Old Dominion University

ODU Digital Commons

Summer 2018

Deep Learning for Segmentation Of 3D Cryo-EM Images

Devin Reid Haslam
Old Dominion University, dhasl002@odu.edu

DEEP LEARNING FOR SEGMENTATION OF 3D CRYO-EM IMAGES

by

Devin Reid Haslam
B.A. May 2017, Old Dominion University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2018

Approved By:

Dr. Jing He (Thesis Advisor)

Dr. Mohammad Zubair (Member)

Dr. Desh Ranjan (Member)

Dr. Jiangwen Sun (Member)

ABSTRACT

DEEP LEARNING FOR SEGMENTATION OF 3D CRYO-EM IMAGES

Devin Reid Haslam
Old Dominion University, 2018
Thesis Advisor: Dr. Jing He

Cryo-electron microscopy (cryo-EM) is an emerging biophysical technique for structural determination of protein complexes. However, accurate detection of secondary structures is still challenging when cryo-EM density maps are at medium resolutions (5-10 Å). Most existing methods are image processing methods that do not fully utilize available images in the cryo-EM database. In this paper, we present a deep learning approach to segment secondary structure elements as helices and β-sheets from medium- resolution density maps. The proposed 3D convolutional neural network is shown to detect secondary structure locations with an F1 score between 0.79 and 0.88 for six simulated test cases. The architecture was also applied to experimentally-derived cryo- EM density regions of 571 protein chains. . The average F1 score for helix detection is 0.747 and 0.674 for β-sheets in a test involving seven cryo-EM density regions. Additionally, we extend an arc-length association method to β -strands and show that this method for measuring error is superior to many popular methods. An interactive tool is also presented that can visualize the results of this arc-length association method.

ACKNOLEDGEMENTS

There are many people who have contributed to the successful completion of this thesis. I extend many thanks to my committee members for their patience and hours of guidance on my research and editing of this manuscript. The untiring efforts of my advisor, Dr. Jing He, deserves special recognition.

NOMENCLATURE

α – Alpha

Å - angstrom

β - Beta

TABLE OF CONTENTS

Chapter 1 - SEGMENTATION OF PROTEIN SECONDARY STRUCTURES USING DEEP

LEARNING

Chapter 2 – SOFTWARE DEVELOPMENT FOR EVALUATION OF ACCURACY IN

PROTEIN SECONDARY STRUCTURE DETECTION FROM CRYO-ELECTRON

MICROSCOPY DENSITY MAPS

LIST OF TABLES

LIST OF FIGURES

LIST OF GRAPHS

CHAPTER 1- SEGMENTATION OF PROTEIN SECONDARY STRUCTURES USING

DEEPT LEARNING

I. INTRODUCTION

Proteins are imperative to living cells. The three-dimensional (3D) structure of a protein determines the function of protein. Cryo-electron microscopy (cryo-EM) is an important technique in molecular structure determination. Using cryo-EM, a growing number of large molecular complexes have been resolved to atomic resolutions [1, 2]. However, for cryo-EM density maps with a medium resolution (5-10 Å), it is much more challenging to recognize detailed molecular features. In most cases, it is not possible to derive atomic structures from these medium resolution images without the knowledge of known atomic structures as templates. When a template structure is available, fitting is used to derive atomic structure [3, 5]. When no suitable template structures are available, matching secondary structures that are detected from the density maps and those predicted from the sequence of the protein may suggest possible topologies of secondary structures [6-10].

The most common secondary structure elements (SSEs) in a medium-resolution density map are α-helices and β-sheets. The major difficulty of detecting secondary structures in such density maps is that the patterns of the SSEs can be indistinguishable from their narrowly located neighbors. Many methods have been developed to detect SSEs at medium resolutions. These approaches are mostly based on image-processing techniques, using cylinder-like templates to detect α-helices and plane-like templates to find β-sheets. The drawbacks of these methods include carefully selected parameters and under-utilizing large amount of existing density maps in the

database. Accurately detected SSEs are important for deriving protein structures from cryo-EM images at medium resolutions [11-17].

Generally, long α-helices, such as those with more than 20 amino acids, can be detected by easily. On the other hand, short α-helices can be easily confused with turns/loops. Similarly, large β-sheets show unique characteristics while small a β-sheet might be confused with an α-helix. Due to the small spacing of β-strands at about 4.5Å, β-strands are generally not visible in a medium-resolution density map. Several methods have been proposed to predict traces of β- strands from segmented β-sheet regions [18, 19]. As machine learning methods continue to show their merit in image processing tasks, several approaches have been taken to solve the problem presented. The authors of [20] used nested K nearest neighbors classifiers to detect α-helices. In addition, methods using support vector machines (SVM) have also been employed to identify α- helices and β-sheets [21]. However, empirically-derived features may not be representative enough to obtain state of the art accuracy. Most recently, Li et al. has shown potential of convolutional neural networks (CNNs) achieving good performance [22]. The main drawback of this method is that is was mostly tested using simulated 3D images rather than experimentally derived density maps. There is no CNN method that has been trained on experimental images.

Convolutional neural networks utilize arranged layers to learn complex features. CNNs have been shown to produce state of the art performance in a variety of image related applications [23-28]. More recently, CNNs have been extended to tasks involving image segmentation with good accuracy [29-31]. CNNs are appealing due to their ability to learn features with trainable parameters in tasks that require nonlinear relationships. Due to these advantages, we explore CNNs to segment secondary structures from cryo-EM 3D density maps.

II. METHODOLOGY

Several challenges are presented when attempting to segment secondary structures from a cryo-EM density map. One of these challenges is the large diversity of proteins in the database. The architecture used to segment these SSEs must be able to learn features from multiple scales. A second challenge existing is the varying sizes of proteins within the database. To overcome this problem, padding and patch-based testing was used. When using patches, training and testing were done with patches of size 48x48x48. A visualization of the patch can be seen in Figure 1. We attempted to find a size that would be small enough to eliminate the need for padding the 3D images, while still being large enough to hold important information when the receptive field is reduced to its smallest window.
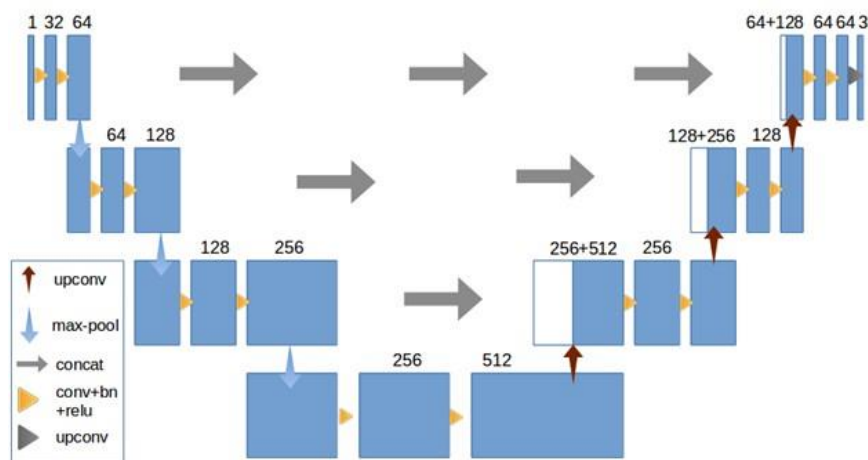
Architecture

Inspired by 3D-UNET [32], a simpler model was implemented. This model consists of an analysis path and a synthesis path. In the analysis path (Figure 2), each layer consists of two 3x3x3 convolutions, both followed by a batch normalization and a relu operation. Each layer in the analysis path is ended by a 3x3x3 max pool with a stride of two. By using a stride of two, we reduced the receptive field by a factor of two at the end of each layer in this path. After three layers that use increasingly more features, the analysis path has ended. The receptive field at the end of the analysis path is now eight times smaller than the original input. The synthesis path is very similar except each layer is ended with a transposed convolution increasing the receptive field by a factor of two. We also concatenate the results of each layer in the analysis path with the results of each synthesis layer. In the last layer we use a 1x1x1 convolution to decrease the amount of output channels to three labels. A more detailed description of the architecture can be seen in Figure 2.
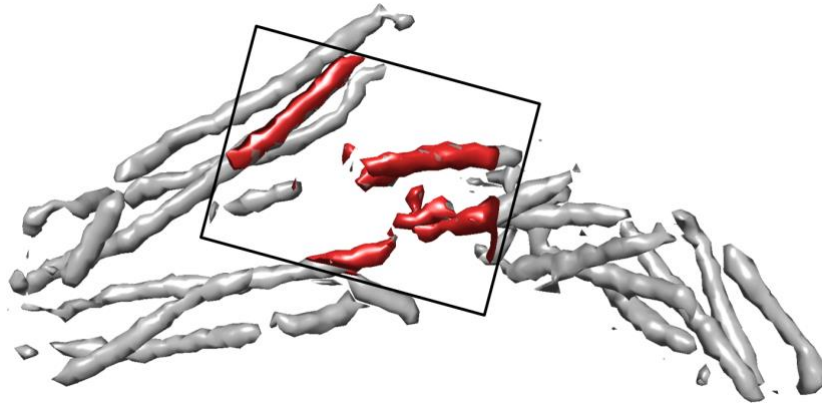
Figure 1. An example of the patch size extracted for training and testing. The patch (red) is superimposed on the simulated 3D image using the atomic structure of protein 2XS1(PDB ID).

A dropout rate of 50% was used during training. Unlike the previous work using a CNN architecture [22], no post-processing was performed, yet the model produces equivalent results as those using post-processing. Naturally, we used softmax with cross entropy to measure loss. In order to optimize this loss function, we employ an Adam optimizer with a 1e-4 training rate.

Figure 2. The 3D U-net architecture.

