Exploring the human kinome using graph representation learning

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

BACKGROUND

- The human proteome contains a vast network of interacting kinases and substrates.
- Some of the kinases in this network have proven to be immensely useful as therapeutic targets, but many are still understudied.
- In this work, we hypothesized that identifying novel interactions for understudied kinases based on their interactions with well studied kinases would provide valuable insights into their biological significance.
- Hence, we use graph representation learning to build a model to predict novel kinases-substrate interactions and then perform a bioinformatic analysis of the predicted interactions to understand the biological role of understudied kinases.

DATASET

- We construct a PTM enzyme-substrate interaction network by filtering for PTM events in humans using NCBI taxon code – 9606 in iPTMnet. The resulting network [Figure 1] had 3033 nodes (proteins) and 8315 edges (interactions).
- Since biological interactions do not happen in isolation, we also use the protein annotation data from Gene Ontology [2] – A controlled hierarchical vocabulary of terms describing molecular functions, biological processes, and cellular components.
- Finally, to test our model, we use the list of understudied kinases from Illuminating the Druggable Genome Project (IDG) [3].

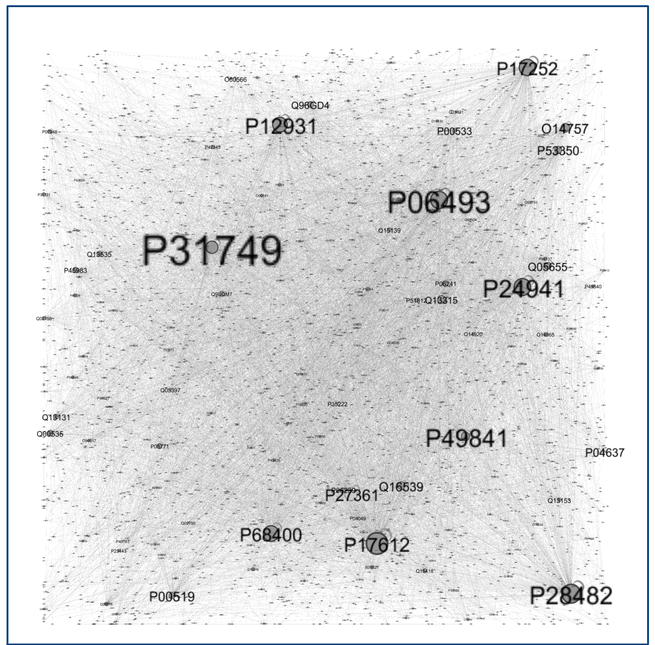
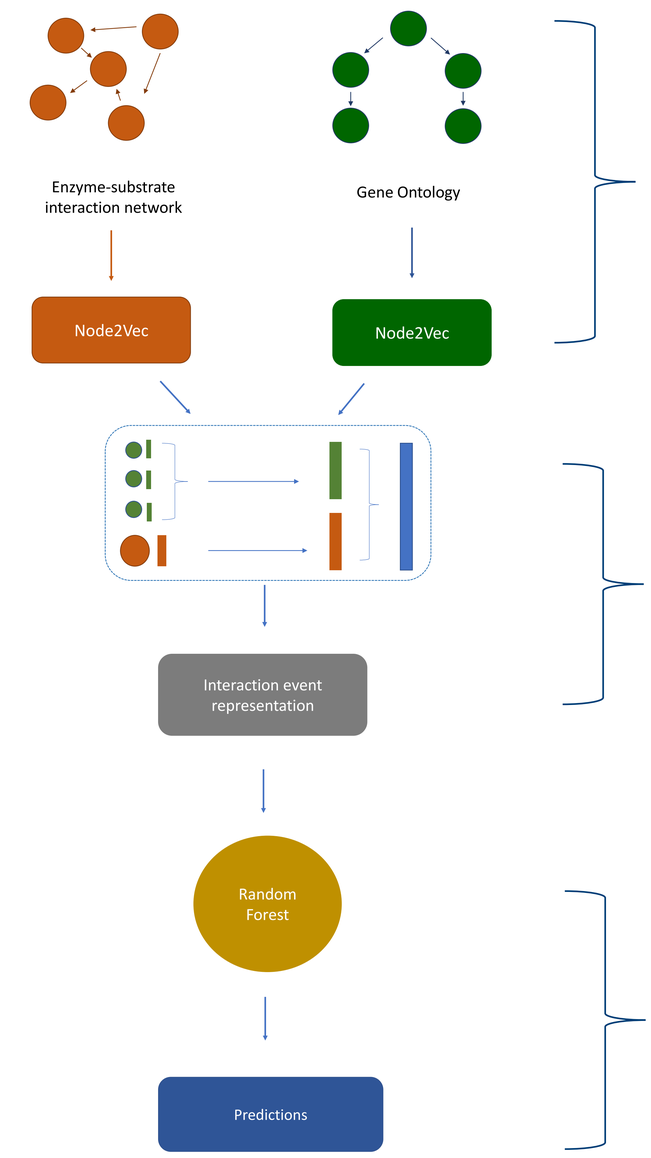


