A Novel Feature-Preserving Spatial Mapping for Deep Learning Classification of RAS Structures

Thomas Corcoran[1,2], Rafael Zamora-Resendiz[1,2], Xinlian Liu[1,2], Silvia Crivelli[2]

[1]Hood College of Frederick Maryland

[2]Lawrence Berkeley National Laboratory

OVERVIEW

A protein's 3D structure is important, since it determines its biological functionality [1]. Understanding subtle features present within protein structures is important to the fields of protein folding prediction and protein-ligand binding prediction. This work leverages the power of Convolutional Neural Networks (CNNs) to classify proteins and extract features from their 3D structures. So far, protein structural information has not been used in CNN studies because a) 3D structural information is difficult to encode into a 2D format for classical CNNs and b) 3D CNNs with deep architectures are slow to train.

Figure. 1 – Aligned HRAS (light and dark blue) and KRAS (white and orange). Dark blue and orange regions show areas of structural divergence between HRAS and KRAS. (Chimera was used for rendering)

About one third of human cancers - in particular thyroid, hepatocellular, pancreatic, lung and colorectal carcinomas - are believed to be driven by mutations in three wild-type RAS proteins: HRAS, NRAS, and KRAS [2]. For a long time, RAS had been considered non-druggable because of its relative small size and globous shape. The Frederick National Laboratory for Cancer Research of the National Cancer Institute has led a large-scale effort involving government, industrial, academic, and community research members in order to develop the methods needed to understand the complex roles that various RAS oncoproteins play in cancer.

DATASET PREPARATION PIPELINE

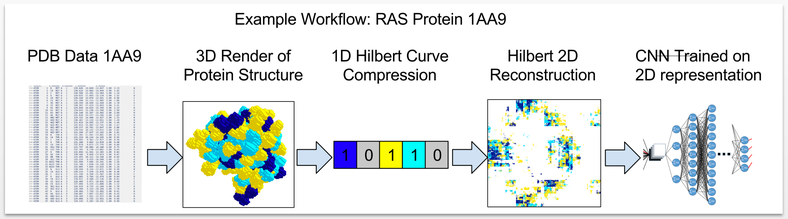
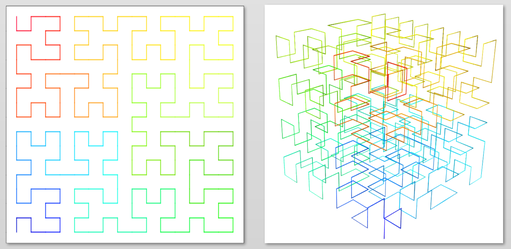


Figure. 2 – High-level data pipeline overview. 3D protein model generation and rendering, Hilbert curve-based transformation to 2D, and ingestion of 2D images by a convolutional neural network.

To generate training data, we built atom-level representations of protein structures using the centroid coordinates of each atom and their Van der Waal radii. These atomic models were then voxelized, and the resulting volumetric matrices were mapped from 3D to 1D and back up to 2D via the Hilbert space-filling curves. Encoded images contain one channel of structural information and three additional channels: the hydrophobic, polar, and charged amino acid residues of a given protein. Curve resolutions were selected so that pixels in the encoded 2D images corresponds to 1 cubic angstrom in the original 3D space. The generated set was augmented by performing random rotations on each native model.

A Space-filling curve maps N-dimensional data down to one dimension while preserving relative adjacency of data points. We selected the Hilbert curve over other such curves (e.g., Z-order) for its superior clustering properties [3]. This allowed us to preserve locality information across mappings.



Figures. 3 & 4 - Hilbert curve traversal of 2D (left) and 3D (right) space.

NETWORK ARCHITECTURES

- Networks were defined in Keras 2.0 using the Tensorflow 1.2 backend.
- 3D and 2D CNN architectures were implemented to compare systems learning our 2D vs. raw 3D data.
- Our 2D net consisted of 1 conv layer with a 5 x 5 kernel learning 64 features, followed by 12 layers of 3 x 3 kernels learning 128 features. One dense layer held 2048 neurons. All layers used RELU activations and batch norm, and some used pooling. Input images are 512x512 in resolution.
- Our 3D net was based on the VoxNet [4], and used 3 conv layers learning 32 features. 1 and 2 used 5x5x5 kernels, the last used a 3x3x3. Pooling and RELUs were used throughout. Inputs are 64x64x64 in resolution.
- The Adam optimizer was used in both nets for its superior adaptability [5].

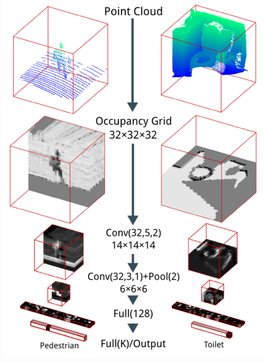


Figure. 5 – Original VoxNet design and basis for our 3D net.

DATASETS

Three distinct datasets were used to establish baseline performance on a known task and to then explore the performance of 2D vs. 3D nets on their respective structural representations:

1. ModelNet 10 Dataset: This is the benchmarking dataset for the ModelNet 3D deep learning project containing 10 classes of everyday objects in 3D CAD model format [6]. We scaled this set from its original ~5,000 example size to ~75,000 examples by applying 15 random rotations to each model. All representations of this set contain one data channel.
2. KRAS-HRAS Dataset: Two classes of proteins, KRAS and HRAS. Based upon coordinate data sourced from the Protein Data Bank. 161 HRAS chains and 79 KRAS chains were augmented with 512 random rotations and additional channel information described in the "Data Pipeline" section in order to generate 122,880 3-channel 2D images.
3. PSI-BLAST Dataset: Sourced by performing a PSI-BLAST search on sequences of KRAS and HRAS mutations, yielding a comparatively difficult classification task. The native set contained 150 non-RAS chains and 364 RAS chains, which we augmented with the same strategies to reach 263,168 3-channel 2D images.

