DBEI DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

DBEI & CCEB RESEARCH DAY #2022ResearchDay

HiG2Vec: hierarchical representations of Gene Ontology and genes in the Poincaré ball

Jaesik Kim, MS[1,2,3], Dokyoon Kim, PhD[1,2], Kyung-Ah Sohn, PhD[3,4]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA; [2]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA; [3]Department of Computer Engineering, Ajou University, Suwon 16499, South Korea; [4]Department of Artificial Intelligence, Ajou University, Suwon 16499, South Korea

Perelman School of Medicine University of Pennsylvania

AJOU UNIVERSITY

Background

- Knowledge manipulation of Gene Ontology (GO) and Gene Ontology Annotation (GOA) can be done primarily by using vector representation of GO terms and genes.
- Previous studies (Onto2Vec[1], OPA2Vec[2], EL embeddings[3], etc.) have represented GO terms and genes or gene products in Euclidean space to measure their semantic similarity using an embedding method such as the Word2Vec-based method to represent entities as numeric vectors.
- However, this method has the limitation that embedding large graph-structured data in the Euclidean space cannot prevent a loss of information of latent hierarchies, thus precluding the semantics of GO and GOA from being captured optimally.
- On the other hand, hyperbolic spaces such as the Poincaré balls are more suitable for modeling hierarchies, as they have a geometric property in which the distance increases exponentially as it nears the boundary because of negative curvature.
- In this study, we propose hierarchical representations of GO and genes (HiG2Vec) by applying Poincaré embedding[4] specialized in the representation of hierarchy through a two-step procedure: GO embedding and gene embedding.

Method

- HiG2Vec proceeds in two steps. First, the model learns representations of GO terms from the GO corpus using Poincaré embedding, and then, it learns representations of genes through fine-tuning using the GOA corpus.
- As a validation of the results, we performed a total of six experiments at both the GO level and the gene level.
- We evaluated the quality of embeddings by **link reconstruction** and **hierarchy reconstruction** and **reconstruct the level of GO terms** in the GO hierarchy using a neural network.
- By using gene embeddings, we predicted **interactions between two genes or gene products** from the STRING and HumanNet v2 databases. Specifically, three kinds of information from interactions were predicted: **existence, score and type**.

Result

- HiG2Vec showed better performance on GO link prediction, GO hierarchy reconstruction, and GO-level reconstruction (Figure 1), and gene-gene interaction predictions: binary, score, type (Figure 2).
- Through experiments, we show that our model represents the hierarchical structure better than other approaches and predicts the interaction of genes or gene products similar to or better than previous studies.
- The results indicate that our proposed method captures the GO semantics and gene semantics from GO and GOA better than other methods because of the hierarchical property of GO.

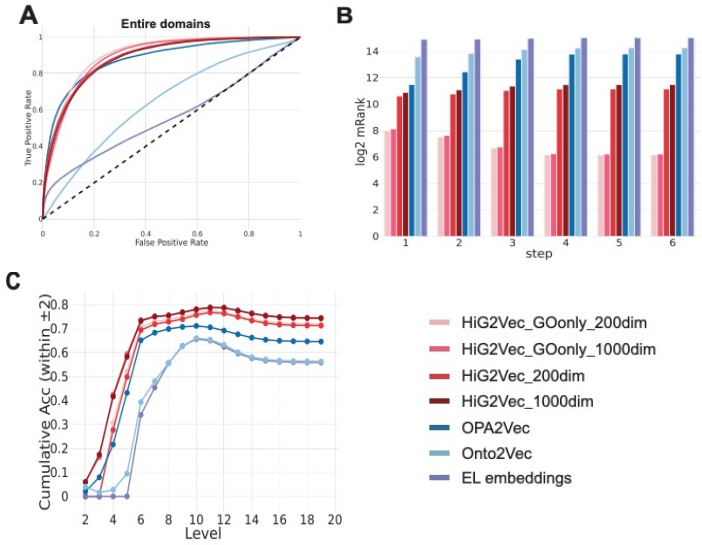

Figure 1. (A) A result of GO link reconstruction for human embeddings: ROC curve. (B) Results of GO hierarchy reconstruction for human GO embeddings: log2 transformed mRank when reconstructing within n-step reachable nodes. (C) Results of GO-level reconstruction for human GO embeddings: cumulative prediction accuracy (allowed within 2 levels as correct)

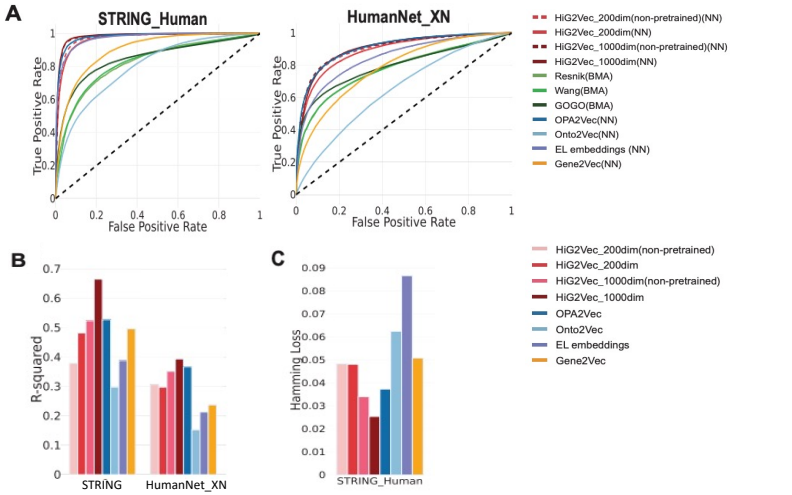

Figure 2. (A) Binary interaction prediction: ROC curves for humans using STRING_Human and HumanNet_XN. (B) Interaction score prediction: the R-squared using STRING_Human and HumanNet_XN. (C) Interaction type prediction using the STRING database: the Hamming loss

Conclusion

- In this study, we proposed a novel embedding model from GO and GOA by adopting representation learning that specializes in the GO hierarchy. The vector representation of GO terms and genes in HiG2Vec is expected to be used for various downstream analyses such as data processing at the gene and protein levels or manipulation of biological knowledge.

References

[1] Smaili,F.Z. et al. (2018a) Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics, 34, i52–i60.
[2] Smaili,F.Z. et al. (2018b) OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics, 35, 2133–2140.
[3] Kulmanov,M. et al. (2019) El embeddings: geometric construction of models for the description logic el þþ. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, fIJCAIg 2019, Macao, China, August 10-16, 2019. pp 6103-6109.
[4] Nickel,M. and Kiela,D. (2017) Poincare´ embeddings for learning hierarchical representations. In: Guyon,I. et al. (eds.) Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., Red Hook, NY, pp. 6338–6347

Published in *Bioinformatics*, Volume 37, Issue 18, 15 September 2021, Pages 2971–2980, https://doi.org/10.1093/bioinformatics/btab193