Learning Causal Relationships in Rare Disease Graphs

1 Introduction

One of the causes of developing the diseases in the human body are the genes which are inherited from the parents. These sets of genes of an individual are called genotypes. Such genotypes are part of DNA[3] and encodes information about the human body, for example encoding of brown hair or blue eyes.

On the other hand, phenotypes/diseases are the visible/outside characteristics of the human body. There are several factors causing such characteristics which include (i) underlying genes (genotypes) and (ii) environmental factors by which the individual's body develops diseases with time.

If a set of genes causes a phenotype (disease) then the relation between genotypes and phenotypes is causal otherwise non-causal. For example: If any individual has a set of genes which causes the genetic disorders like diabetes then the relation between the underlying genes and the disease (diabetes) is considered causal. Otherwise non-causal relation.

As a part of this project, I will focus on predicting the relations between genotypes and phenotypes as causal or non-causal relationships.

Knowing casual or non-causal relationship between genotypes and phenotypes can be helpful in the following ways:

- Predicting phenotypes from gene mutations: In an individual, changes happen to the human body and that information is encoded and passed on to the next generation. Such changes result in gene mutation. Knowing the interaction between genes and phenotypes can help to predict future diseases from gene mutation.
- Identifying the relation between genes and phenotypes, usually takes years of research and cost considerable amount of money. Automating the process could be really helpful.
- Understanding interaction between genes,addiction, and phenotypes: Individuals have varying degrees of tolerance for addiction e.g smoking. Such addiction with time produces unwanted effects on the human body e.g lung cancer (phenotype). Understanding the causality between genotype and phenotype, it is possible to know the risk level of addiction.

To facilitate such study, there is human curated dataset called OMIM[1]. One of the recent papers[2], uses OMIM dataset to predict genes-disease association. Input to the model is a molecular network of proteins present in disease and genes, and output is information on relation between genes and phenotypes. Another paper [4] which have focused on identifying relation between the genes and phenotypes have used sentence from abstracts of the research paper along with automatically extracted genes and phenotypes from the sentence as their input and predict binary relation between genes and phentoype.

2 Dataset

For this project, the data has been collected from the OMIM (Online Mendelian Inheritance in Man)[1]. OMIM is a catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. This dataset

is publicly available for research projects and is continuously updated. There are three categories of information considered in this project: (i) Genes, (ii) Phenotypes, and (iii) Phenotypic Series. Here, genes are the genotypes, phenotypes are the diseases and phenotypic Series is a tabular view of genetic heterogeneity of similar phenotypes across the genome.

All the relationships mentioned between the genotype and phenotype in this dataset are considered to be a causal relationship (positive examples). Because, non-causal relationships are not present in the dataset and are thus, created for this project, by random sampling from the set of available genotypes and phenotypes to build a negative pairs which are not present in the causal relationship.

The dataset will be divided into three datasets (1) Train (2) Development/Validation and (3) Test. From this dataset, Gene-Phenotype pairs (Causal and Non- causal) will be considered for training the model to learn the targeted relationship. This approach will make sure to focus on these pairs and not on other pairs i.e. Gene-Gene, Phenotype-Phenotype, Phenotype-PS, PS-PS, Gene-PS.

Assumption : Non-causal relationship examples are generated considering the data only from OMIM. Though the dataset is continuously updated, I will be working on a snapshot of the data frozen in time.

3 Evaluation Method

There are two possible ways to evaluate the method:

- As predicting link between the genes-phenotypes is considered as the binary classification, evaluation methods for this project are Precision, Recall and F1 score on test dataset. The reason for choosing these metrics is that the dataset can be balanced or skewed depending on the number of the negative examples considered.
- Another possible evaluation method for this project could be ranking where the measures are Precision and HIT at different level. For example, HIT@5 will be as follows: Given the phenotype, I will rank the genes as retrieved documents and if the ground-truth gene occurs in top 5 results, it is considered as HIT at top 5 else miss.

References

[1] Joanna S Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh.Omim. org: leveraging knowledge across phenotype–gene relationships. *Nucleic acids research*, 47(D1):D1038–D1043, 2019.

[2] Sezin Kircali Ata, Min Wu, Yuan Fang, Le Ou-Yang, Chee Keong Kwoh, and Xiao-Li Li. Recent advances in network-based methods for disease gene prediction. *arXiv preprint arXiv:2007.10848*, 2020.

[3] Molly Campbell. *Genotype vs Phenotype: Examples and Definitions*, 2019.

[4] Diana Sousa, Andre Lamurias, and Francisco M. Couto. A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1487–1492, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

A Silver Standard Corpus of Human Phenotype-Gene Relations

Diana Sousa*, Andre Lamurias and Francisco M. Couto

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract

Human phenotype-gene relations are fundamental to fully understand the origin of some phenotypic abnormalities and their associated diseases. Biomedical literature is the most comprehensive source of these relations, however, we need Relation Extraction tools to automatically recognize them. Most of these tools require an annotated corpus and to the best of our knowledge, there is no corpus available annotated with human phenotype-gene relations. This paper presents the Phenotype-Gene Relations (PGR) corpus, a silver standard corpus of human phenotype and gene annotations and their relations. The corpus consists of 1712 abstracts, 5676 human phenotype annotations, 13835 gene annotations, and 4283 relations[1]. We generated this corpus using Named-Entity Recognition tools, whose results were partially evaluated by eight curators, obtaining a precision of 87.01%. By using the corpus we were able to obtain promising results with two state-of-the-art deep learning tools, namely 78.05% of precision. The PGR corpus was made publicly available to the research community.[2]

1 Introduction

Automatic extraction of relations between entities mentioned in literature is essential to obtain knowledge that was already available but required considerable manual effort and time to retrieve. Recently, biomedical relation extraction has gained momentum in several text-mining applications, such as event extraction and slot-filling (Lamurias and Couto, 2019b). Some of the commonly extracted biomedical relations are protein-protein interactions (Papanikolaou et al., 2015), drug-drug interactions (Lamurias et al., 2019) and disease-gene relationships (Kim et al., 2017).

There are a few worth mention systems regarding biomedical Relation Extraction (RE) (Verga et al., 2018), and that specifically focus on the extraction of phenotype-gene relations regarding different species types like plants (Xing et al., 2018) and humans (Collier et al., 2015). The main problem that these systems face is a lack of specific high quality annotated corpora (gold standard corpus), mostly because this task requires not only a considerable amount of manual effort but also specific expertise that is not widely available. A solution to these limitations is to generate the corpus in a fully automated manner (silver standard corpus).

Connecting human phenotypes to genes helps us to understand the origin of some phenotypic abnormalities and their associated diseases. To extract human phenotype-gene relations, both entities, human phenotypes and genes have to be recognized. With genes, as a result of lexical features being relatively regular, many systems can successfully identify them in text (Leaman and Gonzalez, 2008). Even though Named-Entity Recognition (NER) research has significantly improved in the last years, human phenotype identification is still a complex task, only tackled by a handful of systems (Lobo et al., 2017).

To generate a silver standard for phenotype-gene relation extraction, we used a pipeline that performs: i) NER to recognize genes and human phenotype entities; ii) RE to classify a relation between human phenotype and gene entities. First, we gathered abstracts using the PubMed API with manually defined keywords, namely each gene name, *homo sapiens*, and *disease*. Then we used the Minimal Named-Entity Recognizer (MER) tool (Couto and Lamurias, 2018) to extract gene mentions in the abstracts and the Identifying Human Phenotypes (IHP) tool (Lobo et al.,

*dfsousa@lasige.di.fc.ul.pt

[1] Query 1, corresponds to the *10/12/2018* release of PGR

[2] https://github.com/lasigeBioTM/PGR

Proceedings of NAACL-HLT 2019, pages 1487–1492
Minneapolis, Minnesota, June 2 - June 7, 2019.

2017) to extract human phenotype mentions. At last, we used a gold standard relations file, provided by the Human Phenotype Ontology (HPO), to classify the relations obtained by co-occurrence in the same sentence as *Known* or *Unknown*.

To the best of our knowledge, there is no corpus available specific to human phenotype-gene relations. This work, overcame this issue by creating a large and versatile silver standard corpus. To assess the quality of the Phenotype-Gene Relations (PGR) corpus, eight curators manually evaluated a subset of PGR. We obtained highly promising results, for example 87.18% in precision. Finally, we evaluated the impact of using the corpus on two deep learning RE systems, obtaining 69.23% (BO-LSTM) and 78.05% (BioBERT) in precision.

2 PGR Corpus

The HPO is responsible for providing a standardized vocabulary of phenotypic abnormalities encountered in human diseases (Köhler et al., 2017). The developers of the HPO also made available a file that links these phenotypic abnormalities to genes. These phenotype-gene relations are regularly extracted from texts in Online Mendelian Inheritance in Man (OMIM) and Orphanet (ORPHA) databases, where all phenotype terms associated with any disease that is related with a gene are assigned to that gene in the relations file. In this work, we used the relations file created by HPO as a gold standard for human phenotype-gene relations.

We started by retrieving abstracts from PubMed, using the genes involved in phenotype-gene relations and *homo sapiens* as keywords, and the Entrez Programming Utilities (E-utilities) web service (`https://www.ncbi.nlm.nih.gov/books/NBK25501/`), retrieving one abstract per gene (Query 1).

Later, we added the keyword *disease* and filter for abstracts in English (Query 2) [3]. Query 2 represents a more focused search of the type of abstracts to retrieve, such as abstracts regarding diseases, their associated phenotypes and genes.

For each gene, we opted for the most recent abstract (Query 1) and the two most recent abstracts (Query 2).

We opted by searching per gene and not human phenotype or the combination of both terms because this approach was the one that retrieved abstracts with the higher number of gene and human phenotype annotations, in the following NER and RE phases. We removed the abstracts that did not check the conditions of being written in English, with a correct XML format and content. The final number of abstracts was 1712 for Query 1 and 2657 for Query 2 as presented in Table 1. Then we proceeded to use the MER tool (Couto and Lamurias, 2018) for the annotation of the genes and the IHP framework (Lobo et al., 2017) for the annotation of human phenotype terms.

2.1 Gene Extraction

MER is a dictionary-based NER tool which given any lexicon or ontology (e.g., an OWL file) and an input text returns a list of recognized entities, their location, and links to their respective classes.

To annotate genes with MER we need to provide a file of gene names and their respective identifiers. To this goal, we used a list created by the HUGO Gene Nomenclature Committee (HGNC) at the European Bioinformatics Institute (`http://www.genenames.org/`). The HGNC is responsible for approving unique symbols and names for human loci, including protein-coding genes, ncRNA genes, and pseudogenes, with the goal of promoting clear scientific communication. Considering that we intended not only to map the genes to their names but also their Entrez Gene (`www.ncbi.nlm.nih.gov/gene/`) identifiers, we used the API from MyGene (`http://mygene.info/`) with the keyword *human* in species. The MyGene API provides several gene characteristics, including the confidence score for several possible genes that match the query. For this work, we chose the Entrez Gene identifier with a higher confidence score.

After corresponding all gene names to their respective identifiers, we were left with three genes that did not have identifiers (*CXorf36*, *OR4Q2*, and *SCYGR9*). For the first two genes (*CXorf36* and *OR4Q2*), a simple search in Entrez Gene allowed us to match them to their identifiers. For the last gene (*SCYGR9*) we were not able to find an Entrez Gene identifier, so we used the HGNC identifier for that gene instead. We opted to use the Entrez Gene identifiers because of their widespread use in the biomedical research field.

