A knowledge graph-based disease-gene prediction system using multi-relational graph convolution networks

Zhenxiang Gao, PhD[1], Yiheng Pan, M.S.[1], Pingjian Ding, PhD[1], Rong Xu, PhD[1]
[1]Center for Artificial Intelligence in Drug Discovery, Case Western Reserve University School of Medicine, Cleveland, OH, USA

Abstract

Identifying disease-gene associations is important for understanding molecule mechanisms of diseases, finding diagnostic markers and therapeutic targets. Many computational methods have been proposed to predict disease related genes by integrating different biological databases into heterogeneous networks. However, it remains a challenging task to leverage heterogeneous topological and semantic information from multi-source biological data to enhance disease-gene prediction. In this study, we propose a knowledge graph-based disease-gene prediction system (GenePredict-KG) by modeling semantic relations extracted from various genotypic and phenotypic databases. We first constructed a knowledge graph that comprised 2,292,609 associations between 73,358 entities for 14 types of phenotypic and genotypic relations and 7 entity types. We developed a knowledge graph embedding model to learn low-dimensional representations of entities and relations, and utilized these embeddings to infer new disease-gene interactions. We compared GenePredict-KG with several state-of-the-art models using multiple evaluation metrics. GenePredict-KG achieved high performances [AUROC (the area under receiver operating characteristic) = 0.978, AUPR (the area under precision-recall) = 0.343 and MRR (the mean reciprocal rank) = 0.244], outperforming other state-of-art methods.

1. Introduction

Disease related gene detection is important for disease mechanism understanding and treatments[1,2]. Traditional approaches such as genome-wide association studies and linkage analysis were used for discovering disease-related gene[3,4]. In the past few decades, computational methods have been developed to prioritize candidate genes for diseases[5-8]. Network-based computational methods are commonly used for inferring disease-gene associations. The main intuition behind network-based models is that similar genes are more likely to be associated with a similar set of diseases and similar diseases tend to share similar genes[9,10]. Kohler et al. [11] presented a random walk method for prioritization of candidate genes by use of a global network distance measure for definition of similarities in protein-protein interactions. Maji et al.[12] proposed a gene selection algorithm by maximizing the relevance and functional similarity of the selected genes to identify disease related genes. Xu et al. constructed contest-sensitive networks (CSNs)[13] by directly connecting diseases with associated phenotypes and used a network-based ranking algorithm to predict disease genes associations. Chen et al.[14] constructed a gene-centric heterogeneous network based on gene associated phenotypes, functions of the gene products and anatomical location of gene expression, and further developed a graph-based method to predict gene-disease associations. However, due to the complexity and high dimensionality of biomedical data, how to make more effective use of the semantic knowledge embedded in the heterogeneous data resources remains a challenging problem in disease-gene prediction.

Recently knowledge graphs (KGs) have emerged as an effective way to integrate various data sources and model complex semantic relationships[15,16]. KGs have been used for drug repurposing, drug toxicity prediction and drug-target prioritization[17-21]. In this study, we developed a KG-based system for disease gene prediction. We first constructed a knowledge graph that consisted of more than 2,000,000 semantic inter-connections between drugs, genes, diseases, and phenotypic annotations from 10 data resources including publicly available genetic and genomic databases, ontologies, health records of patients, FDA drug labels, published research articles, and clinical trial studies. We then developed a deep learning framework to embed the knowledge graph into low-dimensional latent vectors and utilized these embeddings to predict novel disease-gene associations. We conducted extensive experiments to compare GenePredict-KG with several state-of-the-art models using multiple evaluation metrics.

