Title: Discovering Novel Therapeutic Targets Using Large Biomedical Knowledge Graphs

Authors:

Arwa Raies[1,2], Andrea Pierleoni[3], Oliver Stegle[2,4], Ian Dunham[1,2]

[1] European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

[2] Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

[3] Healx Ltd, Cambridge, United Kingdom

[4] European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany

Abstract:

One of the main challenges in drug discovery is the high attrition rate late in development. Lack of efficacy is one of the main reasons for failures in trial, and more than 79% of drugs at the clinical development stage fail because they are ineffective. However, drugs with genetic validation data are more likely to succeed clinically. Therefore, using the increased amount of human genetic data available today to prioritise the best drug targets can improve the success rate of new drugs development. Particularly, machine learning (ML) approaches are useful to process large volumes of biological data, detect patterns that may not be identified by human insight, and suggest hypotheses of potentially novel target-diseases associations. The goal of the study is discovering new therapeutic targets using ML approaches to accelerate the process of developing new or repurposing existing drugs.

In this study, we apply knowledge graph analysis to exploit heterogeneous data to find novel targets. Nodes in the knowledge graph represent biological entities (e.g., targets, diseases, and drugs), and edges represent their relationships. The graph incorporates data from the Open Targets Platform, which is a database that integrates public domain data to enable target identification and prioritisation, and LINK Platform, which includes associations between targets, drugs, diseases, and biomedical concepts extracted from scientific literature. We also incorporate data from ChEMBL, UniProt, and IntAct. The resulting knowledge graph contains 28,580 targets, 10,493 diseases, 1,423,618 associations between targets and diseases, 476,525 nodes, and 73 million edges.

Graph embedding algorithms are applied to create feature vector representations of nodes in the graph (i.e., embeddings), which represent semantic relationships between entities. We used PyTorch BigGraph (PBG) toolbox to generate the embeddings, which are used to predict if a new link might exist between a pair of nodes. The PBG model ranks the targets such that high scores are given to targets with a high probability to be associated with a disease. The accuracy of the PBG model is assessed using a hold-out testing set, and the model achieved HIT@10, HIT@50, and HIT@100 scores of 43%, 70%, and 83%, respectively in ranking edges between genes and diseases. Additionally, the model ranks associations between differentially expressed genes and tissues with Mean Rank score of 11 and Hit@100 score of 97%. Moreover, the model predicts associations between diseases and variants with HIT@100 score of 72%.

Overall, this study provides a large-scale knowledge graph created for the purpose of targets discovery and provides an accurate model for predicting associations between targets and diseases. The computational approach applied in this study is generic in nature and can be applied or extended to find associations between other biological entities as well.