To the original gene list, we added gene synonyms using a synonyms list file provided by

[3] Query 2, corresponds to the *11/03/2019* release of PGR

Query	Abstracts	Annotations		Relations		
		Phenotype	Gene	Known	Unknown	Total
1 (10/12/2018)	1712	5676	13835	1510	2773	4283
2 (11/03/2019)	2657	9553	23786	2480	5483	7963

Table 1: The final number of abstracts retrieved, number of phenotype and gene annotations extracted and the number of known, unknown and total of relations extracted between phenotype and genes, for Query 1 and 2.

https://github.com/macarthur-lab/gene_lists (expanding the original list almost 3-fold). These synonyms were matched to their identifiers and filtered according to their length to exclude one character length synonyms and avoid a fair amount of false positives. The number of genes in the original gene list was 19194, and by including their synonyms that number increased to 56670, representing a total gain of 37476 genes.

At last, we identified some missed gene annotations that were caught using regular expressions. These missed gene annotations were next to forward/back slash and dashes characters (Example 1).

Example 1. Missed gene annotation because of forward slash.

- **Gene:** *BAX*
- **Gene Identifier:** 581
- **Abstract Identifier:** 30273005
- **Sentence:** According to the morphological observations and DNA fragmentation assay, the MPS compound induced apoptosis in both cell lines, and also cause a significant increase in the expression of **Bax**/Bcl-2.

2.2 Phenotype Extraction

IHP is a Machine Learning-based NER tool, specifically created to recognize HPO entities in unstructured text. It uses Stanford CoreNLP (Manning et al., 2014) for text processing and applies Conditional Random Fields trained with a rich feature set, combined with hand-crafted validation rules and a dictionary to improve the recognition of phenotypes.

To use the IHP system we chose to update the HPO ontology for the most recent version [4]. The annotations that originated from the IHP system were matched to their HPO identifier. There was a total of 7478 annotations for Query 1 and 10973 annotations for Query 2 that did not match any HPO identifier. We put aside these annotations to be confirmed or discarded manually as some of them are incorrectly identified entities but others are parts of adjacent entities that can be combined for an accurate annotation.

We did not use the MER system for phenotype extraction mainly because a more efficient tool for this task was available without the limitations of a dictionary-based NER tool for complex terms as phenotypes are.

2.3 Relation Extraction

After filtering abstracts that did not have annotations of both types, gene and phenotype, we gathered a total of 1712 abstracts for Query 1 and 2656 abstracts for Query 2 as presented in Table 1. The abstracts retrieved by Query 1 were not specific enough for human phenotype-gene relations and therefore about half of them did not contained entities from both types, which we addressed in Query 2, increasing from about 2.5 relations per abstract to about 3.0 relations per abstract.

Using a distant supervision approach, with the HPO file that links phenotypic abnormalities to genes, we were able to classify a relation with *Known* or *Unknown*. For this end, we extract pairs of entities, of gene and human phenotype, by co-occurrence in the same sentence (Example 2). The final number of both *Known* and *Unknown* annotations is also presented in Table 1.

Example 2. Relation extraction.

- **Abstract Identifier:** 23669344
- **Sentence:** A homozygous mutation of **SERPINB6**, a gene encoding an intracellular protease inhibitor, has recently been associated with post-lingual, autosomal-recessive, nonsyndromic **hearing loss** in humans (DFNB91).
- **Gene:** *SERPINB6*
- **Gene Identifier:** 5269
- **Phenotype:** *hearing loss*
- **Phenotype Identifier:** HP_0000365
- **Relation: Known**

[4] *09/10/2018* release

3 Evaluation

To evaluate the quality of the classifier, we randomly selected 260 relations from Query 1 to be reviewed by eight curators (50 relations each, with an overlap of 20 relations). All researchers work in the areas of Biology and Biochemistry. These curators had to evaluate the correctness of the classifier by attributing to each sentence one of the following options: *C* (correct), *I* (incorrect) or *U* (uncertain). The *U* option was given to identify cases of ambiguity and possible errors in the NER phase. We classified as a true positive (TP) a *Known* relation that was marked *C* by the curator, a false positive (FP) as a *Known* relation marked *I*, a false negative (FN) as a *Unknown* relation marked *I* and a true negative (TN) as a *Unknown* relation marked *C*.

3.1 State-of-the-art Applications

The PGR corpus was applied to two state-of-the-art systems that were compared against a co-occurrence (or all-true) baseline method.

3.1.1 BO-LSTM Application

The BO-LSTM system (Lamurias et al., 2019) is a deep learning system that is used to extract and classify relations via long short-term memory networks along biomedical ontologies. This system was initially created to detect and classify drug-drug interactions and later adapted to detect other types of relations between entities like human phenotype-gene relations. It takes advantage of domain-specific ontologies, like the HPO and the Gene Ontology (GO) (Ashburner et al., 2000). The BO-LSTM system represents each entity as the sequence of its ancestors in their respective ontology.

3.1.2 BioBERT Application

The BioBERT system (Lee et al., 2019) is a pre-trained biomedical language representation model for biomedical text mining based on the BERT (Devlin et al., 2018) architecture. Trained on large-scale biomedical corpora, this system is able to perform diverse biomedical text mining tasks, namely NER, RE and Question Answering (QA). The novelty of the architecture is that these systems (BioBERT and BERT) are designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. These feature allows easy adaption to several tasks without loss in performance.

4 Results and Discussion

The final results are presented in Table 2. The inter-curator agreement score, calculated from a total of 20 relations classified by eight curators, was 87.58%. Besides the fact that there were a few incorrectly extracted relations due to errors in the NER phase, that were discarded, the inter-curator agreement is not higher due to the complexity of the sentences where the relations between entities were identified. Even with highly knowledgeable curators in the fields of Biology and Biochemistry, most of them expressed difficulties in deciding which mark to choose on complex sentences that did not necessarily imply a relation between the identified entities (Example 3).

Example 3. Relation marked with *U* (Uncertain).

- **Abstract Identifier:** 27666346
- **Sentence:** **FRMD4A** antibodies were used to probe 78 paraffin-embedded specimens of **tongue squamous cell carcinoma** and 15 normal tongue tissues, which served as controls.
- **Mark:** ***U***

The precision obtained from the test-set (about 6% of the total of relations), was 87.01%. Although we cannot state that this test-set is representative of the overall data-set, this is a solid evidence of the effectiveness of our pipeline to automate RE corpus creation, especially between human phenotype and genes, and other domains if a gold standard relations file is provided. Our lower recall is mostly due to incorrectly retrieved human phenotype annotations by IHP, that can be manually confirmed in a future optimized version of the PGR corpus, as some of them are parts of adjacent entities that can be combined for an accurate annotation.

4.1 Impact on Deep Learning

For BioBERT we used the available pre-trained weights for training and testing of RE model on our corpus. The results of BO-LSTM and BioBERT in the test-set are presented in Table 3. We also measured the performance of the co-occurrence (i.e. assuming all-true) baseline method. This baseline method assumes that all relations in the test-set are *Known* and therefore the recall is 100%. These results are comparable to

Relations		Marked Relations				Metrics		
Known	**Unknown**	**True Positive**	**False Negative**	**False Positive**	**True Negative**	**Precision**	**Recall**	**F-Measure**
77	143	67	86	10	57	0.8701	0.4379	0.5826

Table 2: The *Known* and *Unknown* number of relations selected, the number of true positives, false negatives, false positives and true negatives, and the evaluation metrics for the *Known* relations.

Method	Precision	Recall	F-Measure
Co-occurrence	0.3500	1.0000	0.5185
BO-LSTM	0.6923	0.4200	0.5228
BioBERT	0.7895	0.5844	0.6716

Table 3: Precision, recall and F-measure of the co-occurrence baseline, BO-LSTM and BioBERT.

the ones obtained from the evaluation stage by the curators, and show the applicability of our corpus.

BioBERT significantly outperforms BO-LSTM in all metrics proving that is indeed a viable language representation model for biomedical text mining. Even though the recall for both systems is relatively low, the purpose of this work was mainly to extract correct relations between entities to facilitate Machine Learning (ML), which was achieved by obtaining the precision of 69.23% (BO-LSTM) and 78.95% (BioBERT).

The most relevant metric for a silver standard corpus, directed towards ML tools, is precision. ML tools depend on correct examples to create effective models that can detect new cases, afterwards, being able to deal with small amounts of noise in the assigned labels.

5 Conclusions

This paper showed that our pipeline is a feasible way of generating a silver standard human phenotype-gene relation corpus. The pipeline required the application of two NER tools, and the availability of a list of known relations. We manually evaluated the corpus using eight curators obtaining a 87.01% precision with an inter-agreement of 87.58%. We also measured the impact of using the corpus in state-of-the-art deep learning RE systems, namely BO-LSTM and BioBERT. The results were promising with 69.23%, and 78.95% in precision, respectively. We believe that our pipeline and silver standard corpus will be a highly useful contribution to overcome the lack of gold standards.

Future work includes manually correcting the human phenotype annotations that did not match any HPO identifier, with the potential of expanding the number of human phenotype annotations almost 2-fold and increasing the overall recall. Also, we intend to expand the corpus by identifying more missed gene annotations using pattern matching, which is possible due to our approach being fully automated. Another possibility is the expansion of the test-set for a more accurate capture of the variance in the corpus. For example, we can select a subset of annotated documents in which two curators could work to grasp the complexity of manually annotating some of these abstracts. Further, we intend to use semantic similarity to validate the human phenotype-gene relations. Semantic similarity has been used to compare different types of biomedical entities (Lamurias and Couto, 2019a), and is a measure of closeness based on their biological role. For example, if the *BRCA1* gene is semantically similar to the *BRAF* gene and the *BRCA1* has an established relation with the *tumor* phenotype, it could be possible to infer that *BRAF* gene also has a relation with the *tumor* phenotype, even if that is not evident by the training data. Finally, the effect of different NER systems applied to the pipeline should be studied.

Acknowledgments

We acknowledge the help of Márcia Barros, Joana Matos, Rita Sousa, Ana Margarida Vasconcelos, Maria Teresa Cunha and Sofia Jesus in the curating phase.

This work was supported by FCT through funding of DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), and LASIGE Research Unit, ref. UID/CEC/00408/2019.

References

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan Peter Davis, Kara Dolinski, Selina S. Dwight,

Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Nigel Collier, Tudor Groza, Damian Smedley, Peter N. Robinson, Anika Oellrich, and Dietrich Rebholz-Schuhmann. 2015. Phenominer: from text to a database of phenotypes associated with OMIM diseases. *Database*, 2015:bav104.

Francisco M Couto and Andre Lamurias. 2018. MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(58):1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeongkyun Kim, Jung-jae Kim, and Hyunju Lee. 2017. An analysis of disease-gene relationship from Medline abstracts by DigSee. *Scientific Reports*, 7:40154.

Sebastian Köhler, Nicole Vasilevsky, Mark Engelstad, Erin Foster, et al. 2017. The human phenotype ontology. *2017 Nucleic Acids Research*.

Andre Lamurias and Francisco M Couto. 2019a. Semantic similarity definition. In *Encyclopedia of Bioinformatics and Computtational Biology*, volume 1, pages 870–876. Oxford: Elsevier.

Andre Lamurias and Francisco M Couto. 2019b. Text mining for bioinformatics using biomedical literature. In *Encyclopedia of Bioinformatics and Computtational Biology*, volume 1, pages 602–611. Oxford: Elsevier.

Andre Lamurias, Diana Sousa, Luka A. Clarke, and Francisco M. Couto. 2019. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, 20(1):10.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746.

Manuel Lobo, Andre Lamurias, and Francisco M Couto. 2017. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 7.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford coreNLP natural language processing toolkit. In *ACL*.

Nikolas Papanikolaou, Georgios A. Pavlopoulos, Theodosios Theodosiou, and Ioannis Iliopoulos. 2015. Protein-protein interaction predictions using text mining methods. *Methods*, 74:47–53.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884. Association for Computational Linguistics.

Wenhui Xing, Junsheng Qi, Xiaohui Yuan, Lin Li, Xiaoyu Zhang, Yuhua Fu, Shengwu Xiong, Lun Hu, and Jing Peng. 2018. A gene-phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics*, 34(13):i386–i394.