2. Data and Method

2.1 Data

The knowledge graph was built from publicly available databases as well as knowledge bases that were constructed from published biomedical literature and patient health records using natural language processing and data mining techniques. Figure 1 showed the semantic descriptions and distributions of all relationships in the knowledge graph. We first collected nine types of phenome-level associations from several public databases. Human Phenotype Ontology (HPO) [22] provides a standardized vocabulary of phenotypic abnormalities that have been seen in human disease. We obtained Gene-human phenotype (HP) and Disease-HP pairs from HPO. The SIDER[23] database provides information on marketed medicines and their recorded adverse reactions. We extracted Drug-side effects (SE) associations from the SIDER database. Genotype-Tissue Expression (GTEx)[24] database provides the relationship between genetic variants and gene expression in multiple human tissues and across individuals. We chose a cutoff of 4.0 transcripts per million as threshold[14] to extract genes expressed in each tissue and mapped each tissue to the Uberon Anatomy Ontology[25] to obtain Gene-Uberon Tissue associations. The Gene Ontology Annotation (GOA) [26] database provides annotations to the UniProt Knowledgebase using the standardized vocabulary of the Gene Ontology. We downloaded GO annotations (GOA) from the GO database and extracted Gene-GOA associations. Phenomebrowser[27] is a platform that aggregates phenotype connections with biomedical concepts and provides drug-phenotype dataset which include Mammalian Phenotype (MP) Annotations[28] and Human Phenotype (HP) Ontology associating drugs. We extracted Drug-MP and Drug-HP associations from Phenomebrowser. Mouse Genome Informatics (MGI)[29] database provides access to data on the genetics, genomics, and biology of the laboratory mouse to facilitate the study of human health and disease. We extracted Gene-MP and Disease-MP associations from MGI. DrugBank[30] is a comprehensive online database that provides information on drugs and drug targets. We obtained Drug-Gene associations from DrugBank. The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)[31] captures protein-protein interactions from over 5K different organisms. Gene-Gene associations was extracted from STRING. We also extracted genotypes and diseases from MGI and mapped each mouse gene to its human ortholog using mouse-human orthology to construct Gene-Disease pairs.

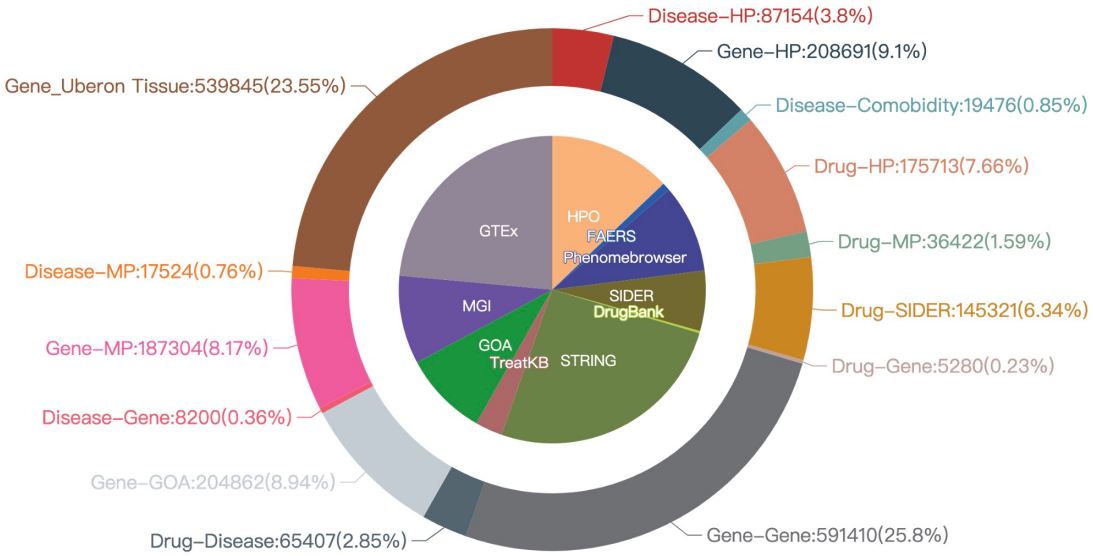

Figure 1. Summary of relationship themes and their distributions in the knowledge graph.

In addition, two types of disease and drug relationships were obtained from published biomedical literature, clinical trial reports, FDA drug labels and patient health records. In our previous studies[32], we extracted disease-comorbidity pairs from FAERS, a large-scale database that contains patient diseases, medications, drug adverse events, demographics of 17 million case reports[33], using Frequent Pattern (FP)-growth algorithm[34]. Our previous studies also constructed TreatKB which included drug-disease treatment relationships mined by NLP techniques from records of patients in FAERS, FDA drug labels, MEDLINE abstracts and clinical trial studies[35,36].

Table 1 highlighted entity types and the number of entities extracted from each database. To integrate data from different data resources, we mapped each entity to an identifier using standard biomedical terminologies. Drug names

were mapped to their active ingredients using PubChem identifiers. Gene symbols were standardized based on Entrez gene[37] identifier provided by the Entrez database. Diseases that were represented using their OMIM (online mendelian inheritance in man)[38] identifiers in MGI and HPO databases were mapped into UMLS (unified medical language system)[39] CUIs (Concept Unique Identifiers)[40]. Disease names in the disease comorbidity knowledge base that we constructed from FAERS have already been mapped to UMLS CUIS in our previous studies[32-33]. Drug entities and disease entities in TreatKB have also been mapped to their CUIs in UMLS[35,36]. Drug and side effect names in SIDER are standardized to their corresponding UMLS CUIs[23]. The standardized knowledge graph contained 73,358 nodes, 2,292,609 edges, 7 node types, and 14 semantic relationships. The distribution of relationships is shown in Figure 1. For example, the "Disease-Gene" relationship occupies 0.36% (8,200/2,292,609) in the knowledge graph.

