Dense Vector Representations of Proteins Based on Gene Ontology Annotations Using Word and Sentence Embedding Tools

JAMES ZHANG[1], Chelsea Ju[2], Muhao Chen[2], Dat Duong[2], Yunsheng Bai[2], Wei Wang[2]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA

2 Department of Computer Science, UCLA

1. Introduction

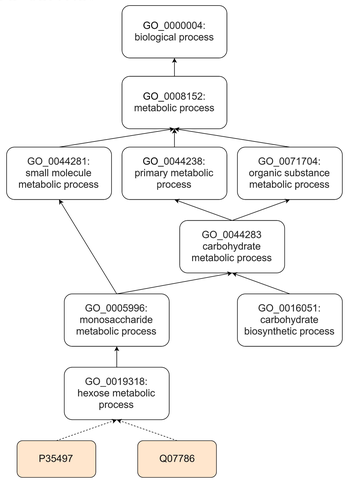
Recent advances of machine learning algorithms and accumulation of large datasets make possible a wide assortment of new predictive tasks in biological and medical settings. In particular, accurate prediction of protein-protein interactions (PPIs) is vital to our understanding of mechanistic pathways within the human body as well as drug design.

However, in order to be analyzed via computational means, proteins must first be embedded into numerical vectors. Recently, the Gene Ontology (GO), a database of biological terms arranged in a hierarchical graph structure, has been considered a potential source of extracting these vector representations of proteins.

Our code is available at our GitHub repository:
https://github.com/jameszha/ProteinEmbedding

2. The Gene Ontology

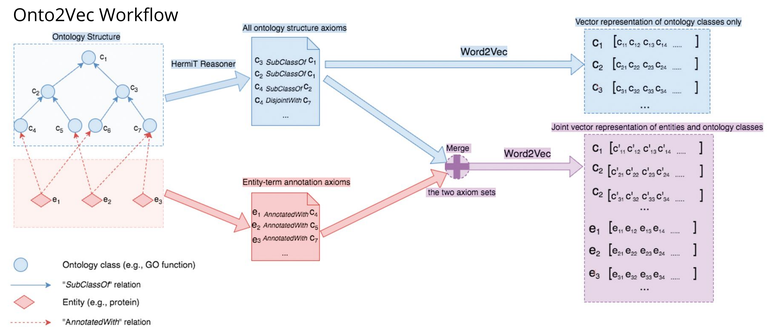
- The GO database is comprised of nearly 50,000 GO terms that each describe biological functions of genes.
- Proteins may be annotated to one or more of these GO terms.



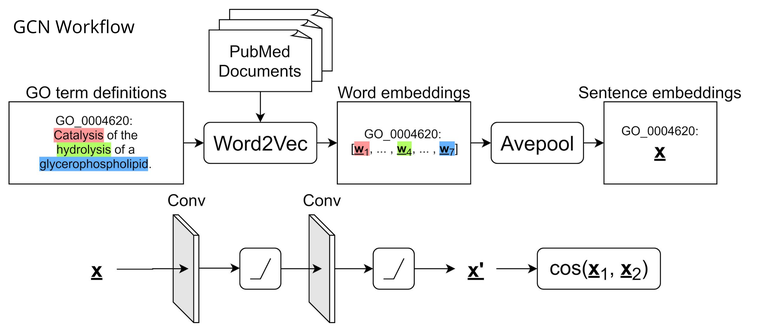
3. Methods

- As a baseline method, sparse binary embeddings were generated directly from the protein annotations.
- Two methods, Onto2Vec[4] and GCN[2], were used to generate dense embeddings of dimension 200.
- The embeddings were evaluated in three different tasks:
  - Enzyme Commission (EC) number classification
  - binary PPI prediction
  - PPI type classification
- We used four classifiers from the scikit-learn Python package:
  - Logistic Regression
  - Random Forest (default sklearn hyperparameters)
  - SVM (linear kernel, C=1)
  - MLP (hidden layers of 800, 200)