Briefings in Bioinformatics
doi.10.1093/bioinformatics/xxxxx
Advance Access Publication Date: Day Month
Manuscript Category

Subject Section

Recent Advances in Network-based Methods for Disease Gene Prediction

Sezin Kircali Ata [1], Min Wu [2], Yuan Fang [3], Le Ou-Yang [4], Chee Keong Kwoh [1] and Xiao-Li Li [2,*]

[1]School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore
[2]Institute for Infocomm Research, 138632, Singapore
[3]School of Information Systems, Singapore Management University, 188065, Singapore and
[4]College of Information Engineering, Shenzhen University, Shenzhen 518060, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Disease-gene association through Genome-wide association study (GWAS) is an arduous task for researchers. Investigating single nucleotide polymorphisms (SNPs) that correlate with specific diseases needs statistical analysis of associations. Considering the huge number of possible mutations, in addition to its high cost, another important drawback of GWAS analysis is the large number of false-positives. Thus, researchers search for more evidence to cross-check their results through different sources. To provide the researchers with alternative and complementary low-cost disease-gene association evidence, computational approaches come into play. Since molecular networks are able to capture complex interplay among molecules in diseases, they become one of the most extensively used data for disease-gene association prediction. In this survey, we aim to provide a comprehensive and up-to-date review of network-based methods for disease gene prediction. We also conduct an empirical analysis on 14 state-of-the-art methods. To summarize, we first elucidate the task definition for disease gene prediction. Secondly, we categorize existing network-based efforts into network diffusion methods, traditional machine learning methods with handcrafted graph features and graph representation learning methods. Thirdly, an empirical analysis is conducted to evaluate the performance of the selected methods across seven diseases. We also provide distinguishing findings about the discussed methods based on our empirical analysis. Finally, we highlight potential research directions for future studies on disease gene prediction.
Key words: disease gene prediction; network-based methods; graph representation learning

Introduction

Genetic diseases are mostly caused by gene mutations, although recent studies reveal that epigenetic factors can also play a role [1]. Among the existing methods, linkage analysis and genome-wide association studies (GWAS) are the most fundamental approaches in disease gene prediction, as they can provide predictive biomarkers through the genetic variation studies among humans, known as single nucleotide polymorphisms (SNPs). Nevertheless, statistical analysis and biological validation of the biomarkers are costly and time consuming, as a large amount of false positives need to be analyzed further [2]. Additionally, these techniques are simply based on direct genotype-phenotype associations. However, biological molecules perform their functions in corresponding pathways in a collaborative fashion. Projecting and characterizing their specific roles and collaborations onto a wired network/graph structure can reveal more useful knowledge and provide more systematical aspects. Furthermore, in a network-based environment, the disease causing factors, such as genetic mutations, epigenetic factors and pathogens, can be tracked more efficiently by chasing network perturbations, i.e., edge or node removals, in the molecular networks [3]. Therefore, molecular networks are efficient and effective data representations, which are able to model complex interplay among the molecules through a wider viewpoint and track the

potential disruptions on the biological pathways due to the disease-causing factors. As such, they have been extensively used by the computational approaches to complement and enrich existing linkage analysis and GWAS studies.

Currently, there are several molecular networks available to describe relationships between genes, such as protein-protein interaction (PPI) networks, gene regulatory networks, gene co-expression and metabolic interaction (MI) networks. Among these networks, PPI networks are the most extensively leveraged for disease-gene association prediction. One key reason is that proteins perform a vast array of critical functions to sustain an organism's well-being. Some of these functions include biochemical reactions, metabolic reaction catalysing, DNA replication, transmission of signals between cells, and maintaining structure for cells and tissues. More specifically, proteins collaborate and interact with each other to perform biological functions, leading to many protein interactions, which can be integrated and modelled as a graph/network data structure. Given a protein interaction, a corresponding gene mutation from one of the two proteins could make existing interaction impossible, and thus loses certain important biological functions and causes diseases. Another reason is, several studies using a PPI network embody the assumption that the position of a protein is not random. Proteins associated with a common set of biological properties tend to have common topological properties in the network such as node degree and centrality [4, 5]. Therefore, PPI networks can be employed in revealing protein-disease associations through the useful network-based features for proteins. However, existing PPI networks are incomplete, i.e., only a fraction of real protein interactions are detected through high-throughput experiments. Likewise they are noisy due to biased experimental evidence towards much studied disease genes [6]. Thus, an integrative approach covering various aspects of proteins such as GWAS, gene-expression, gene ontology, and other domain knowledge is both important and necessary.

In this survey, we focus on reviewing the computational methods leveraging network/graph data for disease gene prediction. First, we introduce the problem definition (i.e., node classification and link prediction) for disease gene prediction with different types of graph inputs. Second, we classify the network-based methods into three categories and provide a brief introduction for these methods. Third, we select representative methods from each category and conduct a comprehensive empirical study on them. The three categories are listed in the following.

- **Network diffusion methods**. The diffusion methods employ random walk techniques for influence propagation in different networks (e.g., PPI networks or phenotype-gene networks) for disease gene prediction.
- **Machine learning methods with handcrafted graph features**. Various features for diseases and genes are first extracted from input graph data, and then fed into traditional machine learning models (e.g., Random Forest) for predicting disease-gene associations.
- **Graph representation learning methods**. Instead of using handcrafted features for disease gene prediction, graph representation learning methods automatically learn the latent features or embeddings for diseases and genes by matrix factorization, graph embedding and graph neural network techniques.

Former surveys on network-based methods [3, 5, 7], which were published about ten years ago, present pioneering endeavours on network analysis and bring to light the importance of network data for disease gene prediction. Nowadays, we witness the usefulness of incorporating network data in several areas ranging from drug discovery to disease gene identification through novel network-based algorithms. In this survey, we aim to bring together up-to-date methods for disease gene prediction and provide a wider perspective to this important problem. In addition, there are two recent surveys which review the recent network embedding efforts on biomedical networks [8, 9]. In [8], the authors present an overview of the existing network embedding methods and their applications in biomedical data science, e.g., pharmaceutical data analysis, multi-omics data analysis and clinical data analysis. However, they do not conduct any empirical evaluation for the introduced methods. In [9], the authors introduce the recent graph embedding methods and their applications in biomedical networks. In particular, they further select 11 graph embedding methods and perform a systematic evaluation on 5 different tasks, i.e., drug-disease association prediction, drug-drug interaction prediction, PPI prediction, protein function prediction and medical term semantic type classification. Yet, they do not touch the task on disease gene prediction and its various state-of-the-art approaches. In this survey, we aim to conduct an empirical analysis on disease gene prediction task using different methods from the above three categories. Besides, we further apply and evaluate recent graph embedding methods including heterogeneous network embedding and multi-view network embedding for disease gene prediction.

The rest of this survey is organized as follows. In Section 'Problem Definition for Disease Gene Prediction', we cast the task of disease gene prediction as an instance of node classification or link prediction based on different graph inputs. Then, we introduce various network-based disease-gene prediction methods in details in Section 'Network-based Methods for Disease Gene Prediction'. Next, in Section 'Empirical Comparison', we perform a comprehensive empirical evaluation for 14 representative methods. Finally, in Section 'Future Perspectives', we discuss potential future research directions in disease-gene association prediction problem and we conclude this paper in Section 'Conclusion'.

Problem Definition for Disease Gene Prediction

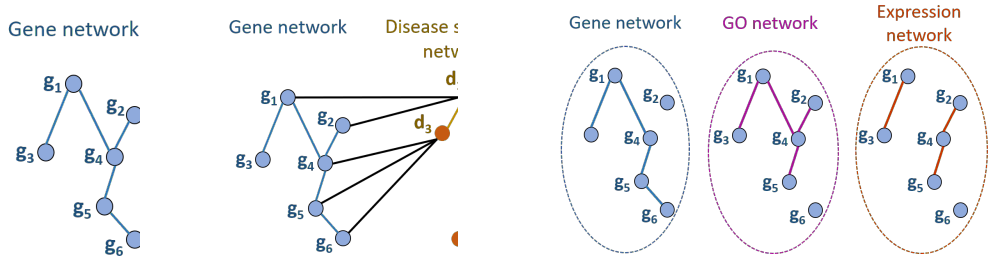
Network-based disease gene prediction leverages graph/network data as inputs to predict disease-causing genes. In particular, different types of graphs have been exploited for this purpose, including homogeneous graphs, heterogeneous graphs and multi-view/multiplex graphs. A homogeneous graph refers to a graph with a single type of nodes and a single type of edges, while a heterogeneous graph contains different types of nodes and edges. In addition, a multi-view or multiplex graph is a collection of graphs with the same set of nodes and different types of edges (e.g., edges from different views). Figure 1 shows the examples of different types of graph inputs. PPI network in Figure 1(a) is a homogeneous graph, while phenotype-gene network in Figure 1(b) is a heterogeneous graph. Figure 1(c) shows a multi-view graph for proteins, containing three views from the perspectives of PPI, GO similarity and gene expression.

Based on the above graph inputs, network-based disease gene prediction can be treated as a node classification or link prediction task. Node classification aims to infer the disease label of the unlabelled genes by utilizing the known disease genes, whereas link prediction aims to predict disease causing genes by utilizing gene-disease associations. Next, we give a formal definition for these two tasks from the perspective of disease gene prediction.

Node Classification

Figure 2(a) shows the node classification tasks, which is to predict the label of the genes of which disease associations are unknown, given known labels on some genes/nodes.

More formally, assume that we have a homogeneous graph $G = (V, E)$ where V is the set of nodes/genes and E shows relationships between nodes. A subset of genes $V_{labeled} \subset V$ represents known disease-causing genes, while $V_{unknown} = V \setminus V_{labeled}$ represents the set of genes of which disease associations are unknown. The formulation of node classification on G is to predict the labels for the nodes in $V_{unknown}$.



(a) A homogeneous graph. (b) A heterogeneous graph. (c) A multi-view graph with three views.

Fig. 1: Different types of graph inputs for network-based methods: (a) homogeneous graph, (b) heterogeneous graph and (c) multi-view graph.

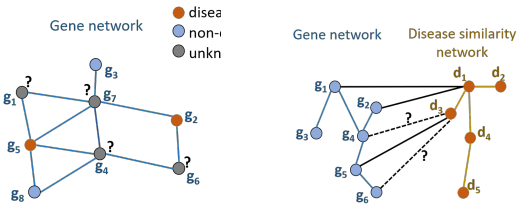
This node classification task in a heterogeneous graph or multi-view graphs can be similarly defined.

Link Prediction

The link prediction is a common task to reveal relationships between the objects especially in recommendation systems and social network analysis. Given a network G, denoted as $G = (V, E)$ and the nodes $v_i, v_j \in V$, the task of link prediction is to provide a measure of proximity/similarity between the nodes v_i and v_j.

In particular, we consider a heterogeneous graph $G = (U, V, E)$ for gene-disease association prediction. The vertices are grouped into two sets U and V, representing the set of genes and the set of diseases, respectively. E includes the edges in U, the edges in V and the edges between U and V (i.e., known disease-gene associations). The goal of gene-disease association prediction is to predict unknown links between U and V.

Figure 2(b) shows the disease-gene association prediction in a heterogeneous disease-gene network. The edges in black show the existing eanwhile, the dashed line pair of g_4 and d_3, which is to be predicted by computational methods.



(a) Node classification. (b) Link prediction.

Fig. 2: Tasks in disease gene prediction.

Network-based Methods for Disease Gene Prediction

In this section, we present a comprehensive review of network-based methods for disease gene prediction, namely diffusion-based methods, traditional feature-based methods and graph representation learning methods as illustrated in Figure 3.

