Predicting protein-protein interaction interfaces using protein coevolution

Stanford | Bioengineering
Schools of Engineering & Medicine

Esha Atolia
Department of Chemical and Systems Biology

Chemical and Systems Biology

Introduction

Proteins are the building blocks of life since they make up the network of machinery that control most of the function of living cells. Therefore, understanding protein-protein interactions (PPI) is vital for describing, characterizing, and manipulating biological systems. Specifically, it is important to be able to identify inter-proteasmal interaction interfaces, i.e. the residues that are in physical contact when two proteins directly bind. Knowing this interface not only provides general mechanistic understanding, but also provides a better avenue for drug discovery. Some proteins that are implicated in disease, such as the Tau protein in Alzheimer's, can be inhibited by disrupting the protein-protein interface.

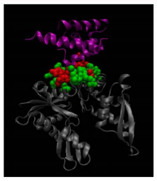


Figure 1. Co-crystal structure of MreB (gray) and RodZ (purple). Residues at the interaction interface that are at a distance of <8 Å (red) and residues that are at a distance of <12 Å (red and green)

Methods

Dataset and Features

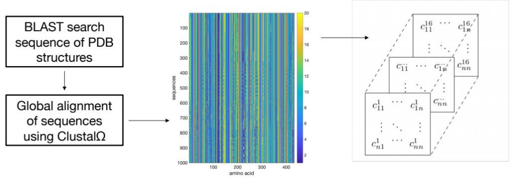
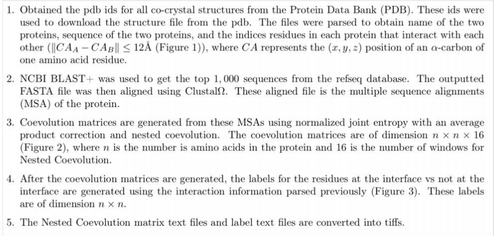

BLAST search sequence of PDB structures
Global alignment of sequences using ClustalΩ
sequences
amino acid
100
200
300
400
500
600
700
800
900
1000
2
4
6
8
10
12
14
16
18
20
c_{11}^{16} ⋯ c_{1n}^{16}
c_{nn}^{16}
$c_{11}^{\cdots}$ ⋯ $c_{1n}^{\cdots}$
$c_{nn}^{\cdots}$
c_{11}^{1} ⋯ c_{1n}^{1}
c_{n1}^{1} ⋯ c_{nn}^{1}

Figure 2. Dataset generation pipeline. The final matrix shown is the structure of the Nested Coevolution matrix. This is the input to CellUNet, a Unet based image segmentation model [1].



1. Obtained the pdb ids for all co-crystal structures from the Protein Data Bank (PDB). These ids were used to download the structure file from the pdb. The files were parsed to obtain name of the two proteins, sequence of the two proteins, and the indices residues in each protein that interact with each other ($\|CA_A - CA_B\| \leq 12$Å (Figure 1)), where CA represents the (x, y, z) position of an α-carbon of one amino acid residue.
2. NCBI BLAST+ was used to get the top $1,000$ sequences from the refseq database. The outputted FASTA file was then aligned using ClustalΩ. These aligned file is the multiple sequence alignments (MSA) of the protein.
3. Coevolution matrices are generated from these MSAs using normalized joint entropy with an average product correction and nested coevolution. The coevolution matrices are of dimension $n \times n \times 16$ (Figure 2), where n is the number is amino acids in the protein and 16 is the number of windows for Nested Coevolution.
4. After the coevolution matrices are generated, the labels for the residues at the interface vs not at the interface are generated using the interaction information parsed previously (Figure 3). These labels are of dimension $n \times n$.
5. The Nested Coevolution matrix text files and label text files are converted into tiffs.

Models

Train (4745)	Test (1,582)	Validation (1,582)

Parameters:
patch = 56
steps = 1000
epoch = 11

Dataset augmentation: flipping, rotation, and adding noise

Cross entropy loss

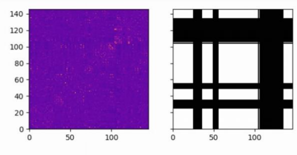


Figure 3: On the left is one layer of the NC matrix, and on the right is the label matrix where black represents the regions of the interface.

