., Rixon, F.J., Chiu, W.: Seeing the herpesvirus capsid at 8.5 Å. *Science* **288**(5467), 877–880 (2000)
4. Zhang, X., Jin, L., Fang, Q., Hui, W.H., Zhou, Z.H.: 3.3 angstrom cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell* **141**, 472–482 (2010)
5. 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**(5), 1033–1044 (2001)
6. Kong, Y., Zhang, X., Baker, T.S., Ma, J.: A structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps. *J. Mol. Biol.* **339**(1), 117–130 (2004)
7. Kong, Y., Ma, J.: A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. *J. Mol. Biol.* **332**(2), 399–413 (2003)
8. Si, D., He, J.: Tracing beta-strands using strandtwister from cryo-EM density maps at medium resolutions. *Structure* **22**(11), 1665–1676 (2014)
9. Si, D., He, J.: Beta-sheet detection and representation from medium resolution cryo-EM density maps. In: *BCB 2013: Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, Washington, D.C., pp. 764–770, 22–25 Sept 2013
10. Si, D., Ji, S., Nasr, K.A., He, J.: A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. *Biopolymers* **97**(9), 698–708 (2012)

11. Dal Palu, A., He, J., Pontelli, E., Lu, Y.: Identification of alpha-helices from low resolution protein density maps. In: Proceeding of Computational Systems Bioinformatics Conference (CSB), pp. 89–98 (2006)
12. Rusu, M., Wriggers, W.: Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. *J. Struct. Biol.* **177**(2), 410–419 (2012)
13. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E.: UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**(13), 1605–1612 (2004)
14. Si, D., He, J.: Combining image processing and modeling to generate traces of beta-strands from cryo-EM density images of beta-barrels. In: Proceeding of EMBC, Chicago, 26–30 Aug 2014