HPC METRICS

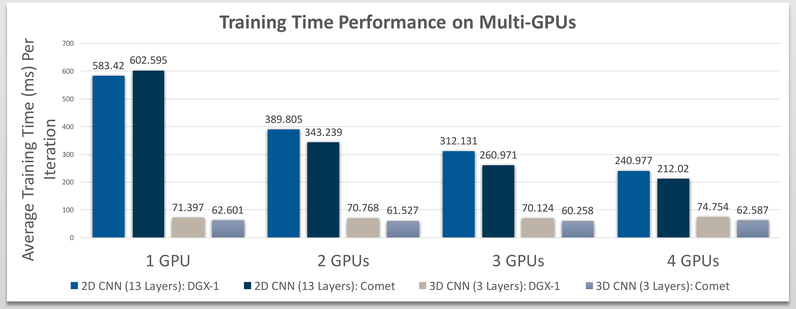


Figure. 6 – Performance variation among architectures on GPU-based systems OLCF's DGX-1 and SDSC's Comet. Bar graph shows average time to update networks on batches of 12 input images for both 3D and 2D network architectures. OLCF's DGX-1 system consists of 2 sets of 4 Tesla P100s while SDSC's Comet GPU node consists of 2 sets of 2 Tesla P100s.

RESULTS

Network Architecture	Dataset	Val Acc	Val Loss	Voted Test Acc
2D CNN *(13 Layers)*	ModelNet10 (15 rand rot)	88.10 %	0.3209	92.56%
3D CNN *(3 Layers)*	ModelNet10 (15 rand rot)	89.21%	0.3559	92.83%
2D CNN *(13 Layers)*	KRAS/HRAS (512 rand rot)	99.99 %	0.0030	100 %
3D CNN *(3 Layers)*	KRAS/HRAS (512 rand rot)	99.91 %	0.0022	100 %
2D CNN *(13 Layers)*	PSI-BLAST (512 rand rot)	93.88 %	0.5100	96.72%
3D CNN *(3 Layers)*	PSI-BLAST (512 rand rot)	85.41 %	0.6268	96.72%

Table. 1 – A comparison of the validation accuracies, the validation losses, and the final voted accuracy measures for each dataset across both 2D and 3D architectures. Voted test accuracy is computed by averaging the softmax classification outputs for each native 3D structure across a given dataset. These results show that our 3D-2D encoding process is capable of supporting 2D CNN training to accuracies matching 3D CNNs operating on raw 3D data. Note that the validation losses serve as an indirect measurement of the difficulty associated with learning from a given dataset.

FUTURE WORK

Our group is working on the development of prototypic volumetric models of cells based on bioimaging datasets, and we plan to investigate the use of CNNs for the classification of normal and diseased breast cells.

Observed differences in time costs when training 2D vs. 3D networks on 2D encodings of structural data versus native 3D structural information also deserves further study, since it appears that there may be cases where operating on 2D representations of 3D structure could be advantageous. Down-sampling of encodings have shown to allow for shallower 2D CNNs performing at comparable accuracies and faster training times when compared to 3D CNNs.

Dataset	Val Acc	Voted Test Acc	Avg Time per Iter.
KRAS/HRAS	98.46 %	100 %	33.25 ms
PSI-BLAST	86. 22%	96.72%	33.25 ms

Table. 2 – Performance of 7-Layer 2D CNN on down-sampled KRAS/HRAS and PSI-BLAST data. 512x512 images were down-sampled using bicubic interpolation to a resolution of 64x64.

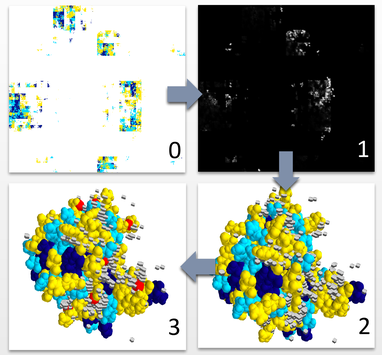


Figure. 7 – 2D structural encodings are used to train a CNN (0). Saliency maps are generated to visualize features relevant to the network (1). Saliencies are superimposed with the original 3D rendering of their corresponding protein (2). Clustering of aggregate point cloud data determines network attention masses (3). Masses may then be correlated with known binding sites and other features of interest.

ACKNOWLEDGEMENTS AND REFERENCES

This work was supported in part by the U.S. DOE WD&E Programs, by the U.S. DOE Office of Science, and by the NSF Blue Waters Project. Computing allocations were provided through NERSC, OLCF, and XSEDE.

REFERENCES:

[1] Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. 2002. Chapter 3, Protein Structure and Function.

[2] Fernández-Medarde A, Santos E. 2011. Ras in Cancer and Developmental Diseases. *Genes & Cancer*.

[3] Bongki Moon, H.V. Jagadish, Christos Faloutsos, Joel Saltz. 2001. Analysis of the Clustering Properties of the Hilbert Space-Filling Curve

[4] D. Maturana and S. Scherer. 2017. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition.

[5] Kingma, Diederik, and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

[6] Wu, Zhirong, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes.