UNIVERSITY OF DELAWARE
DATA SCIENCE INSTITUTE

Predicting novel kinase-substrate interactions using Graph Representation Learning

Sachin Gavali[1], Chuming Chen[1], Julie Cowart[1], Karen Ross[2], Cathy Wu[1,2]

1.Data Science Institute, University of Delaware. Newark DE, 19711, USA
2. Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington DC, 20057, USA

CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

Background

The human kinome consists of 518 protein kinases which play a key role in regulating various biological processes through post-translational modifications (PTM). They have been proven to be immensely useful as a therapeutic target in the development of numerous drugs. Hence, our lab at University of Delaware has developed the iPTMnet knowledge base that combines the information about kinases, substrates, and the corresponding PTM from text mining tools, curated databases, and ontologies[1]. Unfortunately, many of the substrates in iPTMnet have limited information about their corresponding kinases. Identifying new kinase-substrate interactions requires an experimental approach which is time-consuming and expensive. To further exacerbate this, most of the biomedical literature is focused on a limited set of kinases leading to an insufficient understanding of the human kinases. Hence, to help understand these understudied kinases, we have developed a graph powered machine learning model to predict novel interactions between dark kinases and existing well-studied substrates.

Objective

To develop a machine learning model that can exploit existing latent interactions among the proteins in iPTMnet to predict novel kinase-substrate interactions.

Results

The resulting model achieved an F1 score of 0.79 and an AU-ROC of 84% [Figure 2]. It was able to predict 712 novel kinase-substrate relationships among the proteins in iPTMnet. To further evaluate the biological feasibility of the model for studying dark kinases we utilized it to identify kinases for 85 substrates from the “Illuminating the Druggable Genome” project[6]. The model was able to identify novel kinases for 12 substrates.

Acknowledgment

We would like to acknowledge and thank the Data Science Institute at University of Delaware for granting access to the computing resources on the Caviness High Performance Computing Cluster that made this work possible. We would also like to thank the developers of scikit-learn and StellarGraph python libraries for providing well tested implementations of machine learning algorithms used in this work.

References

1. H. Huang et al., “iPTMnet: an integrated resource for protein post-translational modification network discovery,” Nucleic Acids Res., vol. 46, no. D1, pp. D542–D550, Jan. 2018, doi: 10.1093/nar/gkx1104.
2. A. Grover and J. Leskovec, “node2vec: Scalable Feature Learning for Networks,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, Aug. 2016, pp. 855–864, doi: 10.1145/2939672.2939754.

Methods

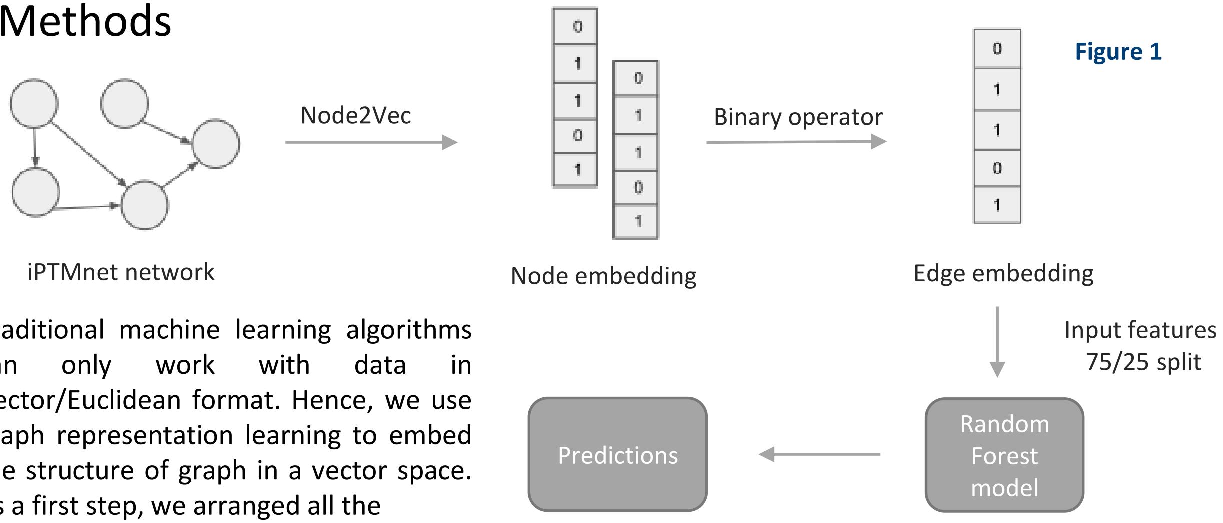
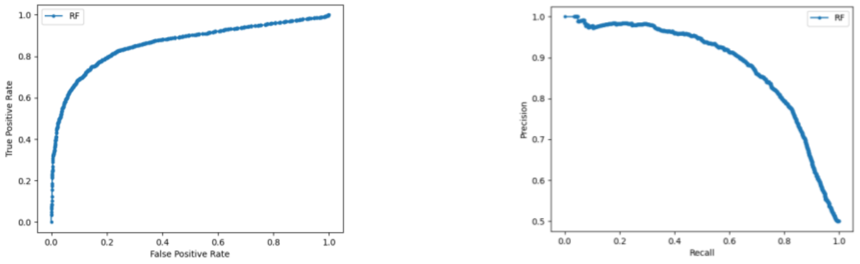


Figure 1

Traditional machine learning algorithms can only work with data in Vector/Euclidean format. Hence, we use graph representation learning to embed the structure of graph in a vector space. As a first step, we arranged all the proteins in iPTMnet in the form of a network of interacting proteins. The resulting network had 3,419 nodes i.e proteins and 8,591 edges i.e interactions. In the second step, a vector representation of the nodes was obtained using the node2vec algorithm[2]. In the third step, a binary operator was applied over the node-vector representations to obtain a vector representation of the edges connecting these nodes. These edge-vector representations were then used as input features to train a random forest (RF) model. The model was trained and evaluated using a three-way cross-validation-test approach. In this approach, 75% of the input edges where used for hyper-parameter tuning, cross-validation and training, and the remaining 25% of the edges where used for testing the resultant model.



ROC-CURVE (AU-ROC)

PR-CURVE (F1)

Figure 2