Onto2Vec Workflow



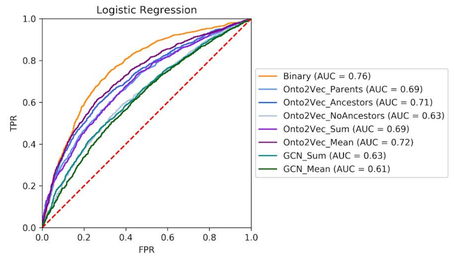
GCN Workflow

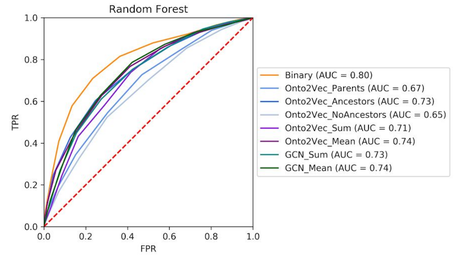


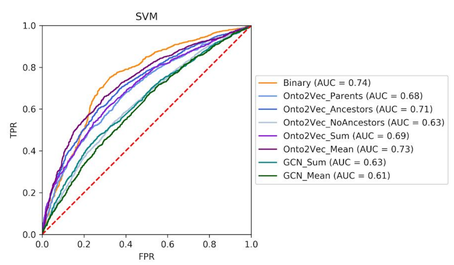
4. Results

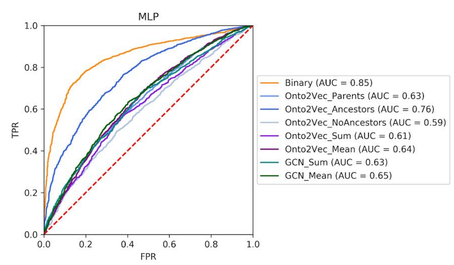
- The binary embeddings generated were of dimension 5609.
- From the datasets used, we generated 938 training samples and 402 test samples for EC classification, 7518 training samples and 3222 test samples for binary PPI prediction, and 7662 training samples and 3284 test samples for PPI types classification.
- For classification into top-level EC categories, Onto2Vec embeddings with annotation propagation up the entire GO performed very well, achieving ROC AUC of 0.92 using a MLP classifier.
- The dense embeddings were less successful in protein-protein interaction tasks, performing worse than the binary embeddings.
- In general, protein embeddings generated from taking the sum or mean of GO term embeddings to which they are annotated performed the most poorly. As a result, the GCN methods were largely unsuccessful.