Table 1. The type-wise distribution of the entities in the original database.

DataSource	Drug	Gene	Disease	GO Terms	MP	UBERON Tissues	HP
HPO	-	4,730	7,172	-	-	-	9,418
SIDER	1.430	-	4,251	-	-	-	-
GTEx	-	16,579	-	-	-	51	-
MGI	-	12,707	4,548	-	9,936	-	-
GOA	-	16,283	-	15,924	-	-	-
Phenombrows	1,429	-	-	-	1,363	-	3,003
DrugBank	984	1,365	-	-	-	-	-
STRING	-	15,664	-	-	-	-	-
FAERS	-	-	924	-	-	-	-
TreatKB	1,973	-	6,988	-	-	-	-

2.2 GenePredict-KG: model, training, evaluation

The overall architecture of GenePredict-KG is shown in Figure 2. The encoder of GenePredict-KG represents entities by aggregating connected entities and relations in the knowledge graph. With node and relation embeddings as the input, the decoder is used to represent the relations by recovering the original triplets in the knowledge graph. More specifically, we first define the knowledge graph as,

$$G = (V, E, X, R) \tag{1}$$

where V and E denote the set of entities and relations, respectively. $T \subseteq V \times E \times V$ denotes the set of triplets, X representes features of nodes, R denotes features of relations. The encoder takes G as input and learns embeddings of entities and relations by aggregating multi-relational information in the knowledge graph. Given an entity v and a relation r, the decoder is input with node embedding h_v and relation embedding h_r generated by the encoder and predicted another suitable entity composing a correct triplet. We formulate the encoding and decoding part in the following sections.

Encoder

The encoder module is based on composition-based multi-relational graph convolutional networks (CompGCN)[42] to aggregate the amount of information from neighboring entities and relations to learn representations of entities and relations. As shown in Figure 2, for capturing multi-hop dependencies in the knowledge graph, the encoder stacks several convolutional layers and treats the knowledge graph as multiple single-relational subgraphs where each subgraph entails a specific type of relations.

For the first layer, the relationship between two adjacent nodes is determined by their relation embedding for each edge type. We define separate filters for each of them. The update equation of node embedding is given as:

$$h_v^1 = f(\sum_{u,r \in N(v)} W_\lambda^1 \psi(x_u, z_r)) \tag{2}$$

where x_u , z_r denote initial features for node u and relation r respectively, f is an activation function, $\psi(\cdot)$ is a composition operator, h_v^1 denotes the updated representation of node v at first layer, and W_λ^1 is relation-specific coefficient matrix at the first layer.

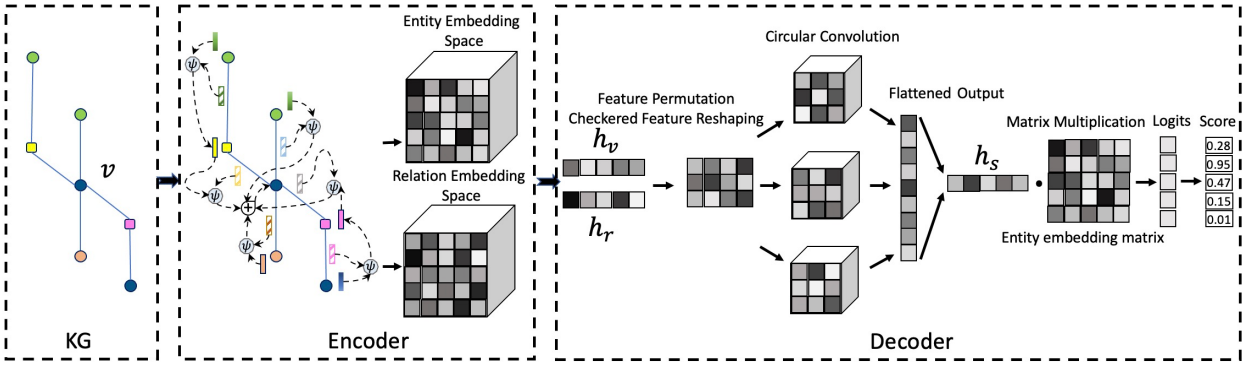

Figure 2 Overview of GenePredict-KG. For encoder, a stack of multiple graph convolutional layers embeds both entities and relations into a low-dimensional embedding space. For decoder, $h_v$and h_r are fed into the model and the model outputs embedding $h_s$through the three operations. Then h_s is matched with all candidate embeddings via inner products. A logistic sigmoid function is used to get predicting scores.

