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

TOM CORCORAN, Hood College, Lawrence Berkeley National Laboratory
RAFAEL ZAMORA-RESENDIZ, Hood College, Lawrence Berkeley National Laboratory
XINLIAN LIU, Hood College, Lawrence Berkeley National Laboratory
SILVIA CRIVELLI, Lawrence Berkeley National Laboratory

We developed a novel algorithm that leverages the power of 2D Convolutional Neural Networks (CNNs) to classify proteins by their 3D structures alone. These structures are mapped to 1D using a space-filling curve before being mapped into 2D images via the application of a complementary curve. Several data encoding and neural network design strategies were explored and evaluated on different GPU-based supercomputers as part of this effort. The results show the effectiveness of the algorithm by correctly classifying encoded elements of three different datasets.

CCS Concepts: Optimization algorithms; Molecular structural biology; Supervised learning by classification

KEYWORDS

Ras, KRas, Space-filling Curves, Deep Learning, ModelNet, 3D CNN

1 INTRODUCTION

Understanding protein structure is important because a protein's 3D structure determines its functionality [1]. This work leverages the power of 2D CNNs to classify proteins and extract features from their 3D structures directly, unlike other recently-proposed deep learning methods which rely on the primary or secondary structures of a protein only [2]. So far, 3D structural protein information has only rarely been used as training data for CNN studies. However, we have designed, implemented, and validated a novel method that maps 3D protein structure renderings to 2D data grids as a preprocessing step for efficient 2D CNN ingestion and training. Our experiments focused primarily on the RAS protein family because it is well-studied and has been linked to various forms of human cancer [3]. Our results show that neural networks trained on our 2D structural encodings can successfully distinguish between two homologous branches within the RAS family, HRAS and KRAS, which are very similar in both sequence and structure [4]. Furthermore, our encoding process also supports training networks on the 10-class ModelNet10 3D object classification dataset, as well as on a more difficult protein-based dataset derived from a PSI-BLAST search on sequences of KRAS and HRAS mutations, a comparatively difficult classification task.

2 EXPERIMENTAL AND COMPUTATIONAL DETAILS

2.1 2D Data Processing Pipeline and Neural Network Architecture

Our data processing pipeline starts with van der Waals radii-based representations of 3D RAS protein structures obtained from the RCSB Protein Data Bank [5]. We voxelized each model to produce a discretized representation. Because the number of available structures in the RAS family is small for the purpose of training a CNN, we perform 512 random rotations of each 3D protein model in order to augment the data and provide multiple views of each protein in our training and testing sets. We applied Hilbert curves to convert the 3D models to 512x512 pixel 2D images [6]. The resulting 2D data grids are encoded with three additional channels of residue properties (hydrophobic, hydrophilic, charged) as binary values (in RGB channels).

The base CNN architecture used in our experiments has 13 convolutional layers, the first of which uses a 5x5 kernel size learning 64 features, with the remaining layers using 3x3 kernels and 128 features. A final dense layer with 2048 neurons and a .5 dropout rate feeds into a softmax output layer. ReLU activations

and batch normalization are used throughout the network, and max pooling operations are performed to gradually reduce the intended 512x512 pixel input image sizes down to 4x4 by the last layer. All networks were implemented using the Keras API on top of the TensorFlow 1.2 back-end library.

We also made use of 3D CNNs in order to compare performance of 2D networks trained upon data generated through our novel encoding process versus volumetric networks operating on raw 3D data. The 3D CNN architecture we used was based upon the VoxNet architecture [7], and made use of 3 convolutional layers learning 32 features each, with kernel sizes of 5x5, 5x5, and 3x3, respectively. ReLU activations and pooling were used throughout the network. For both 2D and 3D architectures the Adam optimizer was used.

2.2 Computing Systems and Training Performance

We used two different supercomputing systems for the network training phase of our work, XSEDE's Comet, which contains nodes with 2 sets of 2 Nvidia P100 GPUs, and ORNL's DGX-1, which contains 2 sets of 4 Nvidia P100s. The performance of our networks on the DGX system is summarized below.