Binary Protein-Protein Interaction

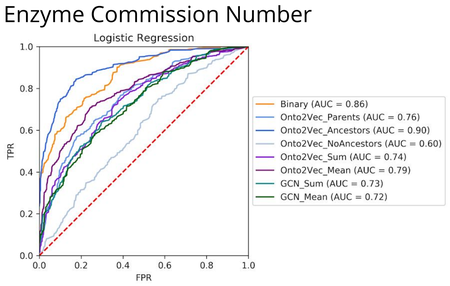


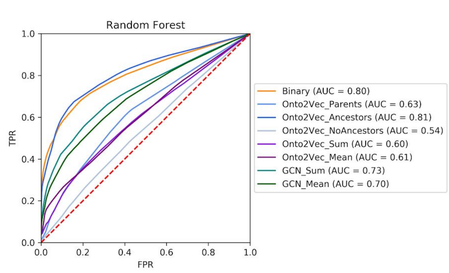


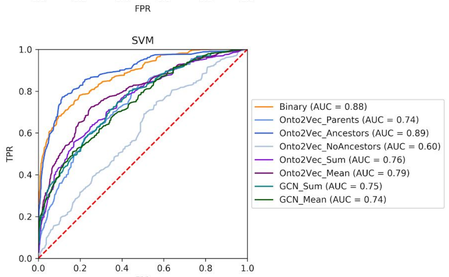


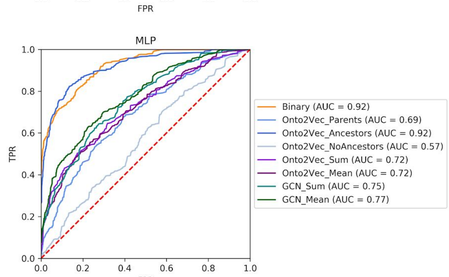


Enzyme Commission Number

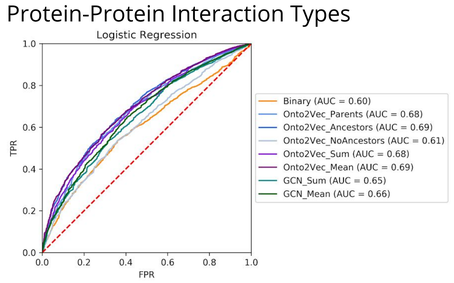


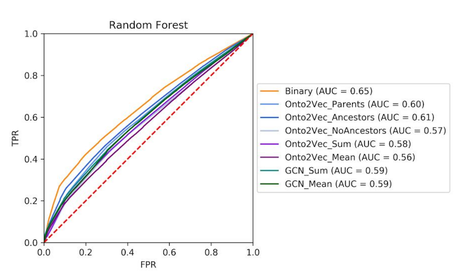


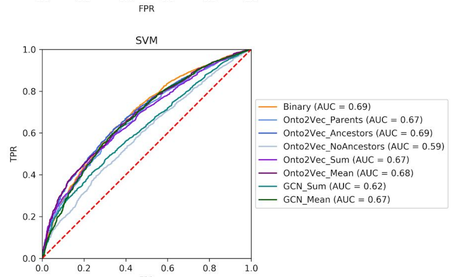


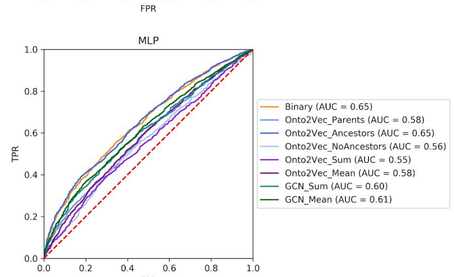


Protein-Protein Interaction Types









5. Discussion

- We hypothesize that dense embeddings may show promise in situations of data scarcity. Due to the high dimensionality of the binary embeddings, a similarly high training sample set size is required. To expand on this project, classifier performance may be examined at varying training set sizes.
- Two advantages of dense embeddings we observed are storage requirements and classifier runtime. Higher dimensionality naturally requires proportionally greater storage space and dramatically increases runtime.
- Future work may involve applying GCN methods to a supergraph containing the GO tree along with protein nodes and annotation edges. Methods like GraphSAGE, which produces node-level embeddings, may serve to yield protein embeddings directly, without need for taking the sum or mean of GO term embeddings.

6. References

1. Chen, M., Ju, C. J., Zhou, G., Zhang, T., Chen, X., Chang, K., . . . Wang, W. (2018). Lasagna: Multifaceted Protein-Protein Interaction Prediction Based on Siamese Residual RCNN. doi:10.1101/501791
2. Duong, D., Ahmad, W. U., Eskin, E., Chang, K., & Li, J. J. (2017). Word and sentence embedding tools to measure semantic similarity of Gene Ontology terms by their definitions. doi:10.1101/103648
3. Hamilton, W. L., Ying, R., & Leskovec, J. (2018). Inductive Representation Learning on Large Graphs. Doi:1706.02216
4. Smaili, F. Z., Gao, X., & Hoehndorf, R. (2018). Onto2Vec: Joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics, 34(13), I52-I60. doi:10.1093/bioinformatics/bty259