Protein Structure Classification Using Spatial Graph Convolutional Neural Networks

Rafael Zamora-Resendiz (LBNL), Silvia Crivelli (LBNL)

Abstract: In the field of deep learning, Convolutional Neural Networks (CNNs) have shown exceptional performance on a wide range of computer vision problems. We propose a novel approach for protein structure classification which takes advantage of graph-based convolutions to learn from graph representations of protein structures. Spatial Graph CNNs were trained on two classification tasks: 1) classifying tumor suppressor genes (TSGs) and proto-oncogenes (OGs) structures, and 2) classifying active and inactive kinase conformations. The experimental results demonstrate promising performance with a 0.95 AUC on the TSG/OG dataset and 0.985 AUC on the active/inactive Kinase dataset.

DATA & PREPROCESSING

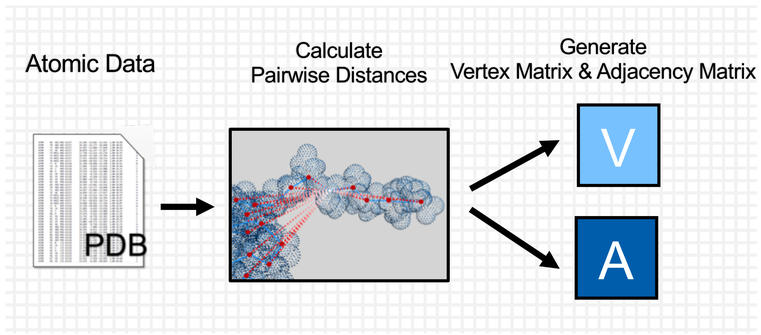

Figure 1. – High–level preprocessing overview. Atomic coordinates of decoy structures were parsed from Protein Data Bank (PDB) files. The pairwise distance between residue alpha carbons were then calculated and define the edge relationships between amino-acid nodes. Node and edge values are stored in vertex feature matricies and adjacency matrices respectively.

- **Dataset 1** contains proto-oncogene and tumor suppressor gene proteins which play major roles in the regulatory mechanisms of tumor growth.
- **Dataset 2** contains examples of kinase in both active and inactive conformation whose catalytic activity is involved in nearly all cellular processes.

Dataset 1

	TSG	OG
Training	832	834
Validation	119	119
Testing	237	238
Total	1188	1191

Dataset 2

	Active Kinase	Inactive Kinase
Training	1241	1115
Validation	177	159
Testing	355	318
Total	1773	1592

Tables 1,2 – Class composition of TSG/OG and Active/Inactive Kinase set. Datasets were divided into training, validation and testing set using a 70%/10%/20% split. Structural data used for model training has been retrieved from the RCSB PDB.

LEARNING FROM PROTEIN GRAPHS

- Each node represents a single residue within a given protein chain and is defined by a one-hot vector of all possible amino-acid types.
- Nodes are fully connected with weighted edges representing the normalized distances between residues in 3D Euclidean space.

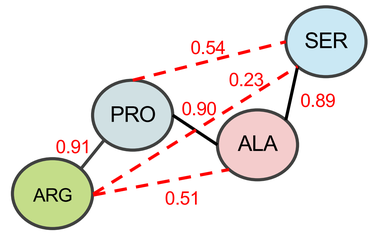

Figure 2. – Diagram of an example protein structure graph. Nodes in close proximity have edge values close to 1.0 with distant nodes having edge values approaching 0.0.

For graphs, latent node values can be learned by applying convolutional filters to all nodes in a local receptive field. Weights are assigned by treating all neighbors equally in graph receptive fields.

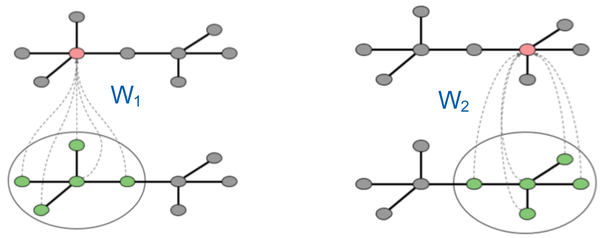

Figure 3. – Diagram of graph convolution operation convolving over a local receptive field.

Different network architectures were explored including residual blocks which have been shown to preform well on limited data. The final architecture employs 5 densely connected graph convolutional layers.

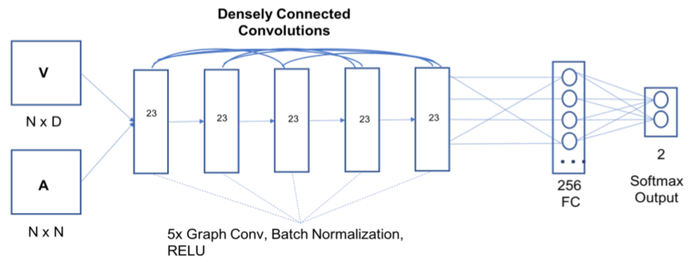

Figure 4. – Network architecture diagram for Spatial Graph CNN. Network training was conducted on LBNL's CAMERA GPU cluster using a Tesla K80 node.

RESULTS

Dataset	Network Arch.	Loss	Accuracy	AUC
TSG/OG	Res Layers	0.398	0.912	0.901
TSG/OG.	DC Layers	0.209	0.951	**0.950**
A/I Kinase	Res Layers	0.374	0.850	0.923
A/I Kinase	DC Layers	0.184	0.950	**0.985**

Table 3,4 – Performance of network on TSG/OG dataset and Active/Inactive Kinase dataset. The table includes final model loss, accuracy and AUC for densely-connected layers and residual block layers network architecture variants.

CONCLUSIONS

- We have described a method of operating on graph representations of proteins structures with explicitly defined spatial relationships between amino acids. This approach provides a more interpretable model, when compared to 3D CNNs operating on voxel representations of structures, through the localization of saliency to specific amino acids in a protein structure.

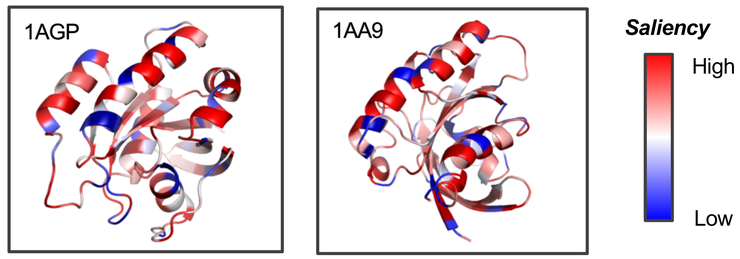

Figure 4. - Saliency of residues produced from trained TSG/OG network. High salient residues are marked in red while low salient residues are marked in blue. Proteins 1AGP and !AA9 of the OG class are presented to show similarity of saliency for structures in the same class.

ACKNOWLEDGEMENTS

This work was supported in part by the U.S. DOE WD&E Programs, by the U.S. DOE Office of Science. Special thanks to Thomas Corcoran of Johns Hopkins Applied Physics Laboratory (and current Berkeley Lab Affiliate), Dr. Xinlian Liu of Hood College (and current Berkeley Lab Affiliate) for their feedback and work on this project. Computing allocations were provided through NERSC, CAMERA, OLCF, and XSEDE.