Initial Training

The presented architecture was initially used to test six simulated 3D images and one experimentally-derived cryo-EM density map. After collecting 31 atomic protein structures from the Protein Data Bank (PDB), we simulated each to 9Å resolution with a 1Å voxel size using UCSF Chimera [33]. Among the 31 3D images, 25 images were used for training, and the remaining six were used for testing. In order to fully utilize the simulated 3D images, each image was rotated around the X, Y, and Z axes with a random angle to produce 35 3D images as additional samples. Conversely, when using experimental data, we have downloaded each cryo-EM density map from Electron Microscopy Data Bank (EMDB) and the corresponding atomic structures from the PDB. Although there is a large number of cryo-EM maps with annotated resolution between 5-10Å, only those with visually good quality were used for training. When evaluating our model on experimental data, we used 42 cryo-EM maps with a total of 67 chains for training. This dataset only utilizes a small portion of the data available. Much of the training data is unique, but there are a few chains in the set that are similar. The experimental data used for training and testing have voxel sizes between 0.82 Å/voxel and 1.86 Å/voxel. We expect the network to learn the characteristics of SSEs even when the voxel size might be different. The results of this small experiment can be found in table 1. The results were good enough to move toward the harder problem of experimental data. This test was done using patch-based training and a batch size of four.

Large Scale Experimental Data Training

Experimental cryo-em data is more challenging to work with than simulated maps. Due to this problem, a larger experimental dataset was collected. This dataset consisted of 125 proteins with 571 chains. While this dataset was much larger than the previous experiments, it still did not

utilize the entirety of the pdb. In this training, we again use visually good maps with annotated

resolution between 5-10Å. Changes from the previous test include a larger batch size of 16 and a

more thorough rotation. This dataset was used for both patch based training and full image

training. The respective results can be seen in table 2 and 3.
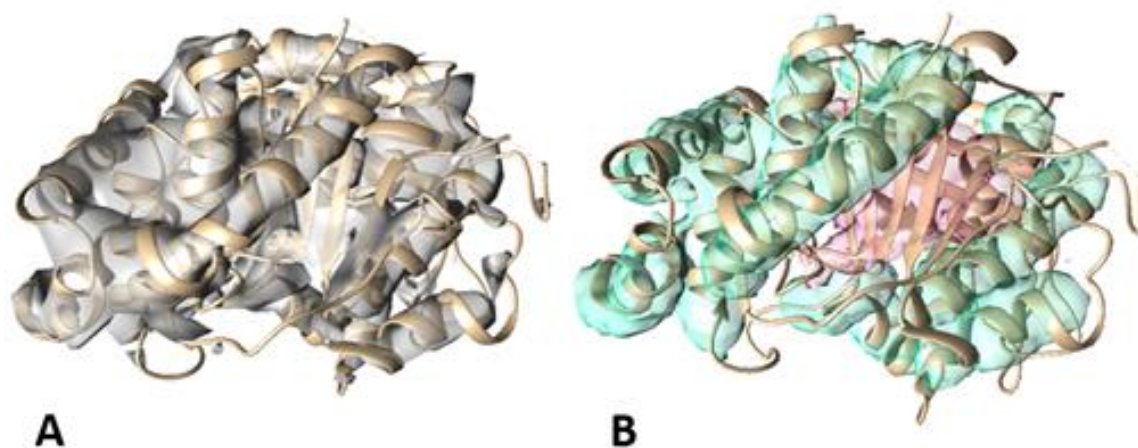
Figure 3. An example of secondary structure segmentation using the CNN architecture. (A) A 3D image simulated using the atomic structure of protein 3j7i_a (PDB ID) (shown in ribbon). (B) The detected helix regions (cyan), and β-sheet regions (pink) are superimposed with the atomic structure (ribbon).
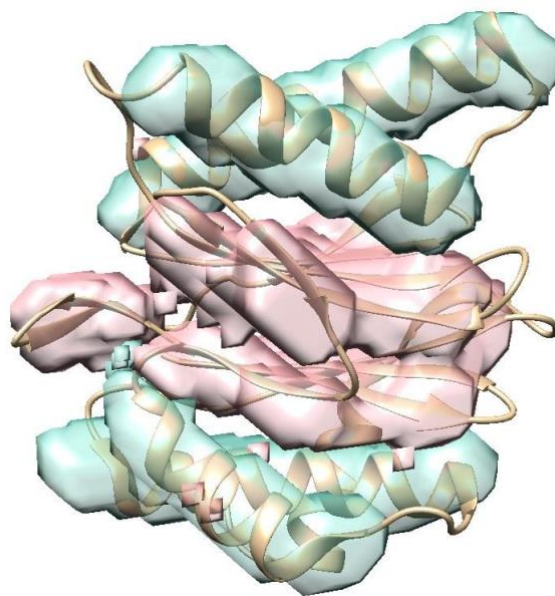
Figure 4. Detected helices and β-sheets from an experimentally-derived cryo-EM density map 1740 (EMD ID). The corresponding atomic structure of protein 3c92(PDB ID) (ribbon) is superimposed.

III. Results and Discussion

Initial Training

An example of secondary structures segmented from a simulated 3D image is shown in Figure 3. This protein 3j7i_a (PDB ID) has 17 helices and nine β-strands (Table 1). Visual inspection shows that both the helix regions and the β-sheet regions were identified correctly using the proposed CNN architecture. When testing, we also use patches of 48x48x48. As an example for 3j7i_a, nine patches of 48x48x48 were randomly selected from the entire density map. The accuracy of detected helix voxels was quantified for each patch using the F1 score. We observed that the F1 scores of different patches in a protein are similar. The averaged F1 score of nine patches in 3j7i_a is 0.806 for helix detection (Table 1). The average F1 score of helix, β-sheet,

and background is 0.789 for all nine patches in protein 3j7i_a. The F1 scores for helix detection are between 0.734 and 0.872 for the six test cases (Table 1). The F1 scores for β-sheet detection are from 0.749 to 0.999. The three cases with the highest F1 scores of β-sheets have small β-sheets with two strands only. The overall 3-class average of F1 scores are between 0.795 and 0.883 for the six simulated test cases.

Due to large amounts of noise found in experimentally-derived cryo-EM density maps, it is much more challenging to identify secondary structures in such images. An example of segmented helices and β-sheets is shown for cryo-EM density map EMD-1740 with 6.2 Å resolution (Figure 4). A chain of the protein 3c92 (PDB ID) was used as an envelope to extract the density region that corresponds to the chain in EMD-1740. This chain consists of five helices and 3 β-sheets, all of which appear to be segmented correctly (Figure 4). In this case, the average F1 score for 14 patches is 0.819 for helix detection, and 0.853 for β-sheet detection. The accuracy for cryo-EM case is comparable, with an overall F1 score of 0.828, to the accuracy of the simulated cases.

Large Scale Experimental Data Training

Table 2 displays the segmentation accuracy for seven cases of cryyo-EM density maps. The results were obtained from the patch-based training using a larger experimental datathan those used in Table 1 . While this dataset is larger than our previous tests, it does not contain all medium resolution chains found in the PDB. With a larger dataset, we hoped that the results from table 2 would exceed the experimental results in table 1. The new test (Table 2) did not produce results comparative to those (Table 1) using the simulated protein density maps. This might be that the experimental cryo-EM density maps are more challenging to learn than simulated density maps. We explored two ways of providing training data using either the patch-based or the full-image

training. In patch-based training, each chopped image has a size of 48x48x48, while each image used in the full-image training often has a larger size. The size in the full-image training varies, but the dimension was chosen to be multiples of 8. The accuracy for segmenting helices and β-sheets is shown as the F1 score (Table 3). With identical training parameters, full image testing produced far better results than patch-based training. As an example, the average F1 score for helix detection increased from 0.583 (patch-based training) to 0.747 (full-image training). The average F1 score for β-sheet detection increased from 0.475 to 0.674 after using full-image training. We noticed that the detection accuracy is better for helices than for β-sheets. A few factors may have contributed to this difference. Generally a β-sheet does not have a strong characteristic shape as for a helix. The imbalance of β-sheets in the training data may also affected the accuracy. The greater accuracy can be explained by the location patch within the image. Many times, a patch chops a secondary structure due to the limited size of a patch. That may obscure the geometry of secondary structures during training. The results indicate that CNNs have potential to extract secondary structures from cryo-EM images.

PDB ID	Patch Number	Helix Number	Strand Number	F1-Helix	F1-Sheet	F1-Background	F1-Avg
3j7i_a	9	17	9	0.806	0.749	0.812	0.789
1T79	10	12	4	0.872	0.766	0.861	0.883
1cv1	8	7	3	0.734	0.878	0.774	0.795
2XS1	8	26	2	0.81	0.998	0.802	0.87
3MK4	7	15	2	0.8	0.999	0.794	0.864
4P1T	7	25	2	0.822	0.998	0.812	0.877
3C92 (1740)	14	182	307	0.819	0.853	0.828	0.833

Table 1. Detection accuracy of three classes (helix, β-sheet, and background) using patch-based training. Row 2 to row 7 are simulated test cases using atomic structures of PDB. Row 8 involves an experimentally-derived test case with its EMDB ID indicated in parentheses.

PDB ID	EMDB ID	Patch Number	Helix Number	Strand Number	F1-Helix	F1-Sheet	F1-Background	F1-Average
3c92_1	1740	15	5	12	0.64	0.61	0.69	0.64
3j6p_B	5931	16	17	19	0.48	0.53	0.51	0.51
5iya_C	8135	9	6	10	0.62	0.35	0.52	0.49
5iya_E	8135	8	10	8	0.56	0.54	0.57	0.55
5vox_B	8724	12	27	6	0.64	0.37	0.66	0.54
5vox_O	8724	11	10	8	0.65	0.45	0.62	0.57

Table 2. Detection accuracy of three classes (helix, β-sheet, and background) using cryo-EM test cases. Seven experimental test cases were used with both the PDB ID (column 1) and EMDB ID (column 2) labeled. These results were obtained through patch based training.

PDB ID	EMDB ID	Helix Number	Strand Number	F1-Helix	F1-Sheet	F1-Background	F1-Average
3c92_1	1740	5	12	0.83	0.87	0.85	0.85
3j5m_A	5779	8	29	0.78	0.84	0.82	0.81
5iya_C	8135	9	6	0.78	0.65	0.61	0.68
5iya_E	8135	8	10	0.68	0.54	0.59	0.61
5vox_B	8724	12	27	0.73	0.61	0.59	0.64
5vox_O	8724	11	10	0.67	0.5	0.62	0.6
5vox_N	8724	6	6	0.76	0.71	0.82	0.76

Table 3. Detection accuracy of three classes (helix, β-sheet, and background) using cryo-EM test cases and full image training. Seven test cases involving experimental cryo-EM density maps were used. These results were obtained through full-image training.