Network Diffusion Methods

Diffusion-based methods working on biological networks are mostly adopted from pioneer graph-based semi-supervised algorithms such as [10, 11]. In disease gene prediction problem, they are widely exploited in the analysis of the biological pathways especially the pathways that are closest to the known disease genes. Starting from known disease genes, they diffuse along the biological network through random walks. Next, we introduce diffusion methods by walking in different networks.

Random Walk in PPI Network

Random walk with restart (RWR) is the most extensively used algorithm especially in prioritization of disease genes. Differently than a basic random walk, it is able to jump back to any seed node with probability r at each iteration. Formally, it is defined as follows:

$$p^{t+1} = (1 - r)W p^t + r p^0, \quad (1)$$

where W is the column-wise normalized adjacency matrix of the network. In particular, p^t is the probability vector of being at node i at time step t in its i-th entry and p^0 is the initial probability vector holding the probabilities of being at known disease nodes (seeds) [12]. Initially, each known disease protein has a uniformly distributed probability as the sum of the probabilities equal to 1. Total number of iterations is determined by the condition which satisfies the L1 norm difference between p^t and p^{t+1} smaller than a pre-defined threshold (e.g., 10^{-6}).

PRINCE [13] adopts RWR to a weighted PPI network through a weight function and also utilizes the disease similarities as prior probabilities. On an unweighted PPI network, PRINCE basically performs random walk with restart (Eq. 1) with adjusted prior probabilities based on the disease similarity scores. VAVIEN [14] is proposed to measure the topological similarity of proteins by formulating their interactions with Pearson correlation coefficient as a topological profile. To compute this profile, they employ the random walk proximity as a feature between seed and candidate proteins, so that the proteins with similar interactions will have a similar topological profile. Then, they prioritize the candidate disease genes based on the topological profile scores. In [15], the authors introduce a method called ORIENT, which prioritizes the candidate disease genes with RWR. They perform RWR on the adjacency matrix where the weights of interactions close to the known disease genes are properly reinforced by computing the shortest path distance to the disease genes. In [16], the authors propose a method called DP-LCC to construct diffusion profiles for the disease genes and the candidate genes separately based on their RWR on the PPI and the phenotype similarity network. Candidate genes are prioritized according to their diffusion profile similarities with the query disease. NPDE [17] utilizes non-disease essential proteins by formulating a dual flow network propagation method to prioritize candidate disease

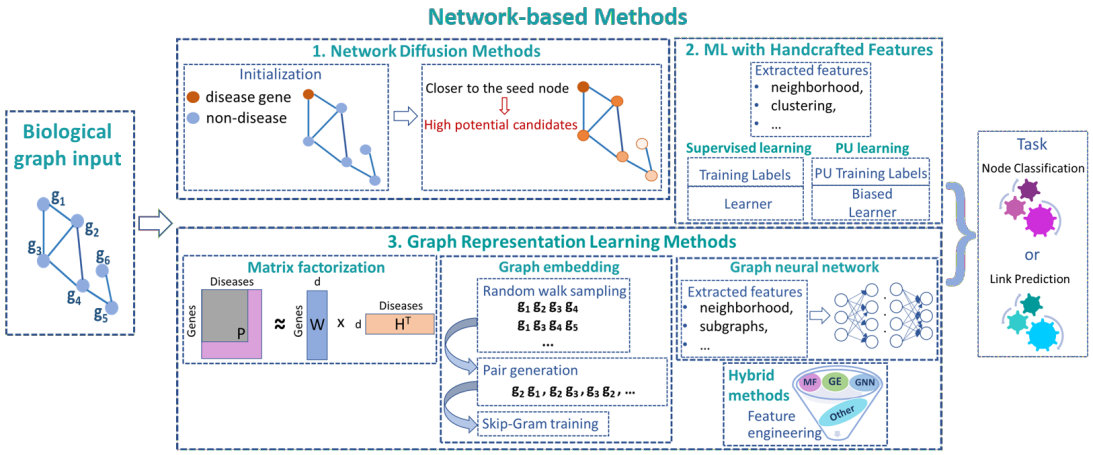


Fig. 3: Pipelines for the categories in network-based studies for disease gene prediction.

genes. It uses the assumption that if non-disease essential proteins exist as neighbors of a candidate protein, the protein is unlikely to cause a disease. Eventually it formulates this assumption as the negative flow in the prior information of the network propagation and the positive flow is allocated for the disease proteins.

Random Walk in Heterogeneous Network

For diffusion methods in heterogeneous networks, we further divide them into two categories with respect to their task as node classification or link prediction.

We start with node classification methods. BioGraph, [18], is a data integration and data mining platform. It utilizes stochastic model of random walks with restarts for a given candidate gene prioritization query as incorporating multiple data sources such as disease, pathway and GO annotation. In [19], two methods, CrossRank and CrossRankStar, formulate disease gene prioritization problem as optimization problems based on network propagation. They model two types of networks, network of networks (NoN) for CrossRank and networks of star networks (NoSN) model for CrossRankStar and both models incorporate tissue specific molecular networks. In SLN-SRW [20], the authors first aim to integrate biomedical data from heterogeneous sources including multiple ontologies and databases. For this propose, a simplified Laplacian normalization based supervised random walk algorithm is employed to learn the edge weights of the integrated network. RWR is then performed on this integrated network for disease gene prediction.

Link prediction methods aim to predict disease-gene associations in heterogeneous phenotype-gene network, where PPI network and phenotype network are connected through known phenotype-gene relationships. RWRH [21] aims to prioritize proteins based on their relevance to disease proteins. Basically, RWRH extends RWR algorithm to the heterogeneous phenotype-gene network through inter and intra transitions between PPI network and phenotype network. It is performed for a given query disease and corresponding disease genes are considered as seed nodes. In [22], a RWRH-based method called RWPCN is proposed to predict and prioritize disease genes on an integrated network comprising human protein complexes, protein interaction network, and phenotype similarity network. The basic difference from RWRH is that RWPCN operate random walks in an additional protein-complex networks. Likewise, a RWRH-based approach to predict malaria-associated genes on an integrated network by integrating human-human, parasite-parasite and human-parasite protein interactions in [23]. BiRW [24], aims to capture circular bigraph patterns based on the assumption that the relation among phenotype-gene associations can be characterized by these patterns. For this purpose, they employ a bi-random walk algorithm and aim to

Table 1. Various network diffusion methods for disease gene prediction.

Categories	Methods	Method descriptions
RW in PPI network (Node classification)	PRINCE [13]	Network propagation on a PPI network.
	VAVIEN [14]	Random walk and correlation on a PPI network.
	ORIENT [15]	Neighbor-favoring RWR on a PPI network.
	DP-LCC [16]	RWR-based diffusion profiles based on PPI and phenotype similarity network.
	NPDE [17]	Network propagation with dual flow on a PPI network.
RW in heterogeneous network (Node classification)	BioGraph [18]	Stochastic RW on an integrated network containing 21 publicly available curated databases.
	CrossRank [19]	Optimization based on network propagation on a heterogeneous network with various aspects.
	SLN-SRW [20]	Supervised RW to learn edge weights of the integrated network and RWR for prediction.
RW in heterogeneous network (Link prediction)	RWRH [21]	RWR on a heterogeneous phenotype-gene network.
	RWPCN [22]	RWR on a heterogeneous network with phenotypes, genes and protein complexes.
	RWRH-Malaria [23]	RWR on cross-species PPI networks for human and parasite.
	BiRW [24]	RWR on a heterogeneous network with a phenotype, gene and the phenotype-gene associations.
	RWRMH [25]	RWR on a multiplex heterogeneous network of protein interactions and disease networks.

reconstruct the missing associations globally through the gene-disease association prediction. In a recent study [25], the authors propose two extensions of RWRH, which are RWRM and RWRMH. Former method performs RWR on a multiplex network (i.e., all nodes are same type) composed of three layers of networks containing PPI, co-expression and pathway associations of proteins. Latter method incorporates a disease-disease network based on phenotype similarities, and gene-disease bipartite associations in addition to the aforementioned multiplex network as multiplex-heterogeneous network so that the random walker can jump to a network containing different sets of edges and nodes. Overall, a summary of different diffusion-based methods is presented in Table 1.

Machine Learning Methods with Handcrafted Features

Supervised machine learning methods

Supervised machine learning methods employ features and/or kernels to integrate various biological concepts, thus have been extensively studied in disease gene prediction. These methods work on the adjacency matrix of the network data. They extract features for genes/proteins based on various graph related measures such as shortest path distance, diffusion kernels, neighborhood with a disease protein, common neighbours, metapaths, metagraphs, etc. In [26], DERanking is proposed to benchmark four different strategies and prioritize candidate genes according to network analysis of their differential expression. The reason of incorporating differential expression as prior knowledge is based on the assumption that the strong disease candidates tend to be surrounded by differentially expressed neighbors. Three of the four benchmarking strategies are distinct random walk-based strategies from exponential diffusion kernel approach and one of them is direct neighborhood analysis. These strategies are performed on four networks comprising functional or psychical interactions of proteins for disease gene prediction. In [27], a method called BRIDGE is introduced for prioritization of disease genes by integrating various gene aspects including PPI, protein sequence data, gene expression data, KEGG database and GO, through a weighting scheme. This scheme is attained through a multiple linear regression model with lasso penalty, which determines the phenotypic similarity between two diseases based on the functional similarities between their associating genes. Accordingly, the model is capable of identifying genes associated with the diseases whose genetic bases are completely unknown. In [28], multiple aspects of proteins including known gene-disease associations, protein complexes, PPIs, pathways and gene expression profiles are integrated by Markov random field (MRF) and Bayesian analysis for disease gene prioritization. Initial prior probability of the MRF is set by Gibbs sampling. Metagraph representations [29] are proposed for disease gene prediction, which utilize both PPI network and biological annotations called *keywords* through heterogeneous subgraphs. By counting the occurrences of proteins within a specific subgraph type in terms of connectivity patterns, feature vectors of proteins are constructed, so that the proteins, which are functionally similar but located far away in a PPI network still have a chance to have similar representations in case they co-occur within the same subgraph type. In [30], the authors employ a multimodal deep belief net named dgMDL to predict disease–gene associations for all known diseases instead of predicting associated genes for a specific disease. This could alleviate the risk of overfitting due to the small number of positives and a large number of features in disease gene prediction problem. They first train two multi-modal DBN, one on PPI network and the other on GO-based similarity network, and then they train a final joint DBN for prediction based on the outputs of the initial DBNs. In [31], the proposed approach Disjunctive Graph Integration (DiGI) merges gene co-expression network, pathways, functional links, phenotype similarity database, co-functional network and PPI network in a single network. Then, DiGI performs a novel node kernel on this single network, which is a decomposition kernel with two strategies, i.e., decomposition by the k-decomposition core and subsequently the clique decomposition. These techniques are used to extract features for each node to be fed into a regularized linear SVM for disease gene prediction.

There are also supervised machine learning endeavours, which work on tissue-specific networks [32, 39, 40]. For instance, the proposed approach NetWAS [32] combines functional interaction network of genes and hierarchically corrected tissue expression to obtain hierarchy-aware tissue-specific knowledge. Then, the tissue-specific knowledge and human data compendium are integrated through regularized Bayesian integration to form tissue-specific functional networks. Finally, using the constructed networks, tissue-specific disease analysis is performed to identify disease associations. Their platform Genome-Scale Integrated Analysis of Networks in Tissues (GIANT) interface [41] further provides the tissue-specific maps and interactive visualizations.

As a final remark for supervised methods on networks, many of them have already employed the implicit semi-supervised learning assumption that samples close to each other tend to share the same labels. That is, the network structures provide intrinsic relationships between nodes (i.e., genes/proteins, diseases, etc.), which reveal valuable insights on related nodes and their tendency to associate with similar labels. The network structures are essentially label-free data that can aid supervised learning. Furthermore, additional multi-omics data (e.g., PPI, GO, gene expression, protein complexes, etc.) can be easily integrated into networks to further improve the performance of supervised learning.