The relation embeddings are also transformed as follows:

$$h_r^1 = W_{rel}^1 z_r \tag{3}$$

where W_{rel}^1 is a learnable transformation matrix which projects all the relations to the same embedding space as nodes and allows them to be utilized in the next layer.

Let h_v^l represents the input vector of the entity v in the l-th layer. If there are a total of l layers in the encoder, the output h_v^{l+1} of the l-th layer is the final embedding of node v. Hence, the output of l-th layer for entity v can be written as follows:

$$h_v^{l+1} = f(\sum_{u,r \in N(v)} W_\lambda^l \psi(h_u^l, h_r^l) \; + \; W_o^l h_v^l) \tag{4}$$

where W_o^l is self-specific coefficient matrix.

Similarly, let h_r^{l+1} denote the representation of a relation r after l layers. Then,

$$h_r^{l+1} = W_{rel}^l h_r^l \tag{5}$$

Decoder

The decoder is based on the InteractE model[43] to infer unseen interactions by defining a score function $\phi(v, r, s)$ for each triplet. This optimization aims to generally score a correct triplet higher than incorrect triplets.

The decoder first utilizes multiple permutations to capture more possible feature interactions preserved in the embeddings of entity and relation. It generates t-random permutations of both h_v and h_r as,

$$P_t \; = \; [(h_v^1, h_r^1); \ldots; (h_v^t, h_r^t)] \tag{6}$$

Then the decoder applies chequer reshapes as the reshaping operation. Thus,

$$\varphi(P_t) \; = \; [\varphi(h_v^1, h_r^1); \ldots; \varphi(h_v^t, h_r^t)] \tag{7}$$

where $\varphi(\cdot)$ is the reshaping function capturing maximum heterogeneous interactions between entity and relation features.

Finally, the decoder stacks the reshaped matrices into a 3D tensor, that is then processed with depth-wise circular convolution. The output of each circular convolution is flattened and concatenated into a vector. The decoder projects this vector to the embedding space. Formally, the score function is defined as follows:

$$\phi(s, r, o) \; = \; g(vec(\, f(\varphi(P_t) \; \oplus \; w \,)\,)\, W_p \,)\, h_s \tag{8}$$

where $\oplus$ denotes depth-wise circular convolution, $vec(\cdot)$ denotes vector concatenation, h_s represents the object entity embedding and W_p is a learnable weight matrix. $w \in R^{k \times k}$ is a convolutional kernel of size k. Functions f and g are chosen to be ReLU and sigmoid respectively.

Model Training

The GenePredict-KG is a knowledge graph embedding model that follows the multi-phase procedure to learn a vector representation for entities and relation of a knowledge graph. First, the model initializes its embeddings with random noise. It then updates them by iterative learning on the training data. In each training iteration (e.g., epoch), the model splits the training data into mini-batches and executes its learning pipeline over each batch. The learning pipeline of the model learns the embeddings of entities and relations by maximizing the scores of true triples and minimizing false triples.

We used the standard cross entropy loss as loss function. These losses were used by Adam optimizers to generate gradients and updated embeddings and parameters. Hyperparameters for GenePredict-KG were tuned using the grid search on the validation set. We tuned the learning rate {0.0001, 0.001, 0.01}, embedding dimensions {100, 200, 400}, Number of GCN Layer {1, 2, 3}, batch size {64, 128, 256}, and dropout {0.1, 0.2, 0.3}. The optimal values were selected based on MRR results. We first checked whether the number of GCN layers affected the performance or not. We found that deeper GCN layers did not highly improve the performance and can largely increase of computational cost. Thus, the number of GCN layers was set to 1. We also investigated the influence of embedding dimensions. Increasing embedding dimensions can boost the performance. But the gain was marginal. Too large dimension of embedding was a burden on memory and computation. Thus, the dimension of both hidden layers and the latent vectors were all set to 200. The size of the mini-batch was set to 128. The learning rate was set to 0.001. To avoid overfitting, we used dropout after each convolution layer with the drop rate of 0.1 in the embedding module and 0.3 in the predicting module. This procedure was performed iteratively for 500 iterations.