IV.   Conclusion

Deriving atomic structures from medium-resolution cryo-EM density maps is challenging. An important step to derive the atomic structure automatically is detecting the location of secondary structures within the density map. We have presented a 3D convolutional neural network for segmentation of secondary structure elements from cryo-EM images. Although CNN has been shown as a powerful image processing method, there is limited work developing CNN architectures that are effective in 3D segmentation problems for protein secondary structure detection from cryo-EM density maps. Using 3D UNET as a guide [32], we have created an encoder decoder architecture employing 3D convolutions to capture features along three dimensions. We show that this version of 3D U-Net can achieve good accuracy on six

experimental density maps. We plan to improve this model and to perform an even larger test using more cryo-EM density maps.

REFERENCES

[1]   Zheng Liu, Fei Guo, Feng Wang, Tian-Cheng Li, and Wen Jiang. 2016. 2.9 Å Resolution Cryo-EM 3-D Reconstruction of Close-packed Virus Particles. *Structure (London, England : 1993)* 24, 2 (February 2016), 319–328. DOI:https://doi.org/10.1016/j.str.2015.12.006

[2]   Xiao-chen Bai, Israel S Fernandez, Greg McMullan, and Sjors HW Scheres. 2013. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* 2, (February 2013). DOI:https://doi.org/10.7554/eLife.00461

[3]   Kwok-Yan Chan, Leonardo G. Trabuco, Eduard Schreiner, and Klaus Schulten. 2012. Cryo-Electron Microscopy Modeling by the Molecular Dynamics Flexible Fitting Method. *Biopolymers* 97, 9 (September 2012), 678–686. DOI:https://doi.org/10.1002/bip.22042

[4]   Gunnar F. Schröder, Axel T. Brunger, and Michael Levitt. 2007. Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. *Structure* 15, 12 (December 2007), 1630–1641. DOI:https://doi.org/10.1016/j.str.2007.09.021

[5]   Willy Wriggers and Stefan Birmanns. 2001. Using Situs for Flexible and Rigid-Body Fitting of Multiresolution Single-Molecule Data. *Journal of Structural Biology* 133, 2–3 (February 2001), 193–202. DOI:https://doi.org/10.1006/jsbi.2000.4350

[6]   Kamal Al Nasr, Lin Chen, Dong Si, Desh Ranjan, Mohammad Zubair, and Jing He. 2012. Building the initial chain of the proteins through de novo modeling of the cryo-electron microscopy volume data at the medium resolutions. 490–497. DOI:https://doi.org/10.1145/2382936.2382999

[7]   Kamal Al Nasr, Desh Ranjan, Mohammad Zubair, Lin Chen, and Jing He. 2014. Solving the Secondary Structure Matching Problem in Cryo-EM De Novo Modeling Using a Constrained K-Shortest Path Graph Algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11, 2 (March 2014), 419–430. DOI:https://doi.org/10.1109/TCBB.2014.2302803

[8]   Kamal Al Nasr, Desh Ranjan, Mohammad Zubair, and Jing He. 2011. Ranking valid topologies of the secondary structure elements using a constraint graph. *J Bioinform Comput Biol* 9, 3 (June 2011), 415–430.

[9]   Devin Haslam, Mohammad Zubair, Desh Ranjan, Abhishek Biswas, and Jing He. 2016. Challenges in matching secondary structures in cryo-EM: An exploration. 1714–1719. DOI:https://doi.org/10.1109/BIBM.2016.7822776

[10]   Tao Ju, Matthew L. Baker, and Wah Chiu. 2007. Computing a family of skeletons of volumetric models for shape description. *Computer-Aided Design* 39, 5 (May 2007), 352–360. DOI:https://doi.org/10.1016/j.cad.2007.02.006

[11] Matthew L. Baker, Tao Ju, and Wah Chiu. 2007. Identification of Secondary Structure Elements in Intermediate-Resolution Density Maps. *Structure* 15, 1 (January 2007), 7–19. DOI:https://doi.org/10.1016/j.str.2006.11.008

[12] A. Dal Palù, J. He, E. Pontelli, and Y. Lu. 2006. Identification of alpha-helices from low resolution protein density maps. *Comput Syst Bioinformatics Conf* (2006), 89–98.

[13] Wen Jiang, Matthew L. Baker, Steven J. Ludtke, and Wah Chiu. 2001. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *Journal of Molecular Biology* 308, 5 (May 2001), 1033–1044. DOI:https://doi.org/10.1006/jmbi.2001.4633

[14] Yifei Kong and Jianpeng Ma. 2003. A Structural-informatics Approach for Mining β-Sheets: Locating Sheets in Intermediate-resolution Density Maps. *Journal of Molecular Biology* 332, 2 (September 2003), 399–413. DOI:https://doi.org/10.1016/S0022-2836(03)00859-3

[15] Mirabela Rusu and Willy Wriggers. 2012. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. *Journal of Structural Biology* 177, 2 (February 2012), 410–419. DOI:https://doi.org/10.1016/j.jsb.2011.11.029

[16] Dong Si and Jing He. 2007. Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps. 764–770. DOI:https://doi.org/10.1145/2506583.2506707

[17] Zeyun Yu and Chandrajit Bajaj. 2008. Computational Approaches for Automatic Structural Analysis of Large Biomolecular Complexes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 5, 4 (October 2008), 568–582. DOI:https://doi.org/10.1109/TCBB.2007.70226

[18] Dong Si and Jing He. 2017. Modeling Beta-Traces for Beta-Barrels from Cryo-EM Density Maps. *BioMed Research International* 2017, (2017), 1–9. DOI:https://doi.org/10.1155/2017/1793213

[19] Dong Si and Jing He. 2014. Combining image processing and modeling to generate traces of beta-strands from cryo-EM density images of beta-barrels. 3941–3944. DOI:https://doi.org/10.1109/EMBC.2014.6944486

[20] Lingyu Ma, M. Reisert, and H. Burkhardt. 2012. RENNSH: A Novel alpha-Helix Identification Approach for Intermediate Resolution Electron Density Maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9, 1 (January 2012), 228–239. DOI:https://doi.org/10.1109/TCBB.2011.52

[21] Dong Si, Shuiwang Ji, Kamal Al Nasr, and Jing He. 2012. A Machine Learning Approach for the Identification of Protein Secondary Structure Elements from Electron Cryo-

Microscopy Density Maps. *Biopolymers* 97, 9 (September 2012), 698–708. DOI:https://doi.org/10.1002/bip.22063

[22] Rongjian Li, Dong Si, Tao Zeng, Shuiwang Ji, and Jing He. 2016. Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. 41–46. DOI:https://doi.org/10.1109/BIBM.2016.7822490

[23] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (January 2013), 221–231. DOI:https://doi.org/10.1109/TPAMI.2012.59

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60, 6 (May 2017), 84–90. DOI:https://doi.org/10.1145/3065386

[25] Tao Zeng, Rongjian Li, Ravi Mukkamala, Jieping Ye, and Shuiwang Ji. 2015. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics* 16, 1 (December 2015). DOI:https://doi.org/10.1186/s12859-015-0553-9

[26] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage* 108, (March 2015), 214–224. DOI:https://doi.org/10.1016/j.neuroimage.2014.12.061

[27] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. 2013. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot and Nassir Navab (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 411–418. DOI:https://doi.org/10.1007/978-3-642-40763-5_51

[28] Y. LeCun, Fu Jie Huang, and L. Bottou. 2004. Learning methods for generic object recognition with invariance to pose and lighting. 97–104. DOI:https://doi.org/10.1109/CVPR.2004.1315150

[29] Viren Jain and Sebastian Seung. 2009. Natural Image Denoising with Convolutional Networks. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio and L. Bottou (eds.). Curran Associates, Inc., 769–776. Retrieved from http://papers.nips.cc/paper/3506-natural-image-denoising-with-convolutional-networks.pdf

[30] Srinivas C. Turaga, Joseph F. Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H. Sebastian Seung. 2010. Convolutional Networks Can Learn to Generate Affinity Graphs for Image Segmentation. *Neural Computation* 22, 2 (February 2010), 511–538. DOI:https://doi.org/10.1162/neco.2009.10-08-881

[31] Tao Zeng, Rongjian Li, Ravi Mukkamala, Jieping Ye, and Shuiwang Ji. 2015. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics* 16, 1 (December 2015). DOI:https://doi.org/10.1186/s12859-015-0553-9

[32] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv:1606.06650 [cs]* (June 2016). Retrieved June 1, 2018 from http://arxiv.org/abs/1606.06650

[33] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. UCSF Chimera?A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25, 13 (October 2004), 1605–1612. DOI:https://doi.org/10.1002/jcc.20084

CHAPTER 2– SOFTWARE DEVELOPMENT FOR EVALUATION OF ACCURACY IN

PROTEIN SECONDARY STRUCTURE DETECTION FROM CRYO-ELECTRON

I. Introduction

The Electron Microscopy Data Bank holds three-dimensional maps with a large range of spatial resolutions. Many density maps that have a resolution between 6-10Å are linked to atomic models found in the Protein Data Bank. In most cases, atomic models are derived directly from low resolution maps or are fitted from similar known structures. At this time, no accurate tools exist for deriving the atomic model of medium resolution maps.

Secondary structure elements such as helices and β-sheets are the most visible structural feature in medium resolution images. Helices are generally visible in images with resolution 10Å. On the other hand, β-sheets tend to become visible at resolutions around 8Å. As more methods are developed to detect secondary structure elements from medium resolution data [1-12], it is becoming more important to quantify the geometry of these detected features.

Accurately measuring the error between detected secondary structures and the corresponding true structure is needed for multiple reasons. While several measurements exist, it is not clear how sensitive the respective methods are. Accurately quantifying error is an important step for enhancement of secondary structure detection methods [13].