FIGURE 1. IPTMNET ENZYME-SUBSTRATE NETWORK

METHODS



Data preparation and graph representation learning

- In the first step we construct an enzyme-substrate interaction network from iPTMnet by including the PTM data for human PTM events.
- To provide a biological context to these interactions we also include the data from Gene Ontology (GO) network.
- Then we use the node2vec [4] algorithm to learn vector representations of the kinases and substrates in the interaction network and GO terms in gene ontology network.

Feature construction

- Since a protein is annotated with multiple GO terms, we condense GO terms annotated to a protein into a single vector by recursively applying Hadamard operator.
- We then concatenated the unified GO terms vector (G) with a vector (P) encoding the protein's representation in enzyme-substrate interaction network to obtain a single feature vector (F).
- To obtain a vector representation of an interaction event (E), we concatenate the feature vectors (F) of the involved proteins.
- We then use the interaction event vector (E) as an input to a random forest model to predict the probability of the event being true or false.

Model Training

- Before we started training the model, we filtered out all the experimentally validated interactions for the IDG kinases from the input dataset to serve as the testing set.
- We then split the remaining dataset into a training (75%) and a validation set (25%). We evaluate the model using AU-ROC and AU-PR on the testing set.
- Once the model is trained, we use it to predict potential interactions involving the IDG kinases and compare our predictions to the experimentally validated edges filtered out earlier to determine the final model performance.

RESULTS

- Our model achieved an AU-ROC of 0.76 and AU-PR of 0.80.
- After validating the model's predictive performance, we studied the highest confidence predictions made by the model for the proteins with the least amount of information.
- We filtered the IDG kinases to include only the kinases that had at least one known interaction in the enzyme-substrate interaction network. We then sorted them according to the number of predicted interactions at a positive probability threshold of 0.8 in a descending order.
- This gave us two candidate kinases - Q9UEE5 (STK17A) and Q9H422 (HIPK3) with 22 and 15 predicted substrates as interaction partners, respectively. We then performed a bioinformatics analysis using STRINGS DB, KEGG and Reactome and discovered that Q9UEE5 (STK17A) might play an important role in cancer activity and Q9H422 (HIPK3) might play an important role in mediating an inflammatory immune response.

Acknowledgement

We would like to acknowledge and thank the Data Science Institute and Center 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 pytorch-geometric python libraries for providing high quality and 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] Consortium, T. G. O. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research 49, D1 (12 2020), D325–D334

[3] Rodgers, G., Austin, C., Anderson, J., Pawlyk, A., Colvis, C., Margolis, R., and Baker, J. Glimmers in illuminating the druggable genome. Nature Reviews Drug Discovery 17, 5 (2018), 301–302.

[4] 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.