Evaluation and Comparison

We conducted cross validation to evaluate the model performance on disease-gene prediction. All disease-gene associations were randomly shuffled five times, and were spitted into training (80%), validation (10%) and test (10%) set in each round. We used each training set to build the model, used each corresponding validation set to optimize the parameter setting of the model, and used the test set to verify the model performance and report results. This procedure was repeated five times. In each testing configuration, we used the known disease-gene interactions as positives, and all other possible combinations between genes and diseases as negatives. We compared our model with several state-of-the-art knowledge graph embedding methods for the task of disease-gene prediction, including TransE[44], and HRGAT[21]. TransE was a knowledge graph embedding model for link prediction that has been applied in drug repurposing and drug target prediction[17,19,20]. HRGAT was a graph neural network-based method to learn knowledge graph embedding for drug-disease and disease-gene predictions.

We also compared GenePredict-KG with traditional network-based methods, including the context-sensitive networks (CSN)-based approach that we have recently developed[13]. The CSNs modeled biomedical relationships by taking into account their context-specific semantics. We have showed that CSN-based approaches performed better than similarity-based network approaches in both disease-gene prediction[13], drug target prediction[45] and drug repurposing[46].

We also provided a comprehensive comparison of GenePredict-KG with two state-of-art disease-gene predicting systems including DL2Vec[14] and SmuDGE[47]. DL2Vec integrated a variety of gene-related information to construct a heterogeneous network including gene associated phenotypes, functions of the gene products and anatomical location of gene expression. It utilized several machine-learning models for disease-gene based on the heterogeneous networks. SmuDGE exploited ontologies and knowledge graphs to learn representations of genes, gene products and diseases, based on the phenotypes they are associated with. These representations were further be used to predict gene-disease associations.

We used several widely used metrics[17,21,45,48] to comprehensively evaluate and compare the performance of GenePredict-KG to other methods. These metrics included Hits@N, the mean reciprocal rank (MRR), area under receiver operating characteristic curve (AUROC) and area under precision-recall curve (AUPRC). Hits@N is the hit percentage of true samples in a test set being ranked by a model within the top N positions, it evaluates the ability for "early recognition" of true predictions. MRR is the average inverse rank for true samples. A higher MRR value indicates a better model.

3. Results

We first tested GenePredict-KG and compared it with several state-of-art knowledge graph embedding methods on the same knowledge graph. TransE was a translation-based method for link prediction, which employs a transitional

characteristic to model relationships between entities. HRGAT was a graph neural network-based method which used global network structure and domain features embedded in knowledge graph to predict new interactions. As shown in Figure 3, GenePredict-KG outperformed the TransE by 7.3% (Hit@1), 5.9% (MRR), 2.0% (AUROC) and 3.4% (AUPR). It also improved over HRGAT with 3.3% on MRR and 2.9% on AUPR. This validated GenePredict-KG captured rich heterogeneous topologic and semantic information preserved in the entity and relation embeddings can better infer disease-gene associations.

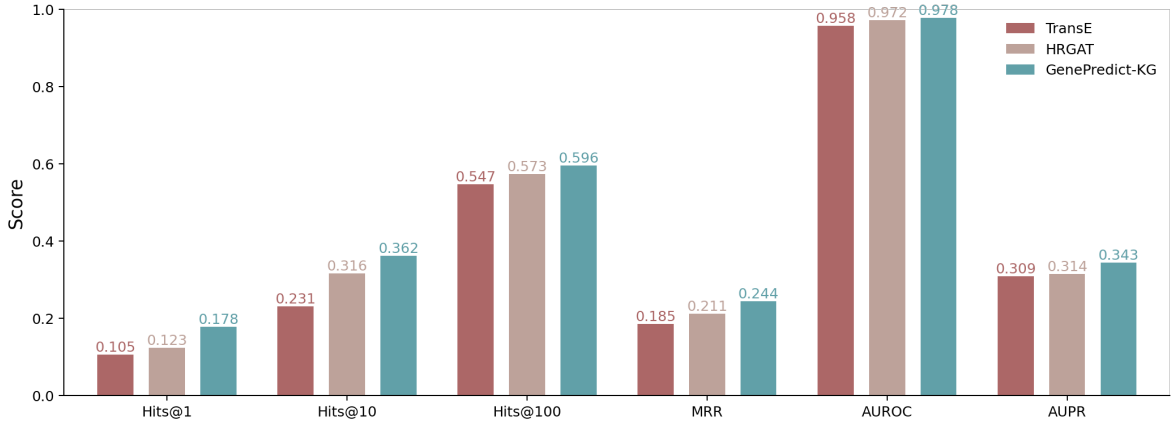

Figure 3 Overall predictive performance of disease-gene associations on the proposed knowledge graph