Arc length association is a sensitive method for measuring discrepancy of secondary structures presented by Zeil *et al.* known as arc-length association [14]. This method calculates both lateral and longitudinal discrepancies. The separation of the two discrepancies allows precise measurement of the length and shift errors that account to most inaccuracy in a detection. In this chapter the idea was applied to β-strands and created an interactive tool to easily visualize this

error. Evidence is presented that this arc-length association method is superior to other methods of measuring error.

II. Methodology

In medium-resolution cryo-EM density maps, the location of helices can be roughly detected by recognizing their cylinder like shape. Although various methods exist to detect the location of helices, we applied *SSETracer* [15] to determine these positions. Using features of local density such as local structure tensor, local thickness, continuity of the skeleton, and density vales, *SSETracer* can detect helices (and β-sheets). In *SSETracer* a detected helix is represented by a set of points located along the central axis of the helix.

The locations of β-strands are much more difficult to detect in medium-resolution cryo-EM images. At medium resolutions, the density map does not clearly represent the location of a β-strand. At 5-10Å resolutions only the location of the entire β-sheet can be resolved. Using features of local density, *SSETracer* was applied to detect the location of these β-sheets. Given an accurate model of the entire β-sheet, we can apply an iterative Bezier method to accurately fit the sheet with a surface representation [16]. Using this derived surface representation to quantify the twist of the beta sheet, a small set of likely β-strand locations can be provided [17]. Using this method, β-strand locations are represented by points located along the central axis of the β-strand.

To compare the set of points detected from the density map with the actual axis of a helix, we must derive the actual axis using the atomic structure. Using the backbone of a helix, the central axis was calculated by averaging the geometric centers of four consecutive amino acids in the helix [14]. The line formed from such points will be shorter than the actual axis due to averaging. Therefore, the ends of each axis is determined by projecting the first and last Cα atom to the first

and last segment of the axis, respectively. To determine the amino acid range of a helix in the atomic model or structure, the DSSP annotation of secondary structures was used. For helices that are shorter than 4 amino acids, we represent the helix with a single point. This point is chosen by calculating the midpoint between all amino acids that make up this short helix.

The calculation of the trace of β-strand's central axis is very similar to the calculation of a helix's actual axis. Using the backbone of a β-strand, the central axis was calculated by averaging two consecutive geometric centers of amino acids in the strand. The geometrical center was calculated using the nitrogen atoms of the amino acid sequence. Just as with helices, we use DSSP annotation of secondary structures and determine the endpoints of the strand by projecting the first and last Cα atom to the axis.

A central axis derived from the atomic structure and the axis detected from the density map is represented by a set of points. The number and spacing of the points are often different for the two axes. In order to make up for these discrepancies, we use a cubic Hermite spline to interpolate each set of points on the axis.
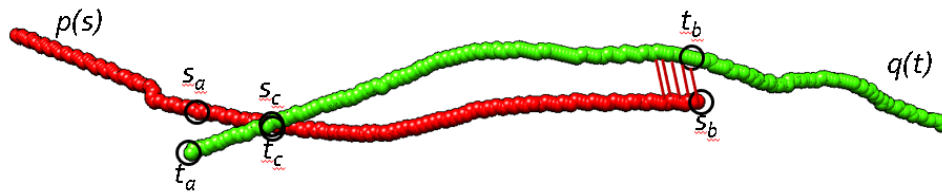
Figure 5. The visualization of the arc-length association method. S_a and t_a represent the first point that can be associated laterally between the two lines. On the other hand S_b and t_b are the last points that can be associated laterally. All points outside these two extremes are calculated as longitudinal discrepancy.
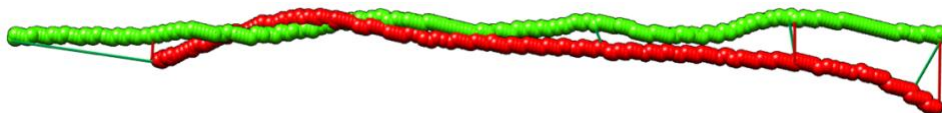
Figure 6. The visualization of as popular error method known as 2-way distance. In this method the distance is calculated from each point on one line to the closest point on the other line. This is repeated for both lines and then averaged.

Sensitivity, sometimes known as the true positive rate, measures the proportion of amino acids on a secondary structure that are close to the detected axis. Specificity, sometimes known as the true negative rate, measure the proportion of correctly undetected amino acids near a secondary structure. More specifically, if a Cα atom is less than 2.5Å from the detected axis, it is considered as belonging to that detection [14]. In calculating Sensitivity, we already know which amino acids belong to a certain secondary structure. Calculating specificity is more difficult because one does not know which amino acids are close enough to a secondary structure to be possibly falsely detected. This is solved by extending the actual axis 15Å in both directions and considering all amino acid 8Å from this extended axis to be falsely detected. Using sensitivity and specificity we calculate an F1 score 2 * ((sensitivity * specificity) / (sensitivity + specificity)). Although we went to great lengths to ensure a consistent F1 score, from examining chartA, it can be seen that the F1 score appears inconsistent. This is due to the great variety of spacing of secondary structures in different proteins. This concept is discussed more thoroughly below.

In order to provide an easy way for researchers to use our proposed arc length association method, we created a tool that can be used in a software known as chimera. This tool provides an easy way for users to visualize the accuracy of their secondary structure detection method.
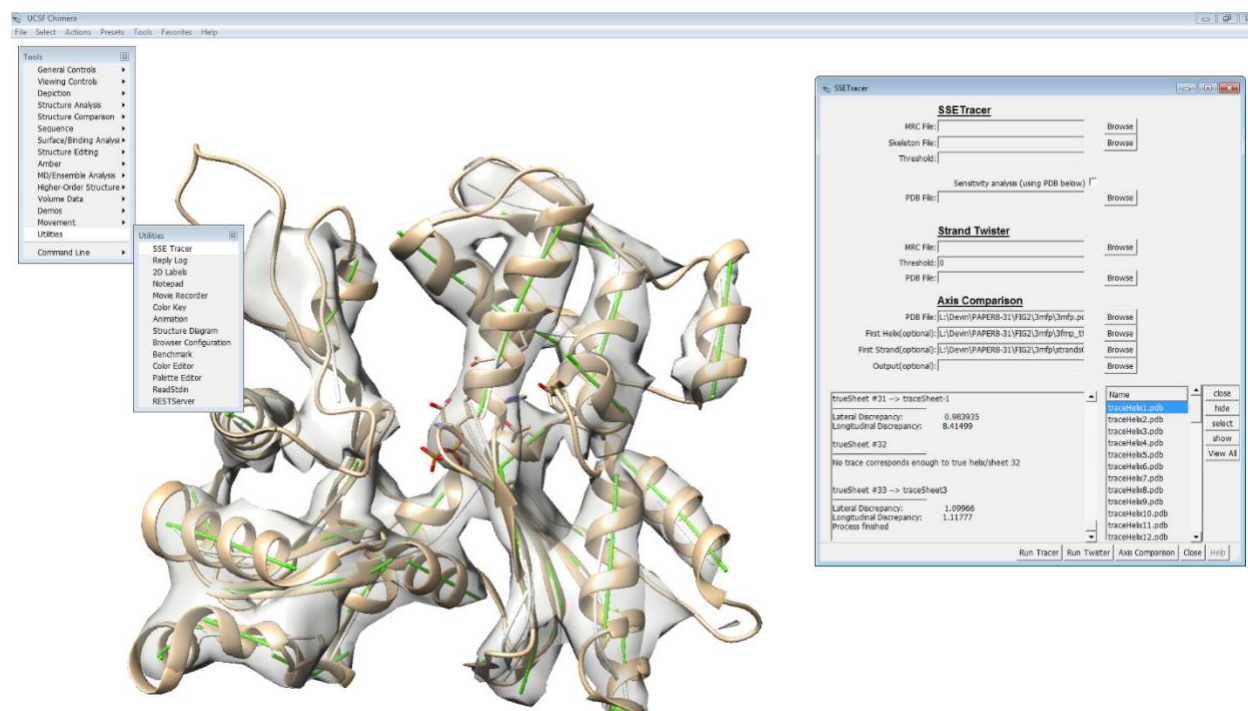
Figure 7. A screenshot of the interactive tool AxisComparison integrated in Chimera, a popular 3D molecular visualization tool. In the GUI seen on the right of the image, a user can see the error calculated and then use the tool to highlight the corresponding place in the image. The image displays the true central axis in green and predicted axes in white.

III. Results and Discussion

For our testing, we worked with both simulated and experimental data. In both cases, the atomic structures were download from the Protein Data Bank. When using experimental data, we downloaded a variety 3D density maps from the PDB of resolutions between 5-10Å. When testing simulated data, each density map was simulated using UCSF Chimera at 10Å. Secondary structures were assigned to the atomic structures using DSSP at the PDB web site. The 18 simulated proteins tested included 54 helices and 53 strands. On the other hands, 9 experimental proteins were tested which included 60 helices and 62 strands.

Both helices and β-strands can be accurately approximated by their central axes. An effective method to compare a secondary structure from a density map with its atomic model is to associate the relative position of their central axes. In order to characterize the effect of length

difference and positional shift we measured lateral and longitudinal discrepancies of the axial lines. In general, we notice that the lateral discrepancies are mostly small, within 2 Å for 96% of the test cases. However, the longitudinal discrepancies are much larger displaying an error of 2 Å for 92% of the test cases. This trend can be clearly seen for both helices and β-strands in simulated and experimental data alike. This result suggests that secondary structure detections are generally positioned in line (providing confidence in the detection), but there are various factors such as map artifacts, conformational variability, and modeling error that effect the accuracy of the longitudinal discrepancy [13].