PU learning methods

Supervised models in disease gene prediction may suffer due to all provided evidence serves as positive samples and the negative samples

Table 2. Machine learning methods with hand-crafted graph features for disease gene prediction.

Categories	Methods	Method descriptions
Supervised learning	DERanking [26]	4 ranking strategies based on whether a gene is surrounded by highly differentially expressed genes.
	BRIDGE [27]	A regression model with lasso penalty to weight 5 genomic data sources.
	IMRF [28]	An improved MRF method on multiple networks, e.g., PPI, pathways, protein complexes, etc.
	Metagraph+ [29]	Using metagraph features extracted from a heterogeneous network with PPI and gene keywords.
	dgMDL [30]	Multi-modal deep belief nets on PPI network and gene network based on GO similarities.
	DiGI [31]	A regularized linear SVM on the features extracted from a novel decomposition kernel.
	NetWAS [32]	Integrated analysis through regularized Bayesian integration of tissue-specific networks with SVMs.
PU learning	ProDiGe [33]	Biased SVM using features derived from multiple data sources [34] including PPI.
	PUDI [35]	Multi-level classification with biased SVM on features extracted from PPI, protein domain and GO.
	CATAPULT [36]	Biased SVM with features derived from walks in a heterogeneous phenotype-gene network.
	EPU [37]	Ensemble PU learning using various features derived from PPI and GO similarity networks.
	PEGPUL [38]	A perceptron ensemble of graph-based PU learning from 3 base classifiers SVM, KNN and CART.

remain unlabeled. Thus, positive-unlabeled (PU) methods arise to tackle this problem. They are usually based on weighting the positive samples and unlabeled samples to promote the prediction performance during the learning. ProDiGe [33] is a PU learning method for prioritization of candidate genes by enabling multiple data source integration such as PPI network and phenotype similarities through multitask learning strategy. In this way, scoring of the genes for prioritization is performed by considering disease-gene pairs instead of the individual genes. PUDI [35] partitions unlabeled set (i.e., negatives) into four sets namely, reliable negative set RN, likely positive set LP, likely negative set LN and weak negative set WN based on features extracted from PPI, GO terms and protein domain. The weighted support vector machines are then used to build a multi-level classifier based on these sets and positive training set. CATAPULT [36] performs a biased SVM to classify gene-phenotype pairs on a combined heterogeneous gene-trait network including gene-phenotype, gene-gene interactions across multiple species. Features for SVM are derived from walks in the network. EPU [37] is an ensemble-based PU learning method, which integrates data from multiple biological sources for training PU learning classifiers consists of weighted KNN, weighted naive Bayes and multi-level support vector machine. It starts with individual label propagation on gene expression network, PPI network and GO similarity network. Then, the obtained gene weights are combined and fed into the ensemble learner. PEGPUL [38] is an ensemble PU learning model to combine a multi-level SVM, a weighted KNN and a weighted decision tree for disease gene prediction. First, it extracts reliable negative genes by utilizing a co-training algorithm. Then, it constructs a similarity graph through metric learning by using PPI, protein domain and GO terms and performs a multi-rank-walk on the constructed graph to propagate labels prior to feeding the genes into the ensemble. A summary of supervised methods and PU learning methods is presented in Table 2.

Graph representation learning methods

Recall that we need to manually extract various features for the methods introduced in Table 2, which is tedious and requires domain knowledge. Recently, graph representation learning methods [42, 43], which can automatically learn the latent features/representations for the nodes, have acclaimed wide attentions in bioinformatics and biomedical applications [8, 9]. In this section, we review graph representation learning methods designed for disease gene prediction. In particular, we divide them into four categories, namely, matrix factorization, graph embedding, graph neural network and hybrid methods as shown in Figure 3.

Matrix factorization

Matrix factorization (MF) techniques have been widely used for link prediction in bioinformatics, e.g., PPI prediction [44], drug-target prediction [45, 46], miRNA-disease association prediction [47, 48], drug-pathway association prediction [49], etc. Here, we introduce various MF methods developed for disease-gene association prediction. Basically, MF methods in gene-disease association prediction, such as inductive matrix completion (IMC) [50], probability-based collaborative filtering (PCFM) [51] and manifold learning [52], aim to learn the latent factors for diseases and genes from the gene-disease association matrix. In addition, they can also specify the factorization process by including additional constraints into the objective function and thus MF techniques are useful for revealing important associations between diseases and genes.

In [50], IMC constructs gene and disease features using different sources including human gene–disease associations, gene-expression from different tissue samples, functional interactions between genes, gene–phenotype associations of other species, disease similarities and disease-related textual data from OMIM database. Then, they form the gene–disease association matrix using these features and formulate an optimization problem to recover unknown low-rank matrix using observations from the constructed associations matrix. Their inductive approach is capable of making predictions for a query disease with no previously known gene associations. In [51], PCFM exploits the gene–gene, gene–disease relationships, gene-disease linkages between orthologous genes and eight non-human species diseases and disease–disease similarity associations. They propose two probability-based collaborative filtering models, one with an average heterogeneous regularization and another with personal heterogeneous regularization using vector space similarity to predict gene-disease associations. The advantage of PCFM than the collaborative filtering is that PCFM enables probabilistic consideration in gene-disease associations instead of binary evaluation. Manifold learning [52] uses gene-disease association data for prediction and assumes that the geodetic distance between any associated gene-disease pairs are shorter than non-associated gene-disease pairs in a lower dimensional manifold. An optimization function is defined based on this assumption and singular value decomposition is employed to solve the optimization problem. Collage [53] applies collective matrix factorization (CMF) to combine a wide range of 14 data sets including RPKM-normalized RNA-seq transcriptional profiles and phenotype ontology annotations. Then, it performs chaining on the learned latent matrices to obtain gene profiles for prioritizing bacterial response genes in Dictyostelium. Similarly, Medusa [54] also builds a collective matrix factorization model for data fusion of a large-scale collections of heterogeneous data and performs chaining on the learned latent matrices to establish connections between non-neighboring nodes in the fusion graph. It formulates the growing of the modules as a sub-modular optimization program. The proposed method is capable of both associating genes with diseases and detecting disease modules. In [55], an unsupervised learning model based on matrix tri-factorization (tri-NMF) framework is proposed to detect disease causing genes from pan-cancer data. In particular, it exploits both the similarities of mutation profiles of different cancer types and gene interaction network. A method called GeneHound [56] adopts Bayesian probabilistic matrix factorization (BPMF) for disease gene prioritization problem. GeneHound uses gene-disease associations as partially observed data and a raw fusion is employed to integrate multiple genomic data sources including literature-based phenotypic and literature-based genomic information. Then, the Bayesian data fusion model jointly learns gene and disease latent factors and corresponding gene and disease-association matrices for disease-gene association prediction.

Graph embedding methods

Graph embedding methods learn low-dimensional and continuous vector representations of nodes through a neural network. For example, SkipGram [57] architecture is an extensively used architecture to construct associations between the node and its neighborhood. The neighborhood of the nodes is extracted through the random walks. Endeavours [58–60], such as DeepWalk [59] and node2vec [58], generate node representations such that the nodes lying within the short random walk distance have similar embeddings. There are also several random walk-based embedding methods proposed for disease gene prediction as follows.

SmuDGE [61] combines disease–phenotype and gene–phenotype associations with interactions between genes to generate a corpus for SkipGram-based representation learning. Then, it builds an artificial neural network (ANN) to predict gene–disease associations. In [62], HeteWalk builds a weighted heterogeneous network by joining six public data sources including PPI, miRNA similarity network, and disease phenotype similarity network and then performs SkipGram based network embedding. It further applies meta-path selection to eliminate potential redundant and misleading information caused by the heterogeneous

Table 3. Graph representation learning methods for predicting disease-gene associations.

Categories	Methods	Method descriptions
Matrix factorization	IMC [50]	Inductive matrix completion incorporating gene and disease features from multiple data sources.
	PCFM [51]	Probability-based collaborative filtering models with different regularization terms.
	Collage [53]	Collective matrix factorization using 14 datasets including genes, GO terms, KEGG pathways, etc.
	Medusa [54]	Collective matrix factorization on a data fusion graph with 16 data matrices.
	Tri-NMF [55]	Matrix tri-factorization on mutation profile similarities for different cancer types and PPI network.
	GeneHound [56]	Bayesian matrix factorization on OMIM associations with genomic and phenotypic data sources.
Graph embedding	SmuDGE [61]	RW-based embedding and an ANN on pair of disease and gene feature vectors.
	HeteWalk [62]	RW-based embedding on a heterogeneous network with genes, miRNAs and diseases.
	HerGePred [63]	RW-based embedding on a heterogeneous network with diseases, symptoms, genes and GO terms.
Graph neural networks	GCAS [64]	Graph convolution on heterogeneous association network for rare diseases (HANRD).
	PGCN [65]	GCN to learn the embeddings for phenotypes and genes in a disease-gene heterogeneous network.
	VGAE [66]	VGAE to learn embeddings for diseases and genes in disease-gene networks for gene prioritization.
Hybrid methods	N2VKO [67]	Combines node2vec embeddings and handcrafted features.
	N2A-SVM [68]	Combines node2vec embeddings with an autoencoder on a PPI network.
	N2Vmotif [69]	High-order PPI structures combining node2vec embeddings and network motifs.
	GCN-MF [70]	Combines GCN with matrix factorization using both gene similarities and disease similarities.
	HNEEM [71]	An ensemble of 6 graph embedding methods in a disease-gene-chemical heterogeneous network.
	DW-GCN [72]	Integrates graph embedding (DeepWalk) and graph convolutional network.

walks with multiple entities. HerGePred [63] also performs SkipGram-based graph embedding on a heterogeneous network consisting of a PPI, disease-protein associations, protein-GO associations, and gene-disease associations. Eventually, HerGePred predicts novel disease-gene associations in two different manners. It can directly calculate the cosine similarity between the embedding vectors of the query disease and the proteins. It can also perform random walk on disease-gene network, which is reconstructed based on the calculated cosine similarities between embeddings.

Graph neural networks

Graph neural network (GNN) is an advanced deep learning model for graph data [73] and has been applied for various bioinformatics tasks [74–76]. Graph convolutional network (GCN), graph attention network (GAT) and graph auto-encoder (GAE) are representative GNN models. For example, GCN aims to learn node embeddings by implementing the convolution operation on a graph based on the properties of neighborhood nodes. Here, we will have a quick review on GNN models for disease gene prediction as follows.

In [64], the authors introduce GCAS, which is an adaptation of graph convolutional network for disease gene prioritization. First, they construct a heterogeneous network consisting of ontological associations as well as curated associations including genes, diseases and pathways from multiple sources. They then perform direct spectral convolution to successively propagate the influence to the neighborhood and infer novel disease-gene associations. In [77], the authors introduce a tool PRIORI-T, which employs disease-gene, phenotype-phenotype and phenotype-disease correlation pairs extracted from a corpus of rare disease MEDLINE abstracts for gene prioritization. These correlations are computed based on Pearson correlation coefficient and used to construct initial correlation network (ICN). Then, GCAS [64] is performed for gene prioritization on this network to obtain ranked gene list for each clinical case. In [65], a graph convolutional network-based disease gene prioritization method called PGCN exploits a heterogeneous phenotype-gene network and the additional information for the nodes (e.g., disease ontology similarity, gene expression data, etc) for disease gene prioritization. In [66], the authors introduce the variational graph auto-encoder (VGAE) for gene-disease association prediction in a heterogeneous disease-gene network.

Hybrid methods

Here, we denote the methods, which combine the representation learning methods with other techniques to derive features, as hybrid methods.

A number of methods combine graph embedding (e.g., node2vec) with other methods for feature engineering. In [67], a method called N2VKO integrates the node2vec embeddings extracted from PPI network with the biological annotations for disease-gene association prediction. In [68], the proposed method N2A-SVM employs node2vec embedding of the genes from a PPI network and then performs dimension reduction with auto-encoder to predict Parkinson's disease genes. In [69], the authors propose to combine graphlet representations with node2vec embeddings for disease gene prediction.

[70] builds a network-based framework for gene-disease association prediction, which exploits disease similarity and gene similarity graphs to construct two GCNs separately for diseases and genes. Then, GCNs are trained by using corresponding disease and gene features and optimized with label information with matrix factorization. In [71], the authors propose a heterogeneous network embedding method, HNEEM, for gene-disease association prediction by ensemble learning. The constructed heterogeneous network consists of gene-disease associations, gene-chemical associations ans disease-chemical associations. HNEEM first extracts graph embeddings with six embedding methods and then feeds these embeddings into a random forest classifier for disease gene prediction. DW-GCN [72] is proposed to combine DeepWalk and GCN for gene-disease association prediction on a heterogeneous disease-gene network. Final predictions are derived using the output of a GCN decoder and the probability distribution derived from DeepWalk. Table 3 summarizes various graph representation learning models for disease gene prediction.

Empirical Comparison

In this section, we conduct experiments to evaluate the performance of various network-based methods for disease gene prediction.

Experimental Setup

Datasets

We collect two networks for proteins, namely, a PPI network from the IntAct database [78] and a protein functional similarity network based

on GO terms [79]. In particular, we first calculate the pairwise functional similarity between proteins based on G-SESAME[1] [80], and subsequently build a graph using the K nearest neighbor (KNN) algorithm. We set $K = 10$ in all of our experiments as its GO KNN graph can help node2vec to achieve the best performance for disease gene prediction. More details about GO KNN graph construction can be found in our supplementary materials. Overall, there are a total of 12,901 nodes (i.e., proteins) in both networks, with 96,845 edges in the PPI network and 107,508 edges in protein functional similarity network.

There are several publicly available databases for gene-disease associations as shown in Table 4. In our experiments, we acquired the associations from the OMIM database. Given a specific disease/phenotype (e.g., Alzheimer's disease), we extract MIM IDs from OMIM Morbid Map and retrieve their corresponding protein IDs from the Uniprot [81] conversion tool. Each protein node is subsequently assigned a binary label to represent whether it is a causative protein for this disease. In our experiments, we focus on seven diseases, namely, Alzheimer's disease (11), breast cancer (24), colorectal cancer (34), diabetes mellitus (37), obesity (19), lung cancer (15) and prostate cancer (17). Note that the number of positive proteins for each disease is included in the parentheses. The data and supplementary materials are available at https://github.com/sezinata/SurveyDGP.

Table 4. Publicly available databases for gene-disease associations.

Name	URL	Latest Update
OMIM [82]	https://omim.org	July, 2020
DisGeNet [83]	https://www.disgenet.org	June, 2020
MalaCards [84]	https://www.malacards.org	March, 2020
COSMIC [85]	https://cancer.sanger.ac.uk	April, 2020
PsyGeNET [86]	http://www.psygenet.org	January, 2018
CTD [87]	http://ctdbase.org	July, 2020

Selected methods for evaluation

We select a subset of methods from different categories for evaluation. For example, we select RWRH [21] from network diffusion methods and IMC [50] from matrix factorization methods. For machine learning methods with hand-crafted features, we select Metagraph+ [29] from supervised methods, and Catapult [36] and ProDiGe [33] from PU learning methods. Furthermore, we select a graph embedding method node2vec [58] and a hybrid method N2VKO [67].

As aforementioned, disease gene prediction is a typical node classification or link prediction task. Therefore, the state-of-the-art social network analysis methods for node classification or link prediction can be exploited for disease gene prediction. In our experiments, we also include HIN2Vec [88], HeGAN [89], MVE [90], mvn2vec [91], DMNE [92] and MANE [93] in our evaluation study.

- ***HIN2Vec*** [88]: A heterogeneous network embedding approach, which samples heterogeneous paths called meta-paths and feeds them into a neural network. We combined PPI network (PPI view) and functional similarity network (GO view) to form a single heterogeneous graph, where edges from different views are assigned to a different type of relation.
- ***HeGAN*** [89]: A heterogeneous network embedding approach, which utilizes the adversarial principle. We perform it on the constructed single network as described in *HIN2Vec*.
- **MVE** [90]: A state-of-the-art multi-view network embedding approach which maintain the collaboration between views by regularizing the Euclidean norm between view-specific embeddings and the final embeddings. The parameter η is used to control the weight of regularization. Since we conduct our experiments on unsupervised models, we adopt the unsupervised version.
- **mvn2vec** [91]: A state-of-the-art multi-view embedding approach. There are two proposed versions. Since they both have similar results we report only mvn2vec-r version which regularizes the Euclidean norm between view-specific embeddings, controlled by a hyperparameter γ. When γ is set to zero, it becomes equivalent to node2vec performed on a single view.
- ***DMNE*** [92]: A multi-view network embedding algorithm which is also capable of generating embeddings for many-to-many node mappings across views. Note that in our dataset only one-to-one mappings exist. We adopt their proximity disagreement formulation, due to its flexible assumption and better empirical performance.
- ***MANE*** [93]: A random-walk sampling-based multi view embedding algorithm, which unifies diversity, first-order collaboration and novel second-order collaboration principles in a framework. There are two hyperparameters α and β to regularize the contribution of the principles.

The datasets that we utilize for the studied methods in this survey are as follows. node2vec and RWR are simply performed on the PPI network. In IMC, Catapult, ProDiGe and RWRH we use the PPI network, phenotype similarity network [94] and the protein-phenotype associations. Metagraph+ and N2VKO leverage the PPI network and protein-keyword associations retrieved from Uniprot [81]. HIN2Vec, HeGAN, MVE, mvn2vec-r, DMNE and MANE utilize GO and PPI networks described in Datasets section. For methods using protein-phenotype associations, we eliminate the test data associations to prevent data leakage.

All methods are performed using the implementations provided by their respective authors unless stated otherwise, and we apply their suggested parameter settings for the hyperparameters. Table 5 shows the source code availability of the selected methods. Next, we briefly summarize the advantages and limitations of these selected methods. IMC, which is an MF method, basically aims to learn the latent factors of gene-disease association matrix. It is capable of predicting disease genes even for a disease that has no known associated genes. However, it generally does not provide global optimal solutions and has difficulty in convergence even for local optimal solutions [95]. Catapult and ProDiGe are useful for unbalanced data through biased SVMs, while they usually require tuning effort. RWRH is a random-walk based method. Although random-walk based methods are powerful in disease gene prediction, their performance might depend on the restart probability. Metagraph+ aims to identify candidate disease genes by capturing similarities through heterogeneous subgraphs incorporating both protein interactions and attributes. However, it is an arduous task to extract subgraph-level features (i.e., metagraphs). node2vec, N2VKO, mvn2vec-r, MVE and MANE are all subject to random walk-based sampling performance. However, mvn2vec-r, MVE and MANE might be able to provide more robust embeddings with the advantage of exploiting multi-view networks. HIN2Vec and HeGAN have the advantage of incorporating meta-paths. However, the former demands more training cost to capture the similarities between nodes and the latter requires a deliberate choice of pre-trained embeddings for initialization.

We adopt two well-known metrics for performance evaluation, namely, area under ROC curve (AUC) and area under precision-recall curve (AUPR). The performance of the models is evaluated through a stratified five-fold cross-validation. Given a specific disease, its causing genes are considered as positive samples and all the remaining genes as negatives. In particular, we randomly divide the positive and negative samples into five

[1] G-SESAME: http://bioinformatics.clemson.edu/G-SESAME/

Table 5. Selected methods for disease gene prediction and their source codes.

Methods	Source Code Availability
IMC	https://bigdata.oden.utexas.edu/project/gene-disease
Catapult	http://www.marcottelab.org/index.php/Catapult
ProDiGe	http://cbio.ensmp.fr/prodige
RWRH	https://github.com/alberto-valdeolivas/RWR-MH
Metagraph+	https://github.com/sezinata/Metagraph
node2vec	https://github.com/aditya-grover/node2vec
N2VKO	https://github.com/sezinata/N2VKO
HIN2Vec	https://github.com/csiesheep/hin2vec
HeGAN	https://github.com/librahu/HeGAN
MVE	https://github.com/mnqu/MVE
mvn2vec-r	https://github.com/sezinata/mvn2vec-code
DMNE	https://github.com/nijingchao/dmne
MANE	https://github.com/sezinata/MANE

groups. For each round, we select one group of positive and negative samples as testing data and the other four groups as training data to calculate AUC/AUPR. Eventually we report the average AUC/AUPR over five rounds as the final AUC/AUPR score. For multi-view graph embedding models (i.e., MVE, mvn2vec-r, DMNE and MANE) we set walk length to 10, number of walks per node to 5, negative sampling size to 10, windows size to 3 and random walk parameters of node2vec (p, q) to 1. All methods have $D = 128$ as the dimension of the final embedding. Differently, for heterogeneous graph embedding model HIN2Vec we set walk length to 30 and number of walks per node to 10. For graph embedding methods (i.e., HIN2Vec, HeGAN, MVE, mvn2vec-r, DMNE and MANE), the learned final embeddings are fed into the logistic regression model.

Results and Discussion

We demonstrate the performance comparison of 14 state-of-the-art methods in Table 6 and Table 7 in terms of AUC and AUPR, respectively. In particular, we group these methods into three categories based on their most proper input data to provide more insight about the approaches, namely homogeneous graph, heterogeneous graph and multi-view graph. In our supplementary materials, we also performed over-sampling with SMOTE [96] for the network embedding methods. The oversampling results as well as the comparison results in terms of ranking-aware metrics are provided in our supplementary materials. Overall, the AUC and AUPR scores in Table 6 and Table 7 are generally correlated, i.e., a method performing better in AUC tend to perform better in AUPR as well, with a Pearson correlation coefficient of about 0.7. In addition, AUPR results are much lower compared to AUC due to the skewness of the datasets, and this is common for disease gene prediction in several studies [55, 70, 71]. Based on the results in Table 6 and Table 7, we can have the following observations.

Firstly, we can observe that RWR and node2vec working on a homogeneous PPI network achieve low performance in terms of both AUC and AUPR as expected. Therefore, we are motivated to integrate different data sources into a heterogeneous network or multi-view network to improve the prediction performance.

Secondly, various methods working on a heterogeneous network perform well for disease gene prediction. ProDiGe and Catapult, as PU learning methods, are very promising to address the imbalanced data issue for disease gene prediction and both incorporate the phenotype similarity network and phenotype-gene associations. In particular, ProDiGe achieves the second best average AUC of 0.7763 and the third best average AUPR of 0.0974 among the 14 selected methods. In addition, RWRH also achieves a stable performance and thus it is a good alternative to the network embedding methods. However, RWRH's performance depends on the tuning process, e.g., the setting of the restart probability as demonstrated in the supplementary materials. N2VKO and Metagraph+ first work on feature extraction/selection in a PPI and keyword network, and then undergo an oversampling procedure. Both of the methods achieve a decent performance. Note that the random walk parameters of node2vec (p, q) in N2VKO are not tuned to maintain consistency with random walk sampling based embedding methods. Heterogeneous graph embedding methods including HIN2Vec and HeGAN hinge on different principles and perform very well. HIN2Vec utilizes random walks and negative sampling with a neural network model, while HeGAN utilizes adversarial learning principle in which a discriminator and a generator compete with each other as in a mini-max game. In particular, HeGAN's performance highly depends on the used pre-trained embeddings as inputs and thus careful initialization is critical in this model.

Lastly, multi-view methods are also robust and useful tools for disease gene prediction. mvn2vec-r and MANE are two random walk-based embedding methods employing Skip-gram for learning the final embeddings, while DMNE utilizes a deep autoencoder in place of the Skip-gram architecture and also employs RW-based sampling to feed into the autoencoder. They can achieve relatively good performance in terms of AUC. In particular, MANE introduces a novel second-order collaboration and combines it with the previously studied principles in a unified framework. Therefore, it has achieved the best performance in average and outperforms ProDiGe and HIN2Vec by 2.5% and 4.8% in AUC, respectively. Meanwhile, MVE is an attention-based supervised algorithm and we employ its unsupervised version for a fair comparison in our experiments. It is thus reasonable to achieve a relatively low performance for MVE.

FUTURE PERSPECTIVES

In this section, we present possible future directions that may address current challenging issues for more accurately predicting disease genes.

Learning with Limited Labeled Data

Learning with limited labelled data has been a challenging task in disease gene prediction. Existing efforts to overcome this difficulty include PU learning and oversampling techniques. For example, PU learning methods [33, 36, 35] select likely-positive samples from the unlabeled data to tackle the problem. Oversampling of minority class samples (e.g., SMOTE [96]) is also a common strategy to address this challenge. However, both strategies might need a tremendous tuning effort for difficult scenarios to attain a satisfactory performance. Developing more efficient and accurate models leveraging these strategies could be a promising direction.

Recently, generative adversarial networks (GAN) have been successfully applied to augment the data for various tasks, e.g., image classification [97, 98], speech recognition [99], etc. It is thus worth investigating GAN-based techniques for disease gene prediction with limited labels in the graph data. In addition, researchers also employ GAN to boost the PU learning [100] and oversampling processes [101]. Therefore, it would be very interesting to see the efforts that combine GAN with PU learning or oversampling for disease gene prediction in the future.

Attention Mechanisms and Data Integration

Graph Attention Networks (GAT) [102], as extension of graph convolutional networks, assign different weights to different neighbours with masked self-attention layers. In particular, this self-attention operation enables the model to focus on more important neighbours. Due to the attention mechanism, GAT have been applied to generate accurate graph embeddings for various tasks in recommendation systems [103, 104]

Table 6. AUC performance comparison among the benchmark methods.

	Disease	Alzheimer	Breast Cancer	Colon Cancer	Diabetes	Lung Cancer	Obesity	Prostate Cancer	Avg
Homogeneous Network	RWR	0.7179	0.7606	0.6241	0.5446	0.4842	0.5056	0.5047	0.5917
	node2vec	0.7539	0.8997	0.8370	0.5103	0.6004	0.5500	0.5888	0.6772
Heterogeneous Network	N2VKO	0.8724	0.8201	0.8469	0.7477	0.7135	0.6427	0.6474	0.7558
	RWRH	0.8518	0.8857	0.9013	0.6138	0.7850	0.5544	0.5655	0.7368
	IMC	0.6688	0.7598	0.6618	0.6145	0.7555	0.6221	**0.8185**	0.7001
	Catapult	0.8241	0.8258	0.8748	0.6211	**0.8431**	0.6144	0.5762	0.7399
	ProDiGe	0.9041	0.8997	0.8875	0.6525	0.7943	**0.7856**	0.5107	0.7763
	Metagraph+	0.8978	0.7725	0.8981	0.7335	0.6709	0.5999	0.6702	0.7490
	HIN2Vec	0.7858	0.8775	0.9225	0.6995	0.7178	0.6850	0.6272	0.7593
	HeGAN	0.8423	0.9199	0.9036	0.6677	0.7224	0.5765	0.5131	0.7351
Multi-view Network	MVE	0.6543	0.8330	0.8520	0.6305	0.5722	0.4078	0.5339	0.6405
	mvn2vec-r	0.8756	0.9002	0.9187	0.6489	0.6482	0.6692	0.5088	0.7385
	DMNE	0.9357	0.7996	0.8496	**0.7526**	0.7152	0.7256	0.4727	0.7501
	MANE	**0.9660**	**0.9276**	**0.9244**	0.7157	0.6951	0.7339	0.6069	**0.7956**

Table 7. AUPR performance comparison among the benchmark methods.

	Disease	Alzheimer	Breast Cancer	Colon Cancer	Diabetes	Lung Cancer	Obesity	Prostate Cancer	Avg
Homogeneous Network	RWR	0.0063	0.0321	0.0140	0.0154	0.0049	0.0530	0.0123	0.0197
	node2vec	0.0411	0.0682	0.0749	0.0048	0.0046	0.0046	0.0060	0.0292
Heterogeneous Network	N2VKO	0.1592	0.0776	0.1185	0.0392	0.1083	0.0618	0.0357	0.0857
	RWRH	0.1760	0.1057	0.1411	0.0085	**0.1203**	**0.1322**	0.0042	0.0983
	IMC	0.0058	0.0094	0.0063	0.0056	0.0103	0.0073	0.0089	0.0076
	Catapult	**0.3537**	0.0718	0.0845	0.0056	0.0761	0.0109	0.0076	0.0872
	ProDiGe	0.3420	0.0732	0.0481	0.0154	0.1114	0.0384	**0.0532**	0.0974
	Metagraph+	0.3018	0.1308	0.1053	0.0159	0.0383	0.0679	0.0032	0.0947
	HIN2Vec	0.0165	0.0987	0.0972	0.0135	0.0517	0.0042	0.0178	0.0428
	HeGAN	0.1442	0.1167	0.1091	0.0129	0.0101	0.0047	0.0028	0.0572
Multi-view Network	MVE	0.0161	0.0622	0.0757	0.0283	0.0150	0.0016	0.0405	0.0342
	mvn2vec-r	0.0275	0.0868	**0.1570**	0.0072	0.0072	0.0794	0.0023	0.0525
	DMNE	0.0603	0.0205	0.0252	0.0123	0.0503	0.0877	0.0018	0.0369
	MANE	0.2277	**0.1889**	0.1252	**0.0435**	0.0850	0.0386	0.0084	**0.1025**

and bioinformatics [105]. Therefore, it is also promising to develop node-level attention mechanisms for diseases and genes in different types of graph inputs for network-based disease gene prediction.

Moreover, it is common to integrate different graph data sources through multi-view techniques [106]. In these multi-view techniques, it is important to reveal the contributions of each network to the final prediction performance, so that it would be possible to prioritize the networks based on their significance. Graph-level attention mechanisms [93] can be useful in multi-view graphs for this purpose. In addition, hierarchical attention mechanisms [107] can also be applied for disease gene prediction by combining both node-level and graph-level attentions.

Sampling Strategies for Multi-View Inference

The study N2VKO [67] demonstrates that node2vec embeddings achieve an inferior prediction performance on some diseases such as the obesity and prostate cancer. The reason is that the disease-associated proteins are scattered in the network with a greater hop-distance between them, which limits the prediction power of a RW-based sampling method (e.g., node2vec). In these cases, it is good to develop models that can utilize other structures in the network. Specifically, this problem can be tackled by sampling strategies which may not consider only the local neighborhood of a node in a network. For example, we can adopt a collaborative sampling strategy, which can consider all the views of multi-view networks or utilize the attributed nodes during the sampling procedure. It would thus minimize the effect of local neighborhoods and help to increase the prediction performance.

Single-cell Data

Recently, single-cell RNA-seq (scRNA-seq) techniques become popular with the advancements in next-generation sequencing. Compared to traditional bulk RNA-seq analysis, they enable cell-level sequencing to capture cell-to-cell heterogeneity. In particular, several methods have been developed to infer gene relationships [108] and gene regulatory networks [109] from single-cell gene expression data. Disease gene prediction can thus benefit from such inferred gene relationships and gene regulatory networks.

In addition, precise identification of cell states and types is crucial to understand the disease related mechanisms so that the scientists can detect correlated expression levels of genes across a homogeneous population of cells [110]. Community detection algorithms such as louvain [111] and its similar version leiden [112] are very popular tools to identify cell clusters. Furthermore, it is possible to model single-cell gene expression data as graphs and we can thus employ graph embedding methods for cell type identification through graph clustering [113].

Explainable Machine Learning Models

Decision trees, linear regressions and logistic regressions are commonly used explainable machine learning models. The rise of the neural networks and deep learning in many areas such as video, speech and text processing

comes with the need of explanation for the "black box" nature of these models. As we know, it is very important for medical disease-related predictions to provide information on why the model performs a certain prediction. However, the graph representation learning methods covered in this review are also lack of interpretability. We are thus highly motivated to develop explainable machine learning models for disease gene prediction.

Recently, knowledge graphs (KG) have been integrated with user-item graphs for accurate and explainable recommendation [104, 114]. A recent work [115] adopts KG and graph neural networks for explainable drug-drug interaction (DDI) prediction. Similarly, we can also construct knowledge graphs for diseases and genes using various side information. For example, we can obtain the medical KG with diseases, drugs and symptoms, and gene ontology KG with genes and their functional annotations. By integrating advanced graph neural network techniques (e.g., GAT) and knowledge graphs, we expect to achieve accurate and explainable predictions for novel disease-gene associations.

Conclusion

Discovering disease causing genes and analyzing their roles in the disease are not only critical for understanding disease formation mechanism, but also extremely important for designing appropriate drugs for corresponding clinical therapies. Linkage analysis and genome-wide association studies (GWAS) form the basis of disease gene prediction. However, they generate a large number of false positives in their statistical analysis of biomarkers. Computational approaches are efficient and complementary tools to help biologists filter out noisy false positives and provide a list of genes which are worth for further clinical study. In this survey, we focus on network-based research, leveraging various networks in their problem formulation for disease gene prediction. We provide an organized, up-to-date overview of state-of-the-art network-based approaches. We also perform an empirical comparison study on different computational methods based on different graph inputs.

Generally, the methods in both heterogeneous network and multi-view network perform very well for disease gene prediction. In particular, multi-view methods with Skip-gram architecture can serve as robust and useful tools, not only for disease gene prediction but also for visualization and clustering purposes. Its low computational cost with a minimal tuning necessity and high prediction performance, make it a good alternative to the other learning methods. These techniques could be further investigated by considering attention mechanisms or different network sampling strategies instead of random walks. Moreover, for network-based disease gene prediction analysis, constructing reliable networks is critical, so future research need to focus on predicting novel interactions and removing noisy interactions. The ultimate success of the disease gene prediction will depend on the *parallel improvements* both in the experimental techniques by biologists and clinicians to provide rich and reliable biological data sets, and in the advanced computational techniques by computer scientists to provide efficient and robust ways to discover novel knowledge and insights from the biological data.

Biographical Note

Sezin Kircali Ata just obtained her PhD degree from School of Computer Science and Engineering Nanyang Technological University (NTU), Singapore and currently is a research fellow at KK Women's and Children's Hospital, Singapore. Her research interests include machine learning, graph mining and bioinformatics. **Min Wu** is currently a senior scientist at the Institute for Infocomm Research (I2R), A*STAR, Singapore. His research interests include machine learning, data mining and bioinformatics. **Yuan Fang** is currently an Assistant Professor at the School of Information Systems, Singapore Management University, Singapore. His research focuses on graph-based machine learning, Web and social media mining and recommendation systems. **Le Ou-Yang** is an assistant professor in the College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China. His research interest includes bioinformatics and machine learning. **Chee-Keong Kwoh** is currently an associate professor at the School of Computer Science and Engineering, NTU, Singapore. His research interests include data mining, soft computing and graph-based inference, bioinformatics and biomedical engineering. **Xiao-Li Li** is currently a department head and principal scientist at I2R, A*STAR, Singapore. His research interests include data mining, machine learning, AI, and bioinformatics.

Key Points

- **Uncovering disease-causing genes is a fundamental objective of human genetics, while computational prediction of disease-genes provides a low-cost alternative.**
- **We classified and reviewed state-of-the-art network-based approaches for disease gene prediction with different types of graph inputs.**
- **We empirically evaluated various selected methods, including some advanced methods for social network analysis, for disease gene prediction.**
- **We also discussed possible future directions that may address current challenging issues for more accurately predicting disease genes.**

References

[1] Fides Zenk, Eva Loeser, Rosaria Schiavo, Fabian Kilpert, Ozren Bogdanović, and Nicola Iovino. Germ line–inherited h3k27me3 restricts enhancer function during maternal-to-zygotic transition. *Science*, 357(6347):212–216, 2017.