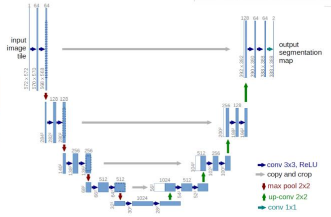
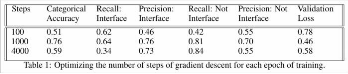


Figure 4. U-net architecture (example for 32x32pixels in the lowest resolution) [8]. Each blue boxcorresponds to a multi-channel feature map. Thenumber of channels is denoted on top of the box.The x-y-size is provided at the lower left edge ofthe box. White boxes represent copied featuremaps. The arrows denote the different operations. [2]

Results

Hyperparameter tuning

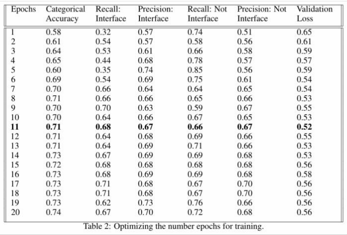


Steps	Categorical Accuracy	Recall: Interface	Precision: Interface	Recall: Not Interface	Precision: Not Interface	Validation Loss
100	0.51	0.62	0.46	0.42	0.55	0.78
1000	0.76	0.64	0.76	0.81	0.70	0.46
4000	0.59	0.34	0.73	0.84	0.55	0.58

Table 1: Optimizing the number of steps of gradient descent for each epoch of training.

Trade-off between the recall and precision

Used step = 1000



Epochs	Categorical Accuracy	Recall: Interface	Precision: Interface	Recall: Not Interface	Precision: Not Interface	Validation Loss
1	0.58	0.32	0.57	0.74	0.51	0.65
2	0.61	0.54	0.57	0.58	0.56	0.61
3	0.64	0.53	0.61	0.66	0.58	0.59
4	0.65	0.44	0.68	0.78	0.57	0.57
5	0.60	0.35	0.74	0.85	0.56	0.59
6	0.69	0.54	0.69	0.75	0.61	0.54
7	0.70	0.66	0.64	0.64	0.65	0.54
8	0.71	0.66	0.66	0.65	0.66	0.53
9	0.70	0.70	0.63	0.59	0.67	0.55
10	0.70	0.64	0.66	0.67	0.65	0.53
11	**0.71**	**0.68**	**0.67**	**0.66**	**0.67**	**0.52**
12	0.71	0.64	0.68	0.69	0.66	0.55
13	0.71	0.64	0.69	0.71	0.66	0.53
14	0.73	0.67	0.69	0.69	0.68	0.53
15	0.72	0.68	0.68	0.68	0.68	0.56
16	0.73	0.68	0.69	0.69	0.68	0.58
17	0.73	0.71	0.68	0.67	0.70	0.56
18	0.73	0.71	0.68	0.67	0.70	0.56
19	0.73	0.62	0.73	0.76	0.66	0.56
20	0.74	0.67	0.70	0.72	0.68	0.56

Table 2: Optimizing the number epochs for training.

After 11 epochs there didnt seem to be a large improvement in recall and precision and there was an increase in validation loss indicating that additional training was leading to overfitting.

Used epoch = 11

Examples of interface classification

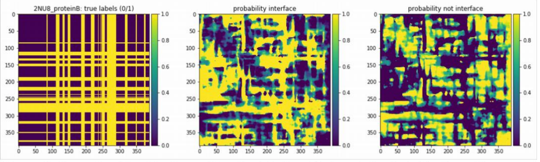


Figure 5. Example of the interface of Succinyl-CoA synthetase beta chain binding with Succinyl-CoAligase [ADP-forming] subunit alpha (PDB: 2NU8).

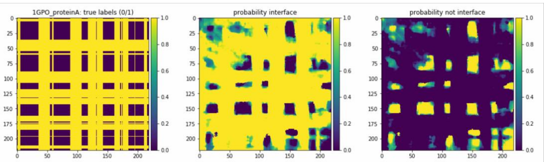


Figure 6. Example of the interface of antibody M41 dimer (PDB: 1GPO).

Conclusion and Future Directions

Conclusions

- Prediction of interacting residues for protein-protein interfaces with UNet shows accuracy on par with current state-of-the-art.
- I show that it is possible to reframe this classification problem as a image segmentation problem, which is novel for the field of protein coevolution and protein interface prediction.

Future Work

- The final label matrix is of size n-by-n but we really just need a final vector of size n to classify all the n amino acids in a protein sequence at being at the interface of not.
- To be able to do this, we can take the encoding part of UNet and instead of building the image back up add some fully connected layers at the end to get a final output of size n.
- There is novel biology to be discovered by looking at the class activation maps of what regions of the original matrix lead to the classification of each residue as at or not at the interface.

References

[1] T. Kudo. Cellunet. https://github.com/braysia/cellunet
[2] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation.CoRR, abs/1505.04597, 2015.