The proposed arc-length association method is more sensitive and accurate than some of the previous ways to measure error. It can be seen from chart A that 2-way error is very closely associate with lateral error. As longitudinal error increases rapidly, 2-way error does not change very much at all. When comparing lateral and longitudinal error to F1 score in chart A, it can be seen that F1 score is very inconsistent. This is mostly due to sensitivity and specificity being calculated by examining the positions of amino acids.
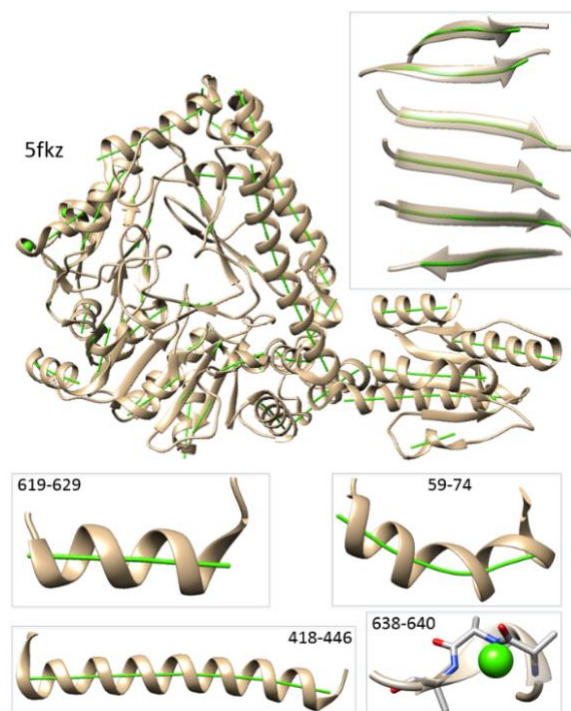
Figure 8. Visualization of the true central axis of all secondary structures in protein 5fkz.
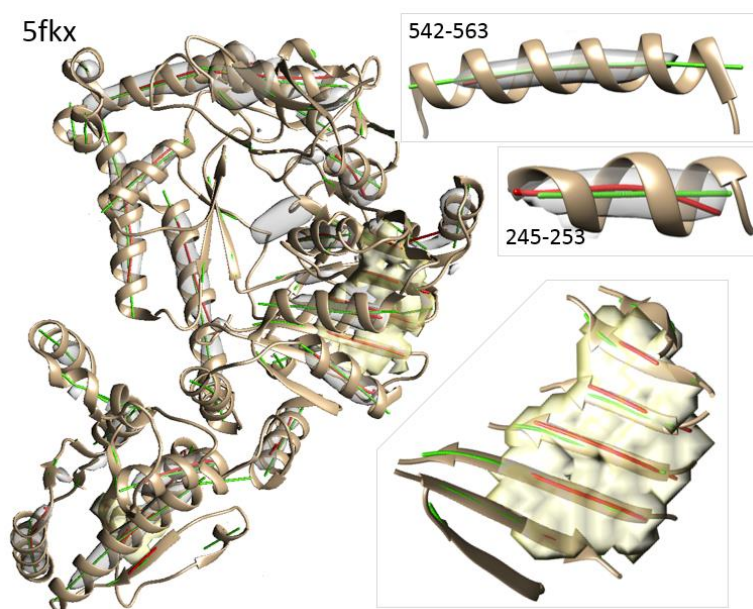
Figure 9. A visualization of predicted secondary structures (red) compared to the true central axis of that secondary structure (green) for experimental image 3204.
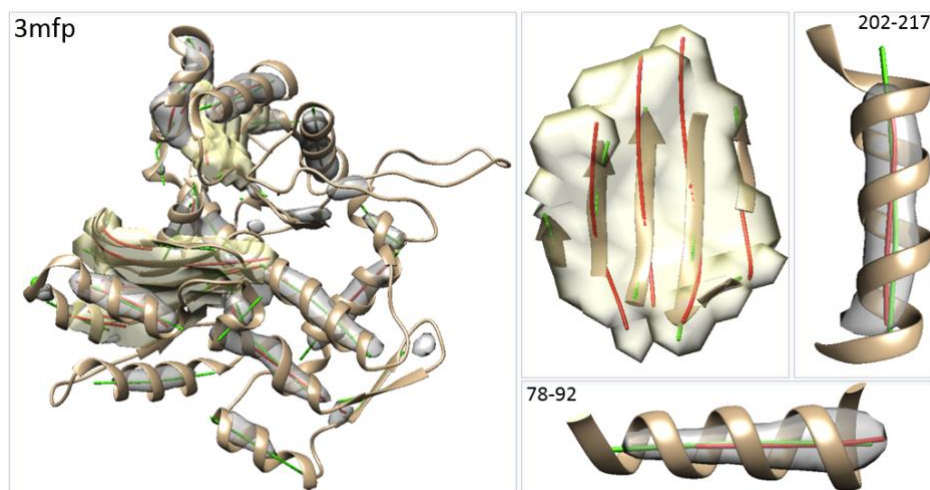
Figure 10. A visualization of predicted secondary structures (red) compared to the true central axis of that secondary structure (green) for simulated image of 3mfp at 8A resolution.
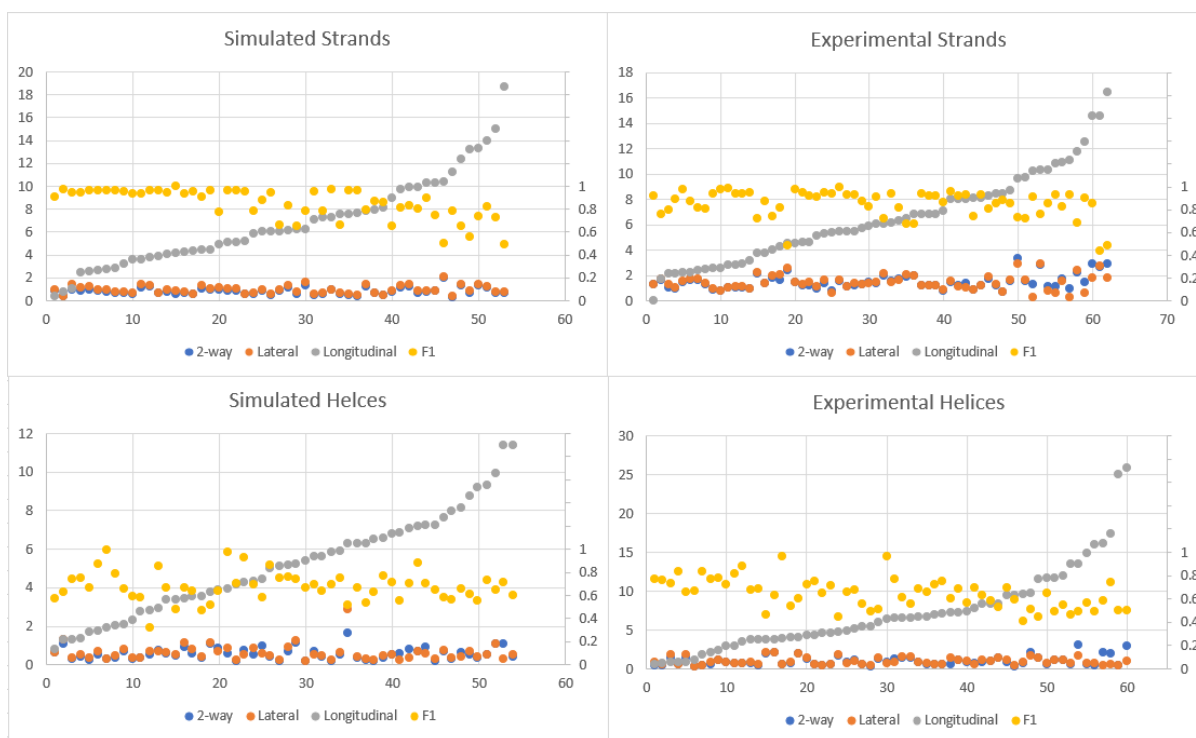
Graph 1. The visualization of error using three different methods on four different datasets. It can be seen that longitudinal (gray) and lateral error(organge) values are much more sensitive than the inconsistent F1 score (yellow) and 2-way distance (blue).

Experimental strands								
	Length True	**Length Detected**	**2-way**	**Lateral**	**Longitudinal**	**Specificity**	**Sensitivity**	**F1**
1237_2GSY_C								
Strand-55-59	39.66	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-67-74	16.41	27.5	0.96	0.33	11.09	0.87	1	0.93
Strand-80-87	25.68	23.31	0.75	0.69	8.43	0.9	0.86	0.88
Strand-123-129	25.6	17.99	0.76	0.91	7.08	0.87	0.86	0.86
Strand-154-159	23.36	2.601	1.14	1.12	5.47	0.86	1	0.93
Strand-349-355	22.97	35.5	1.48	0.62	12.53	0.82	1	0.9
1237_2GSY_G								
Strand-225-236	38.95	27.37	2.23	2.44	11.78	0.92	0.55	0.68
Strand-255-265	36.18	31.74	2.15	2.28	3.77	0.89	0.6	0.72
Strand-270-277	24.38	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-305-316	39.63	35.53	2.03	2.2	6.11	0.83	0.64	0.72
1740_3C92_A								
Strand-37-42	20.49	26.52	1.43	1.45	6.03	0.82	1	0.91
Strand-45-52	26.08	28.36	1.61	1.76	2.28	0.88	0.86	0.87
Strand-162-165	13.81	16.19	1	0.95	3.17	0.9	1	0.95
Strand-209-216	26.62	13.54	2.68	2.74	14.61	0.95	0.29	0.44
Strand-222-223	6.14	1.46	1.79	2	4.07	0.97	0.5	0.74
1740_3C92_B								
Strand-67-71	15.27	19.52	1.69	2.1	4.25	0.9	0.75	0.82
Strand-74-80	23.74	22.38	1.22	1.2	6.87	0.86	1	0.94
Strand-134-142	30.47	25.87	1.26	1.25	6.87	0.86	1	0.92
Strand-145-151	23.07	25.8	1.26	1.25	6.87	0.86	1	0.92
Strand-157-159	9.91	9.72	1.17	0.66	10.87	0.87	1	0.93
1740_3C92_O								
Strand-2-8	23.77	26.32	0.9	0.94	2.55	0.88	1	0.94
Strand-11-16	20.36	22.81	1.16	0.77	10.32	0.91	0.8	0.85
Strand-124-129	16.62	19.47	1.03	1.1	2.85	0.88	1	0.94
Strand-174-179	20.07	17.97	1.07	1.12	2.94	0.9	1	0.94
Strand-183-186	12.04	7	2.44	2.58	4.52	0.95	0.33	0.49
1740_3C92_Q								
Strand-34-38	15.35	7.95	1.95	2	6.85	1	0.5	0.67
Strand-41-44	13.45	19.59	1.59	1.44	6.14	0.89	1	0.94
Strand-98-104	23.29	24.57	1.2	1.33	4.62	0.9	1	0.95
Strand-108-114	23.28	29.14	1.44	1.4	5.86	0.83	0.83	0.83
Strand-118-121	12.79	16.04	1.43	1.05	8.06	0.88	1	0.93
1780_3IZ5_AC								
Strand-73-76	11.89	9.15	1.08	1.35	2.13	1	0.67	0.8
Strand-91-96	19.59	14.6	0.89	0.9	8.1	0.98	0.6	0.74
Strand-117-124	25.97	27.63	1.63	1.71	2.39	0.78	0.86	0.82
Strand-130-136	22.5	26.29	1.37	1.39	3.79	0.77	1	0.87
1780_3IZ5_AH								
Strand-26-29	12.91	6.99	1.54	1.66	5.45	1	1	1
Strand-45-48	13.07	13.79	0.78	0.64	5.4	0.89	1	0.94
Strand-71-72	5.9	16.11	1.27	0.27	10.22	0.83	1	0.91
Strand-93-97	16.65	14.91	1.38	1.62	5.27	0.9	1	0.95
1780_3IZ5_AI								
Strand-17-23	22.46	19.41	1.22	1.31	8.42	0.87	0.83	0.85
Strand-78-86	29.19	23.91	1.24	1.45	4.63	0.96	0.88	0.92
Strand-91-96	19.97	9.72	1.58	1.69	9.77	0.91	0.6	0.72
1780_3IZ5_AS								
Strand-78-84	22.76	16.77	1.23	1.36	5.47	0.87	1	0.93
Strand-89-95	22.53	13.75	1.73	1.93	8.31	0.89	0.67	0.81
Strand-125-129	16.2	13.52	1.33	1.39	2.52	0.89	0.75	0.81
1780_3IZ5_AT								

Strand-18-23	19.16	18.98	1.32	1.33	0.04	0.86	1	0.92
Strand-63-67	16.42	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-72-76	15.68	20.22	1.48	1.47	4.54	0.95	1	0.98
Strand-114-117	12.49	18.75	2.8	2.93	10.31	0.88	0.67	0.76
1780_3IZ5_AY								
Strand-86-89	13.48	19.65	1.22	1.19	8.1	0.88	1	0.93
Strand-115-119	16.57	8.03	1.49	1.57	8.02	0.92	1	0.96
Strand-137-139	10.24	12.51	1.51	1.59	2.27	0.96	1	0.98
Strand-207-208	5.58	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-230-231	6.03	8.78	3.33	2.95	9.64	0.95	0.5	0.73
1829_2WWL_CB								
Strand-148-152	16.31	24.35	1.23	1.14	8.04	0.85	1	0.92
Strand-165-169	16.15	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-181-187	22.99	33.93	1.74	1.6	10.94	0.82	0.83	0.83
Strand-196-202	22.32	18.27	1.33	1.34	5.77	0.9	0.83	0.87
1829_2WWQ_OA								
Strand-24-30	22.08	23.8	1.69	1.71	1.72	0.88	0.67	0.76
Strand-35-40	19.5	21.37	1.03	1.03	2.84	0.98	1	0.99
Strand-47-52	19.63	22.71	0.97	1.12	5.1	0.84	1	0.91
Strand-90-93	11.82	15.65	2.93	1.86	16.44	0.95	0.33	0.49
1829_2WWQ_SA								
Strand-6-11	17.85	31.52	2.92	1.86	14.54	0.9	0.8	0.85
Strand-69-78	31.07	35.04	1.53	1.68	8.71	0.81	0.89	0.85
Strand-10-109	32.39	38.73	1.76	1.61	6.34	0.77	0.89	0.82
1829_2WWQ_TA								
Strand-31-33	9.66	2.48	1.89	2.06	6.47	1	0.5	0.67
Strand-59-63	15.88	18.02	1.01	1.07	2.13	0.79	1	0.89
Strand-80-84	16.05	18.11	0.8	0.84	2.6	0.95	1	0.98
			Simulated Strands					
	Length True	**Length Detected**	**2-way**	**Lateral**	**Longitudinal**	**Specificity**	**Sensitivity**	**F1**
1A4i_A_SheetB								
Strand-37-43	22.95	11.3	0.33	0.37	11.19	0.4	0.67	0.78
Strand-66-72	22.48	3.29	0.7	0.76	18.65	0.96	0.33	0.49
Strand-96-99	13.14	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1A4i_A_SheetC								
Strand-5-7	8.72	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-167-171	15.63	10.98	0.79	0.98	4.02	0.89	1	0.94
Strand-191-195	16.08	3.27	1.3	1.45	12.34	0.95	0.5	0.65
Strand-211-214	12.72	9.34	0.65	0.8	2.87	0.93	1	0.96
Strand-232-235	12.52	6.06	0.55	0.69	5.83	0.95	0.67	0.78
Strand-267-269	9.64	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1A8D_SheetB								
Strand-42-45	12.71	4.54	0.46	0.56	7.53	0.92	1	0.96
Strand-70-73	12.56	9.29	0.91	1.23	2.54	0.93	1	0.96
Strand-206-209	12.45	4.44	0.58	0.7	7.25	0.93	0.67	0.78
1A8D_SheetE								
Strand-110-115	19.16	3.55	0.66	0.73	14.98	0.93	0.6	0.73
Strand-127-133	23.42	14.89	0.66	0.7	7.95	0.92	0.83	0.87
Strand-136-142	23.09	15.29	0.52	0.59	7.11	0.91	1	0.95
Strand-148-154	22.83	0.93	0.67	0.83	21.22	0.96	0.17	0.28
1A12_A_SheetA								
Strand-36-41	19.94	9.21	0.79	0.9	10.27	1	0.8	0.9
Strand-61-63	9.4	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-399-406	26.14	11.47	1.13	1.27	13.95	0.93	0.71	0.82
Strand-410-417	26.6	19.99	0.83	0.93	6.01	0.89	0.86	0.88
1A12_A_SheetB								

Strand-69-74	19.12	8.17	0.87	0.9	10.32	1	0.6	0.75
Strand-78-83	19.78	14.92	0.56	0.57	4.35	0.91	1	0.95
Strand-88-92	16.75	8.54	0.43	0.47	7.69	0.93	1	0.96
1A12_A_SheetC								
Strand-121-126	19.74	11.08	0.44	0.5	8.11	0.93	0.8	0.86
Strand-130-135	20.06	15.72	0.66	0.71	3.9	0.92	1	0.96
Strand-140-144	16.39	8.61	0.96	0.98	7.26	0.93	1	0.97
Strand-166-168	9.41	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1A12_B_SheetJ								
Strand-69-74	19.28	5.47	1.37	1.43	13.28	0.95	0.6	0.74
Strand-78-83	19.95	15.65	1.27	1.35	3.75	0.92	1	0.96
Strand-88-92	16.71	11.79	0.92	1.01	4.46	0.93	1	0.96
Strand-113-115	9.43	2.65	0.55	0.73	6.21	0.94	0.5	0.65
1AKY_SheetA								
Strand-6-11	19.97	14.22	0.56	0.62	5.24	0.91	1	0.95
Strand-31-34	12.13	6.38	0.89	1.08	5.12	0.92	1	0.96
Strand-86-89	12.84	6.38	0.89	1.08	5.12	0.92	1	0.96
Strand-114-119	18.89	12.54	0.53	0.58	6.01	0.88	1	0.94
Strand-197-201	15.97	5.55	1.26	1.44	9.89	0.93	0.75	0.83
1AOP_Sheet1								
Strand-89-91	8.83	2.19	0.85	0.95	6.04	0.96	0.5	0.66
Strand-347-351	15.82	8.92	1.18	1.37	6.11	0.92	0.75	0.83
Strand-354-361	24.85	20.82	0.59	0.67	3.59	0.86	1	0.93
Strand-389-393	16.47	13.3	0.85	0.93	2.61	0.92	1	0.96
Strand-397-404	24.8	21.78	0.85	1.18	2.44	0.89	1	0.94
1ATG_SheetB								
Strand-83-87	16.18	6.42	0.78	0.88	9	0.95	0.5	0.65
Strand-108-111	12.45	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-141-144	12.33	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-162-165	12.96	4.44	1.27	1.43	7.81	0.95	0.67	0.79
Strand-179-181	10.11	1.93	0.57	0.63	7.52	0.97	0.5	0.66
1AZO_SheetA								
Strand-77-83	22.49	17.61	0.64	0.76	4.21	0.87	1	0.93
Strand-120-126	21.93	17.71	1.11	1.42	3.6	0.88	1	0.93
Strand-141-144	13.13	2.14	1.96	2.12	10.38	0.98	0.33	0.5
1B5E_B_SheetD								
Strand-39-42	12.54	7.07	1	1.14	4.96	0.92	0.67	0.77
Strand-126-128	9.38	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Strand-151-157	28.75	18.6	1.18	1.33	9.69	0.89	0.75	0.81
Strand-160-167	25.46	18.78	1.35	1.58	6.28	0.85	0.71	0.78
Strand-208-212	15.89	11.97	0.71	0.75	3.23	0.9	1	0.95
1BUP_SheetA								
Strand-7-10	12.8	11.82	0.38	0.43	0.78	0.93	1	0.97
Strand-17-22	20.14	15.4	0.6	0.85	4.15	1	1	1
Strand-25-28	12.3	11.87	0.83	1	0.42	0.84	1	0.91
Strand-141-146	19.91	16.54	0.77	0.98	2.78	0.93	1	0.96
Strand-168-174	21.28	19.63	0.99	1.4	1.04	0.89	1	0.94
1E0M_A_Sheet1								
Strand-8-13	19.57	5.72	0.71	0.85	13.24	0.92	0.4	0.56
Strand-17-23	21.33	16.28	1.08	1.34	4.45	1	0.83	0.91
Strand-26-30	15.25	4.83	0.65	0.85	9.9	0.86	0.75	0.8
Experimental Helices								
5fKx_A_3204								
Helix-12-29	25.34	24.23	0.68	0.72	3.45	1	0.76	0.88
Helix-38-49	15.7	5.57	0.35	0.41	9.52	0.65	0.55	0.59
Helix-59-74	7.53	11.32	1.97	2.06	3.79	0.72	0.33	0.46

Helix-90-94	8.02	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-109-193	29.12	23.21	0.34	0.38	5.4	0.33	0.9	0.49
Helix-155-178	17.99	12.91	0.59	0.63	4.63	0.66	0.77	0.71
Helix-184-188	9.52	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-195-209	15.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-218-232	10.19	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-245-253	7.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-276-292	20.72	22.45	1.31	1.39	4.27	0.67	0.79	0.72
Helix-314-322	19.58	12.15	0.79	0.89	6.63	0.74	0.64	0.69
Helix-336-344	11.98	12.13	0.58	0.61	0.86	0.7	1	0.83
Helix-385-396	7.61	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-402-416	16.53	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-419-446	11.62	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-510-521	7.47	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-542-563	7.75	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-565-575	17.02	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-577-582	21.59	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-585-610	40.61	15.01	0.43	0.47	24.95	0.65	0.41	0.5
Helix-619-629	16.97	11.93	0.52	0.56	4.37	0.77	0.73	0.75
Helix-638-641	31.33	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-669-683	7.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5g4f_A_3436								
Helix-27-34	11.16	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-70-78	10.78	9.26	1.44	1.88	0.82	0.88	0.63	0.73
Helix-103-108	7.44	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-114-124	7.66	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-195-212	7.96	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-236-249	13.92	22.36	2.12	0.49	16.09	0.46	0.76	0.58
Helix-256-263	19.53	5.41	0.63	0.68	13.46	0.92	0.31	0.46
Helix-267-282	11.92	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-291-296	22.21	17.46	0.68	0.79	4.01	0.41	0.8	0.54
Helix-359-372	9.24	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-387-392	21.87	21.6	0.26	0.31	1.11	0.5	1	0.67
Helix-398-422	8.94	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-428-435	19.9	21.35	0.47	0.5	1.8	0.71	1	0.83
Helix-438-448	8.08	25.52	1.98	0.55	17.44	0.59	1	0.74
Helix-465-469	35.37	30.1	0.38	0.47	4.54	0.52	0.88	0.65
Helix-472-489	12.53	12.79	0.77	0.75	3	0.69	1	0.82
Helix-491-495	14.91	15.97	0.79	0.7	4.91	0.57	0.8	0.66
Helix-513-525	6.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-533-542	17.07	15.23	0.7	0.64	7.73	0.68	0.73	0.7
Helix-545-559	10.16	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-568-573	7.29	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-586-599	18.41	22.84	1.2	0.84	6.49	0.63	1	0.77
Helix-618-623	13.44	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-637-652	20.92	16.53	0.76	0.89	3.77	0.59	0.79	0.68
Helix-659-667	10.52	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-671-686	18.68	14.06	0.52	0.61	3.97	1	0.92	0.96
Helix-697-708	8.48	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-713-726	22.2	12.09	0.89	1.16	9.49	1	0.53	0.7
3C91_C_1733								
Helix-21-33	18.68	15.12	0.84	0.9	2.98	0.78	0.67	0.72
Helix-60-64	7.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-81-104	33.3	16.66	0.39	0.69	16.06	0.45	0.52	0.49
Helix-108-123	23.64	30.01	0.85	0.7	6.37	0.93	1	0.96
Helix-168-180	16.77	12.4	0.46	0.53	3.77	0.58	0.83	0.69
Helix-185-202	24.11	16.12	0.9	1.03	7.33	0.71	0.47	0.57

Helix-225-233	11.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5KBV_A_8232								
Helix-22-36	20.79	13.92	1.36	1.51	6.53	0.81	0.43	0.56
Helix-54-68	20.9	13.97	1.38	1.59	6.5	0.79	0.5	0.61
Helix-81-92	16.43	14.79	1.16	1.07	7.21	0.76	0.64	0.69
Helix-117-128	17.05	18.15	1.16	1.01	5.18	0.74	0.64	0.68
Helix-142-156	18.7	18.08	1.46	1.84	0.87	0.63	0.69	0.66
Helix-179-284	8.26	18.02	2.15	1.63	9.76	0.71	0.4	0.51
Helix-197-212	22.3	13.33	1.34	1.41	8.36	0.81	0.4	0.53
Helix-252-263	16.16	9.49	1.3	1.44	6.03	0.83	0.36	0.51
Helix-279-301	33.35	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-324-334	15.17	9.95	1.72	1.76	4.79	0.89	0.3	0.45
Helix-411-415	6.61	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-423-437	20.3	15.74	1.91	1.97	4.02	1	0.43	0.6
Helix-461-469	9.62	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-482-487	8.38	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-513-517	8.23	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-569-580	23.79	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-595-630	6.73	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-636-642	16.24	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-653-662	26.18	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-664-677	10.44	7.58	1.12	1.17	2.38	0.74	0.83	0.78
Helix-685-695	12.97	4.28	0.97	0.99	8.35	0.85	0.44	0.58
Helix-706-714	18.93	13.04	0.52	0.61	5.39	0.47	0.69	0.56
Helix-742-757	14.89	2.8	1.34	1.43	11.49	0.88	0.3	0.45
Helix-757-768	12.28	15.43	2.11	2.15	3.79	0.87	0.5	0.63
Helix-789-820	20.83	34.3	3.06	1.65	13.46	0.48	0.5	0.49
4UWA_A_2751								
Helix-75-88	17.37	5.11	1.07	1.18	11.65	0.69	0.38	0.49
Helix-303-313	6.27	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-383-386	7.82	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-394-421	40.98	35.88	0.62	1.44	7.2	0.48	0.81	0.6
Helix-437-453	23.75	21.6	0.59	0.83	2.06	0.68	0.88	0.77
Helix-466-496	22.4	48.31	2.91	0.93	25.91	0.34	0.93	0.5
Helix-508-531	21.47	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-534-552	10.18	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-553-592	24.82	24.08	0.48	0.53	0.73	0.64	0.94	0.76
Helix-864-890	14.74	5.96	0.89	1.07	8.33	0.84	0.5	0.63
Helix-909-935	9.95	6.65	0.62	0.56	7.1	0.79	0.71	0.75
Helix-960-973	19.6	18.31	0.4	0.8	0.56	0.63	1	0.77
Helix-978-1007	11.43	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1023-1046	22.58	15.51	1.11	1.12	12	0.73	0.44	0.55
Helix-1646-1660	34.69	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1662-1676	7.86	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1705-1725	10.14	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1806-1817	11.49	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1819-1827	10.09	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1835-1863	42.89	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1935-1948	6.87	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-1953-1977	10.24	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-2003-2020	20.86	5.46	0.64	0.7	14.91	0.84	0.43	0.57
Helix-2026-2046	20.71	8.57	0.6	0.74	11.62	0.74	0.57	0.65
Helix-2061-2074	29.79	22.61	0.6	0.7	6.69	0.58	0.75	0.66
Helix-2093-2109	15.35	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-2114-2145	11.8	4.62	0.58	0.64	6.96	0.7	0.75	0.72
Helix-2149-2173	16.06	13.59	0.77	0.83	9.66	0.55	0.32	0.41

Simulated Helices 8A								
	Length True	**Length Detected**	**2-way**	**Lateral**	**Longitudinal**	**Specificity**	**Sensitivity**	**F1**
1FLP								
Helix-4-19	22.65	22.49	0.61	0.64	0.78	0.41	0.93	0.57
Helix-21-35	20.02	20.05	0.7	0.78	2.09	0.54	0.86	0.66
Helix-37-41	5.99	8.75	0.97	0.58	4.44	0.65	1	0.58
Helix-59-76	25.89	28.67	0.35	0.36	2.78	0.41	1	0.58
Helix-82-97	22.44	23.31	0.76	0.72	2.96	0.78	0.93	0.85
Helix-103-116	19.47	24.44	0.94	0.56	7.24	0.57	0.92	0.7
Helix-124-138	20.94	24.61	0.87	0.71	3.89	0.49	0.93	0.64
1HG5								
Helix-19-30	16.37	22.42	1.09	0.31	11.39	0.58	0.91	0.71
Helix-38-50	18.27	20.8	0.56	0.27	6.86	0.39	0.92	0.55
Helix-55-67	18.52	24.13	0.83	0.38	7.08	0.54	1	0.7
Helix-71-89	25.5	19.92	0.2	0.22	5.1	0.69	0.83	0.75
Helix-89-100	15.95	11.94	0.54	0.52	6.77	0.64	0.8	0.71
Helix-114-142	42.22	32.36	0.34	0.42	9.22	0.42	0.79	0.55
Helix-160-180	31	35.03	0.66	0.44	8.12	0.55	0.85	0.66
Helix-184-188	7	5.1	1.07	1.3	1.31	0.84	0.5	0.63
Helix-190-222	47.82	51.36	0.59	0.82	3.54	0.48	0.94	0.64
Helix-228-258	46.2	45.12	0.51	0.67	2.82	0.19	0.93	0.32
Helix-260-264	6.98	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1HZ4								
Helix-4-25	31.35	27.31	0.35	0.43	3.54	0.31	0.95	0.47
Helix-27-41	21.42	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-46-66	28.7	22.09	0.53	0.66	5.92	0.71	0.79	0.75
Helix-66-84	27.26	24.35	0.3	0.34	2.32	0.42	1	0.59
Helix-86-104	27.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-106-124	26.84	24.53	0.24	0.33	1.71	0.54	0.89	0.67
Helix-130-146	24.25	17.08	0.34	0.45	6.6	0.73	0.81	0.77
Helix-148-163	22.81	16.84	0.17	0.18	5.38	0.52	0.93	0.67
Helix-171-186	22.36	16.63	0.68	0.91	5.16	0.78	0.73	0.76
Helix-187-203	24.09	17.13	0.26	0.29	6.29	0.39	0.88	0.54
Helix-208-226	27.08	23.03	0.57	0.66	3.37	0.54	0.89	0.67
Helix-228-239	16.04	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-250-264	20.95	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-266-284	27.3	20.11	0.17	0.22	6.5	0.5	0.83	0.63
Helix-286-306	28.44	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-306-325	27.63	21.05	0.2	0.22	5.84	0.61	0.79	0.69
Helix-327-332	8.72	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-333-347	20.39	11.11	1.08	1.11	9.96	0.75	0.57	0.65
Helix-351-366	22.36	13.77	0.31	0.34	7.96	0.43	0.8	0.56
1LWB								
Helix-4-12	12.17	12.42	0.31	0.31	1.91	0.97	1	0.99
Helix-16-29	19.38	15.43	0.49	0.54	3.37	0.32	1	0.48
Helix-29-37	3.07	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-57-75	27.17	22.37	0.51	0.84	4.36	0.58	0.83	0.69
Helix-76-97	32.93	23.69	0.51	0.71	8.75	0.52	0.76	0.61
Helix-100-120	28.82	28.55	1.63	2.9	6.28	0.46	0.6	0.52
3C91_H								
Helix-48-71	33.95	26.2	0.67	0.71	7.21	1	0.78	0.88
Helix-75-90	23.16	18.88	0.56	0.85	3.96	1	0.93	0.97
Helix-130-142	16.77	11.07	1.14	1.25	5.23	1	0.58	0.74
Helix-147-166	28.22	26.05	0.93	1.16	3.41	0.62	0.74	0.67
Helix-188-200	18.11	13.69	1.08	1.14	3.77	0.43	0.67	0.52
1P5X								
Helix-12-28	24.05	14.01	0.51	0.53	9.33	0.78	0.69	0.73