We compared GenePredict-KG with CSN-Rank based on the same datasets that CSN-Rank used. CSN-Rank predicted disease-gene associations based on a random walker in context sensitive networks and achieved a better performance than similarity-based network approaches[13]. GenePredict-KG aimed to learn low-dimensional representations of entities and relations and further utilized these representations for disease-gene predictions. Results were shown in Figure 4. The GenePredict-KG outperformed CSN-Rank in terms of Hits@N and MRR. In addition, GenePredict-KG consistently got a better performance, with 0.4% higher AUROC and 3.3% higher AUPR than CSN-Rank. It indicated that GenePredict-KG can better preserve the topologic and semantic feature information to low-dimensional entity embeddings which enhanced disease-gene prediction.

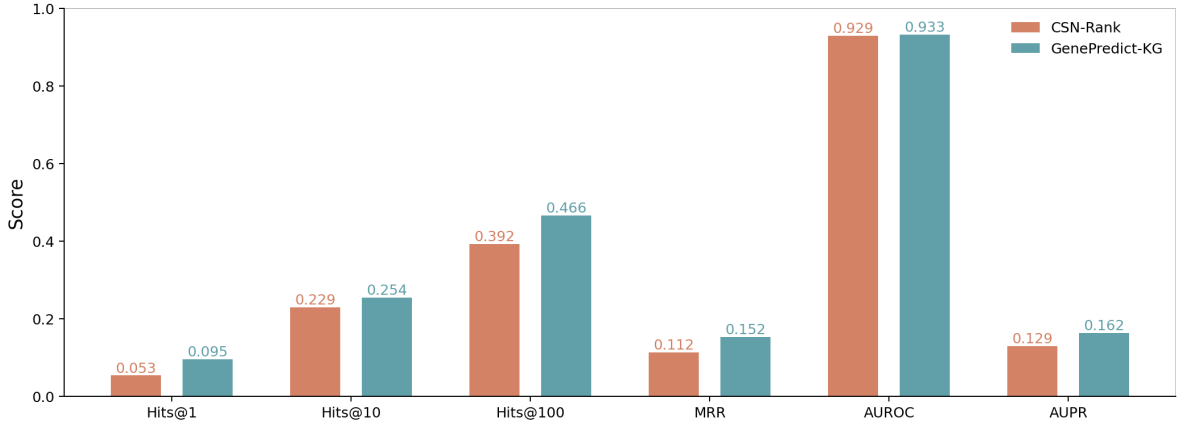

Figure 4 Evaluation results for predicting gene-disease associations using CSNs.

To further evaluate the proposed system performance, we compared GenePredict-KG with state-of-art disease-gene prediction systems. Chen et al.[14] constructed a heterogenous networks (HNs) and used several machine learning models to predict disease-gene associations. They conducted cross-validation for disease-gene prediction. Within each split, they used 10% of the data as the test data to report the results and used the other 90% data to train the model and tune its parameters. We selected DL2Vec and SumDGE with excellent performance on HNs as baselines. The HNs shared the same disease-gene associations with our proposed knowledge graph. We followed the same data splitting process to divide the disease-gene associations in our proposed knowledge graph and trained GenePredict-KG. Results were summarized in Table 2. DL2Vec got the highest score of Hits@10. Our GenePredict-KG model improved upon DL2Vec's Hits@1 by a large margin of 14.1%, and upon SumDEG's Hits@1 by a large margin of 15.0%.

GenePredict-KG also achieved a competitive performance with DL2Vec and SumDEG in terms of AUROC. Compared with HNs, GenePredict-KG had more gene-related and disease-related edge types that can form more paths between genes and diseases, which potentially improved predicting performance. In addition, the embeddings of genes and disease learned by our model effectively integrated heterogeneous information from multiple type of entities and relations, hence, was useful for providing accurate disease-gene predictions.

Table 2. Disease-gene predicting performance of GenePredict-KG and two baseline models.