Table 1. KRAS vs. HRAS Training Performance Across Network and System Architectures

Network Architecture	Voted Accuracy	Batch Time 1 GPU	Batch Time 2 GPUS	Batch Time 3 GPUs	Batch Time 4 GPUs
2D CNN	100%	583.42 ms	389.81 ms	312.13 ms	240.98 ms
3D CNN	100%	71.39 ms	70.77 ms	70.12 ms	74.75 ms

3 RESULTS: CLASSIFICATION OF HRAS AND KRAS

Our experimental results successfully validated the effectiveness of our mapping algorithm via correct classification of KRAS and HRAS protein models, for which a ~123,000 element training and test set was prepared using 72 KRAS and 156 HRAS chains sourced from the RCSB Data Bank. By mapping the 2D attention maps from our trained networks back into the original 3D protein model space and superimposing them with their corresponding 3D models, we observed some structurally and biologically-relevant segments, including the active binding site, which seems to be the focus of the network when it is making its classifications. This result demonstrates the feature-preserving nature of our 3D-2D mapping algorithm and highlights the value this approach holds for biomedical research, potentially as a method useful for rapid exploration of complex structural feature spaces. Furthermore, our tests show that networks trained on 2D encoded versions of the PSI-BLAST search result set, as well as 2D encoded versions of the ModelNet10 3D model set, are also able to classify strongly, although these results are not included here for space reasons.

4 CONCLUSIONS

This project has provided a new vector for the deep learning of high-dimensional data. Because the attention maps of our networks highlight areas of interest during classification, we hope to find common structural features that are unique to a group or family of proteins sharing a similar function. This information can be useful to better understand RAS signaling in cancer progression. For example, some mutations in RAS genes are considered to be higher-risk than others [8]. It will be of great clinical interest if we can identify structural and functional differences between such mutations or mutations of lower risk and wild-types of RAS. Furthermore, if a structure with a higher-risk mutation can be re-classified by the neural network to a lower-risk class after binding with an inhibitor drug, that will be indicative that the drug works. Indications that networks training on our 2D encodings may outperform their 3D counterparts in terms of training time while retaining accuracy when image sizes are down-sampled are also deserving of investigation.

ACKNOWLEDGMENTS

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 under the Student Internship Program. Computing allocations were provided through NERSC, OLCF, and XSEDE.

REFERENCES

[1] Jeremy M. Berg, John L. Tymoczko, Lubert Stryer. 2002. *Biochemistry* (5th ed.). W.H. Freeman, New York, NY.
[2] Jie Hou, Badri Adhikari, Jianlin Cheng. 2017. DeepSF: deep convolutional neural network for mapping protein sequences to folds. Retrieved August 1, 2017 from https://arxiv.org/abs/1706.01010.
[3] A. Fernández-Medarde, E. Santos. Ras in Cancer and Developmental Diseases. *Genes & Cancer*, 2, 3 (2011), 344-358. DOI:10.1177/1947601911411084
[4] E. Castellano, E. Santos. Functional Specificity of Ras Isoforms: So Similar but So Different. *Genes & Cancer*, 2, 3 (2011), 216-231. DOI:10.1177/1947601911408081
[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.
[6] Bongki Moon, H.V. Jagadish, Christos Faloutsos, Joel Saltz. Analysis of the Clustering Properties of the Hilbert Space-Filling Curve. *IEEE Transactions on Knowledge and Data Engineering*, 13, 1 (January / February 2001).
[7] D. Maturana and S. Scherer. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. IROS2015. Retrieved 1 August, 2017 from http://www.dimatura.net/publications/voxnet_maturana_scherer_iros15.pdf.
[8] Rebecca Kirk. Nature Reviews Clinical Oncology. *Nature Reviews*, 8, 1 (January 2011). DOI:10.1038/nrclinonc.2010.204.

Received August 2017