Helix-34-43	13.81	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-43-54	17.02	4.89	0.44	0.52	11.4	0.87	0.45	0.6
Helix-66-70	7.47	9.33	0.75	0.53	4.29	0.86	1	0.93
Helix-85-103	27.23	25.85	0.43	0.52	1.36	0.64	0.89	0.75
Helix-105-123	27.67	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-140-153	19.24	21.04	0.7	0.51	5.61	0.59	0.85	0.69
Helix-171-186	23.06	21.67	0.3	0.36	1.31	0.59	1	0.74
Helix-192-202	14.45	6.36	0.72	0.75	7.61	0.7	0.5	0.58
Helix-205-243	19.49	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1XQO								
Helix-2-15	19.17	11.28	0.2	0.28	7.26	0.57	0.77	0.65
Helix-18-27	14.4	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-29-43	19.83	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-43-57	20.98	18.43	0.38	0.45	2.04	0.68	0.93	0.79
Helix-64-78	21.38	16.53	0.21	0.24	4.17	0.57	0.93	0.7
Helix-83-95	17.84	12.14	0.43	0.49	4.99	1	0.75	0.86
Helix-101-113	17.39	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-121-134	18.67	11.75	0.33	0.42	6.28	0.65	0.69	0.67
Helix-139-158	27.93	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-172-182	14.99	12.82	0.53	0.71	1.76	0.77	1	0.87
Helix-188-195	11.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-195-211	23.65	17.43	0.43	0.47	5.63	0.5	0.88	0.64
Helix-213-229	23.49	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Helix-238-245	12.37	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 4. The dataset seen in graph 1. Three error methods used on predicted secondary structures. Helices were generated with SSETracer while β -strands were generated by an iterative surface fitting program.

4. Conclusion

It is important to have an accurate method to calculate the discrepancy of predicted secondary structures. After examining the results of three popular error methods, we have determined that arc-length association is superior due to its sensitivity. Additionally, this method provides important information about the lateral and longitudinal discrepancies. We have applied this arc-length association method to β –strands and also created an interactive tool used to visualize these discrepancies. The interactive tool was developed as a software plugin in Chimera, a popular molecular visualization tool to provide easy access to the cryo-EM community.

REFERENCES

[1] Rossmann M.G. 2000. Fitting atomic models into electron-microscopy maps. Acta Crystallogr. D Biol. Crystallogr. 56(Pt 10), 1341–1349

[2] Wriggers W., and Birmanns S. 2001. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. J. Struct. Biol. 133, 193–202

[3] Schröder G.F., Brunger A.T., and Levitt M. 2007. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15, 1630–1641

[4] Lindert S., Alexander N., Wotzel N., et al. 2012. EM-fold: De novo atomic-detail protein structure determination from medium-resolution density maps. Structure 20, 464–478

[5] Al Nasr K., Chen L., Si D., et al. 2012. Building the initial chain of the proteins through de novo modeling of the cryo-electron microscopy volume data at the medium resolutions. Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. Orlando, FL, ACM: 490–497

[6] Al Nasr K., and He J. 2016. Constrained cyclic coordinate descent for cryo-EM images at medium resolutions: Beyond the protein loop closure problem. Robotica 34, 1777–1790

[7] Al Nasr K., Ranjan D., Zubair M., et al. 2014. 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. 11, 419–429

[8] Al Nasr K., Sun W., and He J. 2010. Structure prediction for the helical skeletons detected from the low resolution protein density map. BMC Bioinform. 11(Suppl. 1), S44

[9] Baker M.L., Abeysinghe S.S., Schuh S., et al. 2011. Modeling protein structure at near atomic resolutions with Gorgon. J. Struct. Biol. 174, 360–373

[10] Baker M.L., Ju T., and Chiu W. 2007. Identification of secondary structure elements in intermediate-resolution density maps. Structure 15, 7–19

[11] Rusu M., and Wriggers W. 2012b. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. J. Struct. Biol. 177, 410–419
Si D., and He J. 2014. Tracing beta-strands using strandtwister from cryo-EM density maps at medium resolutions. Structure 22, 1665–1676

[12] Lindert S., Alexander N., Wotzel N., et al. 2012. EM-fold: De novo atomic-detail protein structure determination from medium-resolution density maps. Structure 20, 464–478

[13] Wriggers W., and He J. 2015. Numerical geometry of map and model assessment. J. Struct. Biol. 192, 255–261

[14] Zeil, S., Kovacs, J., Wriggers, W., He, J. "Comparing an Atomic Model or Structure to a Corresponding Cryo-electron Microscopy Image at the Central Axis of a Helix", *Journal of Computational Biology* , 24(1), 52-67, 2017.

[15] Si D., and He J., 2013. 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, DC, 764–770

[16] Poteat, M., He, J. "Modeling Beta-sheets using Iterative Bezier Surface Fitting on Cryo-EM Density maps" *Molecular Based Mathematical Biology*

[17] Si D., and He J. 2014. Tracing beta-strands using strandtwister from cryo-EM density maps at medium resolutions. Structure 22, 1665–1676

Devin Reid Haslam

OLD DOMINION UNIVERSITY, COMPUTER SCIENCE DEPARTMENT, NORFOLK, VA 23529

dhasl002@odu.edu   devinhaslam.com   github.com/dhasl002   linkedin.com/in/devin-haslam

Summary of Qualifications

- Experience developing deep convolutional neural networks for semantic segmentation of noisy data
- Proficient in several programming languages including C, C++, Java, and Python(Tensorflow)
- Passionate for machine learning, computer vision, and data mining

Education

Old Dominion University   *Norfolk, Virginia*

MASTER OF SCIENCE IN COMPUTER SCIENCE   *July 2018*
- Awarded the outstanding research and creativity grant in addition to a full tuition scholarship.
- **3.75 GPA**

BACHELOR OF SCIENCE IN COMPUTER SCIENCE   *May 2017*
- Received a full tuition scholarship and graduated with a **3.6 GPA** from the honors college.

Experience

Research Assistant - Old Dominion University   *Apr. 2016 - Present*

SOFTWARE ENGINEER & MACHINE LEARNING RESEARCHER
- Developed a deep convolutional neural network used to label pixels in a 3D biological image.
- Constructed a heuristic shortest path algorithm (**C**) to match secondary structures from a 3D protein image and a 2D protein sequence.
- Improved the runtime of a serial algorithm by using parallel computing (**CUDA**).
- Developed an interactive extension (**Python**) for an imaging application that displayed the results of a C++ executable.

Teaching Assistant - Old Dominion University   *Jul. 2017 - Present*

C++ CLASSROOM INSTRUCTOR
- Taught lectures, developed assignments, and graded exams for Problem Solving and Programming II.

Skills

Languages: C++, C, C#, Java, Python, Html, SQL, Javascript, IOS, R, LaTex, Matlab, Prolog, SML
Libraries: TensorFlow, CUDA, Eigen, OpenGL, ASP.NET
Other Skills: Git, Windows, Linux, Bash Scripting, P2P

Projects

Restricted Boltzmann Machine for Music Generation
- Created a Restricted Boltzmann Machine that generates a chord in the style of training music.
- Used Google's **TensorFlow** library and Gibbs Sampling to ensure that chords are constructed with randomness.

Recurrent Neural Network for Speech Recognition
- Created a recurrent neural network with **TensorFlow**, in order to recognize a spoken digit from one to ten.
- Utilized a Long short-term memory network to capture dependencies between items in a spoken word.

Publications

- Haslam, D., Zubair, M., Ranjan, D., Biswas, A., and He, J., **Challenges in Matching Secondary Structures in Cryo-EM: An Exploration**. IEEE BIBM 2016 Workshop: Computational Structural Bioinformatics Workshop. Shenzhen, China, 15-18 December 2016.
- Haslam, D., El Mesalami, A., and Ibrahim, S., **Color Restoration Survey and an Overdetermined System for Color Retrieval from Faded Images**. Journal of Image and Vision Computing. (Accepted)
- Haslam, D., Sazzed, S., and He, J. **Pattern Recognition Tools in Medium-resolution Cryo-EM Density Maps and Low-resolution Cryo-ET Density maps**. 14th International Symposium on Bioinformatics Research and Applications. Beijing, China, 8-11 June 2018
- Li, R., Si, D., Haslam, D., Ji, S., and He, J. **Detection and Evaluation of Protein Secondary Structure Patterns from 3-dimentional Cryo-Electron Microscopy Images using Deep Learning**. BioImage Informatics Conference. Banff, Canada, 19-21 September 2017.