Model	Hits@1	Hits@10	Hits@100	AUROC
SumDGE*	0.028	0.214	0.582	0.973
DL2Vec*	0.037	0.255	**0.637**	0.976
GenePredict-KG	**0.178**	**0.362**	0.596	**0.978**

*Results were adopted from the reference[16].

4. Discussion and Conclusion

In this study, we built a knowledge graph-based predicting system, called GenePredict-KG, to integrate diverse information from different sources to predict novel disease-gene associations. We first constructed a large-scale knowledge graph which contained over 2,000,000 interconnections between more than 70,000 entities including drugs, genes, diseases, and phenotypic annotations. We then developed a knowledge graph-based predicting system, GenePredict-KG, to predict disease-associated genes. GenePredict-KG was validated on two datasets and compared with state-of-the-art methods using several evaluation metrics. We demonstrated that GenePredict-KG achieved high performances and outperforming other state-of-art methods.

GenePredict-KG has several limitations that warrants further investigation. First, the underlying knowledge graph of GenePredict-KG is un-weighted and the encoder integrated information from all neighboring entities with equal weights to learn knowledge graph embeddings. However, not all neighbors have the same contribution to the prediction target. Currently we are experimenting with adding a self-attention layer to the encoder to assign different weights to neighbors to learn more robust embeddings enhancing the accuracy of disease-gene prediction. Second, we currently treated all genes that are not known to be associated with a given disease as negatives. That can result in data class imbalance (negatives outnumber positives) as well as a high rate of false negatives. In the future, we will utilize several rational sampling methods[49,50] to reduce data imbalance to improve predicting performance. Third, the predicted disease-gene relationships are associational. Additional works are necessary to further establish the causal relationships between genes and diseases as well as if a gene is a driver or an effector for a given disease. Finally, we have evaluated the performance of GenePredict-KG using standard cross-validation based on known disease-gene associations. However, how GenePredict-KG in identifying truly novel (and unknown) disease-gene associations remains unknown.

In summary, GenePredict-KG incorporated a large amount of semantic knowledge from multiple data resources and the knowledge graph embedding was able to learn low-dimensional representations of entities and relations, and utilized these embeddings to infer new disease-gene interactions. GenePredict-KG achieved high performance and outperformed state-of-art methods, indicating that it has high potential in discovering disease associated genes.

Acknowledgments

This work has been supported by NIH National Institute of Aging R01 AG057557, R01 AG061388, R56 AG062272, National Institute on Alcohol Abuse and Alcoholism (grant no. R01AA029831), National Eye Institute (EY029297), National Institute on Drug Abuse (UG1DA049435, CTN-0114), American Cancer Society Research Scholar Grant RSG-16-049-01 – MPC, The Clinical and Translational Science Collaborative (CTSC) of Cleveland (UL1TR002548-01)

References

1. Wang L, Wu M, Wu Y, Zhang X, Li S, He M, ... Li J. Prediction of the Disease Causal Genes Based on Heterogeneous Network and Multi-Feature Combination Method. Comput Biol Chem. 2022;107639.
2. Shu J, Li Y, Wang S, Xi B, Ma J. (). Disease gene prediction with privileged information and heteroscedastic dropout. Bioinform. 2021;37: i410-i417.

