Editor: Dana Milan

Write Up

Abstract

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).
As part of our efforts to solve the problem, we follow a deep learning approach to segment helices from good to medium resolution density maps (4 and 6 Å). In particular, we build a convolutional neural network (CNN) classifier that predicts the label (helix or non-helix) for every individual voxel in 3D cryo-EM image.
The proposed 3D convolutional neural network is shown to detect alpha helices locations with F1 score between 0.69-0.86 for four simulated test cases in 4 Å resolution and F1 score between 0.59-0.86 for four simulated test cases in 6 Å resolution.

Introduction and Motivation

Proteins perform most of the work of living cells with unique and stable three-dimensional (3D) structures, which determine their function. Cryo-electron microscopy (cryo-EM) is an experimental technique with increasing popularity to study the structures of protein complexes. Through cryo-EM, many molecular complexes, such as ribosome and viruses, have been resolved to near atomic resolutions. However, for cryo-EM density maps at medium resolutions such as 5–10Å, detailed molecular features are not resolved. It is a challenging problem to derive atomic structures from such density images, which in most cases require known atomic structures as templates. When such templates are not available, possible topologies of secondary structures can be inferred by matching of secondary structures that are detected from the 3D image and those predicted from the sequence of the protein.

The major difficulty in detecting secondary structures from images of medium resolution is that the spatial shape patterns of secondary structure elements (SSEs) at medium resolution are hard to distinguish from their narrowly located neighbors. Many methods have been developed to detect SSEs at medium resolutions, which are mostly based on image processing techniques. An α-helix is often identified using cylinder-like templates or carefully-designed cylinder-like features. In general, long α-helices can be detected rather accurately by these methods. However, short α-helices appear to be similar to other SSEs, such as turns/loops, in density images at medium resolution. Other drawbacks of these methods include carefully selected parameters and under-utilizing large amount of existing density maps in the database. Hence arises the need for more accurate method for α-helices identification in cryo-EM images at medium resolution.

Convolutional neural networks (CNNs) are a type of fully trainable models that learn a hierarchy of features through nonlinear mappings between multiple stacked layers. CNNs have been widely used in a variety of image related applications and have achieved state-of-the-art performances. Recently, attempts have been made to extend these models to the field of image segmentation, leading to improved performance. CNNs are appealing due to their ability to learn features with trainable parameters in tasks that require nonlinear relationships. Therefore, in our project we employ a CNN to segment α-helices from cryo-EM 3D density maps.

Algorithm Description

Data

We used the following 3 sets of proteins, for which we synthesized cryo-EM maps both in 4Å and 6Å resolutions:
66 proteins of EF hand superfamily.
200 proteins of Class All Alpha.
200 proteins of Class Alpha and Beta (a/b).

To deal with various sizes of proteins within the dataset, we train and test with patches of size 32x32x32.
To balance the data, we used about 1/7 of the patches that contain less than 15% positive voxels, and all the other patches.

Net

We implemented a model similar to 3D-Unet, which consists of an analysis path and a synthesis path.
The analysis path consists of four layers, each layer consists of two consecutive3x3x3 convolutions. The first convolution is followed by a relu operation. Both followed by a batch normalization. At the end of each analysis layer, a 3x3x3 max pool with a stride of two is applied to reduce the receptive field by a factor of
two. The receptive field at the end of the analysis path is now sixteen times smaller than the original input.
In the synthesis path, each layer consists of a 3x3x3 transposed convolution and a 3x3x3 convolution, both followed by batch normalization and a relu operation. We also concatenate the results of each layer in the analysis path with the results of each synthesis layer. When training the network, we employ an
Adam optimizer with 0.001 training rate.

Prediction

To predict helices in the entire protein, we split the cryo-EM map into 32x32x32 disjoint patches and use the model to predict each of them.

Experimental Results

Verification and Analysis of Correctness of Results

As the step of balancing data in our training process relies on randomity, we trained the network several times on each of the following training sets[1], randomly dividing the sets to training set and evaluation set. Our best networks achieved the following metrics upon evaluation:

Training Set	True Positive Voxels	False Positive Voxels	True Negative Voxels	False Negative Voxels	Recall	Precision	Accuracy	F1 Score
EF Hand Superfamily 4Å	989038	479515	3890963	145507	0.872	0.673	0.886	0.76
Class All alpha 4Å	2538919	1117082	16147949	774354	0.766	0.694	0.908	0.729
Class Alpha and Beta 4Å	1068840	1234163	14923193	402988	0.726	0.464	0.907	0.566
EF Hand Superfamily 6Å	1051152	341049	6254874	250013	0.808	0.755	0.925	0.781
Class All alpha 6Å	2664738	1018019	24473629	1531422	0.635	0.724	0.914	0.676
Class Alpha and Beta 6Å	911061	1183831	20802927	826213	0.524	0.435	0.915	0.475

For eight simulated whole-protein test cases synthesized with Chimera, the above-mentioned networks have achieved the following best results upon prediction:

Test Case (PDB ID)	Network trained on:	True Positives	False positives	True Negatives	False Negatives	Recall	Precision	Accuracy	F1 Score
1k40 (Class All Alpha, Four Helical Up and Down Bundle Fold, 4 Å)	EF Hand 4 Å	6909	668	276896	2047	0.771	0.912	0.991	0.836
	Class All alpha 4 Å	8176	1766	275798	780	0.913	0.822	0.991	0.865
	Class Alpha and Beta 4 Å	8755	2815	274749	201	0.978	0.757	0.989	0.853

Test Case (PDB ID)	Network trained on:	True Positives	False positives	True Negatives	False Negatives	Recall	Precision	Accuracy	F1 Score
1naf (Class All Alpha, Spectrin Repeat-like Fold, 4 Å)	EF Hand 4 Å	6954	851	504132	1975	0.779	0.891	0.995	0.831
	Class All alpha 4 Å	8035	1711	503272	894	0.9	0.825	0.995	0.861
	Class Alpha and	8738	2750	502233	191	0.979	0.761	0.994	0.856

[1] Our training sets are available at: https://github.com/NirShalmon/AlphaHelixDetection

	Beta 4 Å								
Test Case (PDB ID)	**Network trained on:**	**True Positives**	**False positives**	**True Negatives**	**False Negatives**	**Recall**	**Precision**	**Accuracy**	**F1 Score**
1jfj (Class All Alpha, EF-hand Superfamly, 4 Å)	EF Hand 4 Å	60231	18891	888865	12698	0.826	0.761	0.968	0.792
	Class All alpha 4 Å	60754	18667	889089	12175	0.833	0.765	0.969	0.798
	Class Alpha and Beta 4 Å	65373	20815	886941	7556	0.896	0.758	0.971	0.822

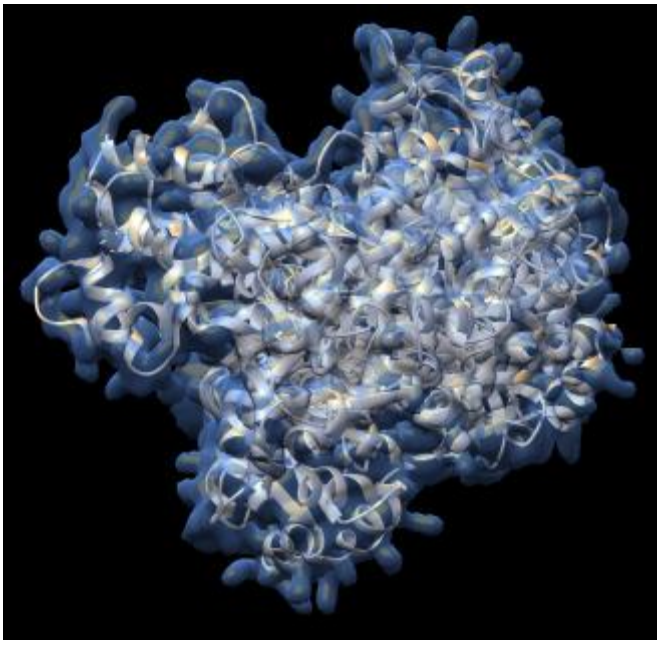
Test Case (PDB ID)	**Network trained on:**	**True Positives**	**False positives**	**True Negatives**	**False Negatives**	**Recall**	**Precision**	**Accuracy**	**F1 Score**
1o5t (Class Alpha and Beta, Adenine Nucleotide Alpha Hydrolase-like Fold, 4 Å)	EF Hand 4 Å	14787	8696	540363	4330	0.774	0.63	0.977	0.69
	Class All alpha 4 Å	17675	14274	534785	1442	0.925	0.553	0.972	0.692
	Class Alpha and Beta 4 Å	18800	18714	530345	317	0.983	0.501	0.967	0.664