3. Xiang J, Zhang NR, Zhang JS, Lv XY, Li M. PrGeFNE: predicting disease-related genes by fast network embedding. Methods. 2021; 192:3-12.
4. Luo P, Li Y, Tian LP, Wu FX. Enhancing the prediction of disease–gene associations with multimodal deep learning. Bioinform, 2019; 35: 3735-3742.
5. Ata SK, Wu M, Fang Y, Ou-Yang L, Kwoh CK, Li XL. Recent advances in network-based methods for disease gene prediction. Brief. Bioinformatics. 2021; 22: bbaa303.
6. Luo P, Tian LP, Chen B, Xiao Q, Wu FX. Ensemble disease gene prediction by clinical sample-based networks. BMC Bioinform. 2020; 21: 1-12.
7. Liu Z, Hu J. Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction. Methods. 2016; 93: 119-127.
8. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nature reviews Genetics. 2009; 10: 392-404.
9. Erten S, Bebek G, Koyutürk M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. J Comput Biol. 2011; 18: 1561-1574.
10. Zhu J, Qin Y, Liu T, Wang J, Zheng X. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. BMC Bioinform. 2013; 14: 1-11.
11. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008; 82: 949-958.
12. Maji P, Shah E, Paul S. RelSim: An integrated method to identify disease genes using gene expression profiles and PPIN based similarity measure. Inf Sci. 2017; 384: 110-125.
13. Chen Y, Xu R. Context-sensitive network-based disease genetics prediction and its implications in drug discovery. Bioinform, 207; 33(7): 1031-1039.
14. Chen J, Althagafi A, Hoehndorf R. Predicting candidate genes from phenotypes, functions and anatomical site of expression. Bioinform. 2021; 37(6): 853-860.
15. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, ... Hamilton W. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. arXiv preprint arXiv. 2021:2102.10062.
16. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020; 18: 1414-1428.
17. Moon C, Jin C, Dong X, Abrar S, Zheng W, Chirkova RY, Tropsha A. Learning Drug-Disease-Target Embedding (DDTE) from knowledge graphs to inform drug repurposing hypotheses. J Biomed Inform. 2021; 119: 103838.
18. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. Bioinform. 2020; 36(2): 603-610.
19. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. J Biomed Inform. 2021; 115: 103696.
20. Zhang X, Che C. Drug Repurposing for Parkinson's Disease by Integrating Knowledge Graph Completion Model and Knowledge Fusion of Medical Literature. Future Internet. 2021; 13(1): 14.
21. Zheng S, Rao J, Song Y, Zhang J, Xiao X, Fang EF, ... Niu Z. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in bioinformatics, 2021; 22(4): bbaa344.
22. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, et al. The human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet. 2015; 97:111–24.
23. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016; 44(D1): D1075-D1079.
24. Zhang X, Che C. Drug Repurposing for Parkinson's Disease by Integrating Knowledge Graph Completion Model and Knowledge Fusion of Medical Literature. Future Internet. 2021; 13(1): 14.
25. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012; 13(1): 1-20.
26. Consortium GO, et al. Gene ontology consortium: going forward. Nucleic Acids Res. 2015; 43(D1): D1049-56.
27. OntoSIML. http://phenomebrowser.net/archive/sider_drug_phenotype.txt, 2021: Phenomebrowser.
28. Smith CL, Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med. 2009; 1(3): 390-399.
29. Eppig JT, Smith CL, Blake JA, Ringwald M, Kadin JA, Richardson JE, Bult CJ. Mouse Genome Informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. Syst Genet. 2017: 47-73.
30. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, ... Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 2008; 36: D901-D906.

31. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015; 43(D1): D447–D452.
32. Zheng C, Xu R. Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data. BMC bioinform. 2018; 19(17): 85-93.
33. FAERS. https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/instructions-requesting-individual-case-reports.
34. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM sigmod record, 2000; 29(2): 1-12.
35. Xu R, Wang Q. Automatic signal extraction, prioritizing and filtering approaches in detecting post-marketing cardiovascular events associated with targeted cancer drugs from the FDA Adverse Event Reporting System (FAERS). J Biomed Inform. 2014; 47: 171-177.
36. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC bioinform. 2013; 14(1): 1-11.
37. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005; 33: D54-D58.
38. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005; 33: D514-D517.
39. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. (1998). Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc. 1998; 5(5): 421-431.
40. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Yearb Med Inform. 1993; 2(01): 41-51.
41. MetaMap. https://doi.org/metamap.nlm.nih.gov.
42. Vashishth S, Sanyal S, Nitin V, Talukdar P. Composition-based multi-relational graph convolutional networks. arXiv preprint arXiv. 2019:1 911.03082.
43. Vashishth S, Sanyal S, Nitin V, Agrawal N, Talukdar P. (2020, April). Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions. AAAI. 2020; 34(3): 3009-3016.
44. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. NIPS, 2013;26.
45. Zhou M, Zheng C, Xu, R. Combining phenome-driven drug-target interaction prediction with patients' electronic health records-based clinical corroboration toward drug discovery. Bioinform. 2020; 36: i436-i444.
46. Zhou M, Wang Q, Zheng C, John Rush A, Volkow ND, Xu R. (). Drug repurposing for opioid use disorders: integration of computational prediction, clinical corroboration, and mechanism of action analyses. Molecular psychiatry, 2021; 26: 5286-5296.
47. Alshahrani M, Hoehndorf R. Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes. Bioinformatics, 2018; 34(17): i901-i907.
48. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics, 2019; 35(24): 5191-5198.
49. Tyagi S, Mittal S. Sampling approaches for imbalanced data classification problem in machine learning. ICRIC. 2020: 209-221.
50. Jian C, Gao J, Ao Y. A new sampling method for classifying imbalanced data based on support vector machine ensemble. Neurocomputing, 2016; 193: 115-122.