Test Case (PDB ID)	**Network trained on:**	**True Positives**	**False positives**	**True Negatives**	**False Negatives**	**Recall**	**Precision**	**Accuracy**	**F1 Score**
1ll (Class All Alpha, Saposin-like Fold, 6 Å)	EF Hand 6 Å	6675	1350	356742	1522	0.814	0.832	0.992	0.823
	Class All alpha 6 Å	7356	2266	355826	841	0.897	0.764	0.992	0.826
	Class Alpha and Beta 6 Å	7131	2931	355161	1066	0.87	0.709	0.989	0.781

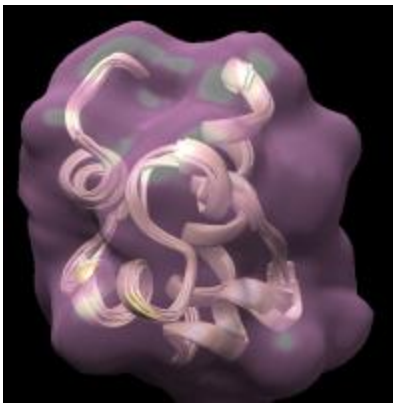
Test Case (PDB ID)	**Network trained on:**	**True Positives**	**False positives**	**True Negatives**	**False Negatives**	**Recall**	**Precision**	**Accuracy**	**F1 Score**
1sgg (Class All Alpha, SAM domain-like Fold, 6 Å)	EF Hand 6 Å	10163	2155	309712	1095	0.885	0.825	0.988	0.84
	Class All alpha 6 Å	10786	2683	309184	1095	0.908	0.8	0.988	0.851
	Class Alpha and Beta 6 Å	10785	6134	305733	1096	0.908	0.637	0.978	0.749
Test Case (PDB ID)	**Network trained on:**	**True Positives**	**False positives**	**True Negatives**	**False Negatives**	**Recall**	**Precision**	**Accuracy**	**F1 Score**
1c7v (Class All Alpha, EF-hand superfamily, 6 Å)	EF Hand 6 Å	15880	568	336734	568	0.965	0.779	0.986	0.862
	Class All alpha 6 Å	16182	5137	336115	266	0.984	0.759	0.985	0.857
	Class Alpha and Beta 6 Å	16361	10643	330609	87	0.995	0.606	0.97	0.753

Test Case (PDB ID)	Network trained on:	True Positives	False positives	True Negatives	False Negatives	Recall	Precision	Accuracy	F1 Score
1ykg (Class Alpha and Beta, Flavodoxin-like Fold, 6 Å)	EF Hand 6 Å	20752	16274	475542	648	0.97	0.56	0.967	0.71
	Class All alpha 6 Å	21015	14580	477236	385	0.982	0.59	0.971	0.737
	Class Alpha and Beta 6 Å	21075	28459	463359	325	0.984	0.425	0.944	0.594

A visual result for 1jfj, predicted by the network trained on dataset of 4 Å maps of Alpha and Beta Class: The gray area is the 3D shape of the protein. The blue area is the area predicted to be part of a helix.



A visual result for 1sgg, predicted by the network trained on dataset of 6 Å maps of All Alpha Class:



Practical Runtime Analysis

We measured the runtime of the training pipeline on various sizes of map datasets:

Number of proteins	Number of patches	Runtime (seconds)
50	273	40
100	664	91
200	1421	194
400	2809	388

As expected, the number of patches used in training and the runtime in seconds are approximately linear in the number of proteins.

Conclusions

Identification of secondary structure of proteins is challenging because of their structural similarities in 3D space. We demonstrate the use of a 3D U-Net CNN to segment α-helices in cryo-EM density maps.
We show that this version of 3D U-Net can achieve relatively good accuracy in a test of eight simulated density maps.

Future improvement of the model may include better architectures and algorithms to obtain a better segmentation accuracy.
In addition, a better balancing of the data may lead to better results, as well as a different approach for combining the predicted patches to a whole protein prediction.