[2] Sora Yoon et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Research*, 46(10):e60–e60, 03 2018.

[3] Xiujuan Wang, Natali Gulbahce, and Haiyuan Yu. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, 10(5):280, 2011.

[4] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140–n/a, 2007.

[5] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome Research*, 18(4):644–652, 2008.

[6] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 2015.

[7] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56, 2011.

[8] Chang Su, Jie Tong, Yongjun Zhu, Peng Cui, and Fei Wang. Network embedding in biomedical data science. *Briefings in Bioinformatics*, 21(1):182–197, 12 2018.

[9] Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*,

36(4):1241–1251, 2020.

[10] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 912–919. AAAI Press, 2003.

[11] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: Fast solutions and applications. *Knowl. Inf. Syst.*, 14(3):327–346, March 2008.

[12] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949 – 958, 2008.

[13] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 6(1):1–9, 01 2010.

[14] Sinan Erten, Gurkan Bebek, and Mehmet Koyutürk. Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J Comput Biol*, 18(11):1561–1574, Nov 2011.

[15] Duc-Hau Le and Yung-Keun Kwon. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. *Computational Biology and Chemistry*, 44:1 – 8, 2013.

[16] Jie Zhu, Yufang Qin, Taigang Liu, Jun Wang, and Xiaoqi Zheng. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. *BMC Bioinformatics*, 14(5):S5, 2013.

[17] Shunyao Wu, Fengjing Shao, Jun Ji, Rencheng Sun, Rizhuang Dong, Yuanke Zhou, Shaojie Xu, Yi Sui, and Jianlong Hu. Network propagation with dual flow for gene prioritization. *PLOS ONE*, 10(2):1–15, 02 2015.

[18] Anthony ML Liekens, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, and Jurgen Del-Favero. Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, 12(6):R57, 2011.

[19] Jingchao Ni, Mehmet Koyuturk, Hanghang Tong, Jonathan Haines, Rong Xu, and Xiang Zhang. Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model. *BMC Bioinformatics*, 17(1):453, Nov 2016.

[20] Jiajie Peng, Kun Bai, Xuequn Shang, Guohua Wang, Hansheng Xue, Shuilin Jin, Liang Cheng, Yadong Wang, and Jin Chen. Predicting disease-related genes using integrated biomedical networks. *BMC Genomics*, 18(1):1043, 2017.

[21] Yongjin Li and Jagdish C. Patra. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

[22] Peng Yang, Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Inferring gene-phenotype associations via global protein complex network propagation. *PLOS ONE*, 6(7):1–11, 07 2011.

[23] Yang Chen and Rong Xu. Network-based gene prediction for plasmodium falciparum malaria towards genetics-based drug discovery. *BMC Genomics*, 16(7):S9, 2015.

[24] MaoQiang Xie, YingJie Xu, YaoGong Zhang, TaeHyun Hwang, and Rui Kuang. Network-based phenome-genome association prediction by bi-random walk. *PLOS ONE*, 10(5):1–18, 05 2015.

[25] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anais Baudot. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505, 07 2018.

[26] Daniela Nitsch, Joana P. Gonçalves, Fabian Ojeda, Bart de Moor, and Yves Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460, 2010.

[27] Yong Chen, Xuebing Wu, and Rui Jiang. Integrating human omics data to prioritize candidate genes. *BMC Medical Genomics*, 6(1):57, 2013.

[28] Bolin Chen, Jianxin Wang, Min Li, and Fang-Xiang Wu. Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics*, 7(2):S2, 2014.

[29] Sezin Kircali Ata, Yuan Fang, Min Wu, Xiao-Li Li, and Xiaokui Xiao. Disease gene classification with metagraph representations. *Methods*, 131:83–92, 2017.

[30] Ping Luo, Yuanyuan Li, Li-Ping Tian, and Fang-Xiang Wu. Enhancing the prediction of disease-gene associations with multimodal deep learning. *Bioinformatics*, 35(19):3735–3742, 2019.

[31] Van Dinh Tran, Alessandro Sperduti, Rolf Backofen, and Fabrizio Costa. Heterogeneous networks integration for disease-gene prioritization with node kernels. *Bioinformatics*, 36(9):2649–2656, 01 2020.

[32] Casey S. Greene, Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, Jun 2015.

[33] Fantine Mordelet and Jean-Philippe Vert. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12(1):389, 2011.

[34] Tijl De Bie, Léon-Charles Tranchevent, Liesbeth MM Van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–i132, 2007.

[35] Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 08 2012.

[36] U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods, Inderjit S. Dhillon, and Edward M. Marcotte. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLOS ONE*, 8(5):1–17, 05 2013.

[37] Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and See-Kiong Ng. Ensemble positive unlabeled learning for disease gene identification. *PloS one*, 9(5), 2014.

[38] Gholam-Hossein Jowkar and Eghbal G Mansoori. Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification. *Computational biology and chemistry*, 64:263–270, 2016.

[39] Victoria Yao, Rachel Kaletsky, William Keyes, Danielle E. Mor, Aaron K. Wong, Salman Sohrabi, Coleen T. Murphy, and Olga G. Troyanskaya. An integrative tissue-network approach to identify and test human disease genes. *Nature Biotechnology*, 36(11):1091–1105, December 2018.

[40] Yuanfang Guan, Dmitriy Gorenshteyn, Margit Burmeister, Aaron K. Wong, John C. Schimenti, Mary Ann Handel, Carol J. Bult, Matthew A. Hibbs, and Olga G. Troyanskaya. Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLOS Computational Biology*, 8(9):1–12, 09 2012.

[41] Aaron K Wong, Arjun Krishnan, and Olga G Troyanskaya. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Research*, 46(W1):W65–W70, 05 2018.

[42] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, 2018.

[43] H. Cai, V. W. Zheng, and K. Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge & Data Engineering*, 30(09):1616–1637, 2018.

[44] Hua Wang, Heng Huang, Chris Ding, and Feiping Nie. Predicting protein–protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *Journal of Computational Biology*, 20(4):344–358, 2013.

[45] Yong Liu, Min Wu, Chunyan Miao, Peilin Zhao, and Xiao-Li Li. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Computational Biology*, 12(2):e1004760, 2016.

[46] Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(3):646–656, 2016.

[47] Xing Chen, Lei Wang, Jia Qu, Na-Na Guan, and Jian-Qiang Li. Predicting mirna–disease association based on inductive matrix completion. *Bioinformatics*, 34(24):4256–4265, 2018.

[48] Zi-Chao Zhang, Xiao-Fei Zhang, Min Wu, Le Ou-Yang, Xing-Ming Zhao, and Xiao-Li Li. A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics*, 36(11):3474–3481, 2020.

[49] Chun-Chun Wang, Yan Zhao, and Xing Chen. Drug-pathway association prediction: from experimental results to computational models. *Briefings in Bioinformatics*, 2020.

[50] Nagarajan Natarajan and Inderjit S. Dhillon. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, 30(12):i60–i68, Jun 2014.

[51] Xiangxiang Zeng, Ningxiang Ding, Alfonso Rodríguez-Patón, and Quan Zou. Probability-based collaborative filtering model for predicting gene–disease associations. *BMC Medical Genomics*, 10(5):76, Dec 2017.

[52] Ping Luo, Li-Ping Tian, Bolin Chen, Qianghua Xiao, and Fang-Xiang Wu. Predicting gene-disease associations with manifold learning. In *Bioinformatics Research and Applications*, pages 265–271, Cham, 2018.

[53] Marinka Žitnik, Edward A Nam, Christopher Dinh, Adam Kuspa, Gad Shaulsky, and Blaž Zupan. Gene prioritization by compressive data fusion and chaining. *PLoS Computational Biology*, 11(10), 2015.

[54] Marinka Zitnik and Blaz Zupan. Jumping across biomedical contexts using compressive data fusion. *Bioinformatics (Oxford, England)*, 32(12):i90–i100, 2016.

[55] Jianing Xi, Ao Li, and Minghui Wang. A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. *Neurocomputing*, 296:64–73, 2018.

[56] Pooya Zakeri, Jaak Simm, Adam Arany, Sarah ElShal, and Yves Moreau. Gene prioritization using bayesian matrix factorization with genomic and phenotypic side information. *Bioinformatics*, 34(13):i447–i456, 2018.

[57] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[58] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 855–864. ACM, 2016.

[59] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani, editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 701–710. ACM, 2014.

[60] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, 2015.

[61] Mona Alshahrani and Robert Hoehndorf. Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, 34(17):i901–i907, 2018.

[62] Yun Xiong, Mengjie Guo, Lu Ruan, Xiangnan Kong, Chunlei Tang, Yangyong Zhu, and Wei Wang. Heterogeneous network embedding enabling accurate disease association predictions. *BMC Medical Genomics*, 12(10):186, 2019.

[63] K. Yang, R. Wang, G. Liu, Z. Shu, N. Wang, R. Zhang, J. Yu, J. Chen, X. Li, and X. Zhou. Hergepred: Heterogeneous network embedding representation for disease gene prediction. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1805–1815, 2019.

[64] Aditya Rao, Saipradeep VG, Thomas Joseph, Sujatha Kotte, Naveen Sivadasan, and Rajgopal Srinivasan. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Medical Genomics*, 11(1):57, Jul 2018.

[65] Yu Li, Hiroyuki Kuwahara, Peng Yang, Le Song, and Xin Gao. Pgcn: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. *bioRxiv*, 2019.

[66] Vikash Singh and Pietro Liò. Towards probabilistic generative models harnessing graph neural networks for disease-gene prediction. *CoRR*, abs/1907.05628, 2019.

[67] Sezin Kircali Ata, Le Ou-Yang, Yuan Fang, Chee-Keong Kwoh, Min Wu, and Xiao-Li Li. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. *BMC Systems Biology*, 12(9):138, 2018.

[68] Jiajie Peng, Jiaojiao Guan, and Xuequn Shang. Predicting parkinson's disease genes based on node2vec and autoencoder. *Frontiers in genetics*, 10:226–226, Apr 2019.

[69] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. *Pac Symp Biocomput*, 23:111–122, 2018.

[70] Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, and Panos Kalnis. GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 705–713. ACM, 2019.

[71] Xiaochan Wang, Yuchong Gong, Jing Yi, and Wen Zhang. Predicting gene-disease associations from the heterogeneous network using graph embedding. In Illhoi Yoo, Jinbo Bi, and Xiaohua Hu, editors, *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA,*

November 18-21, 2019, pages 504–511. IEEE, 2019.

[72] Lvxing Zhu, Zhaolin Hong, and Haoran Zheng. Predicting gene-disease associations via graph embedding and graph convolutional networks. In Illhoi Yoo, Jinbo Bi, and Xiaohua Hu, editors, *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019*, pages 382–389. IEEE, 2019.

[73] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–21, 03 2020.

[74] Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, and Fei Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics*, 06 2019. bbz042.

[75] Ruichu Cai, Xuexin Chen, Yuan Fang, Min Wu, and Yuexing Hao. Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. *Bioinformatics*, 2020.

[76] Yahui Long, Min Wu, Chee Keong Kwoh, Jiawei Luo, and Xiaoli Li. Predicting human microbe-drug associations via graph convolutional network with conditional random field. *Bioinformatics*, 2020.

[77] Aditya Rao, Thomas Joseph, Vangala G. Saipradeep, Sujatha Kotte, Naveen Sivadasan, and Rajgopal Srinivasan. Priori-t: A tool for rare disease gene prioritization using medline. *PLOS ONE*, 15(4):e0231728, Apr 2020.

[78] Sandra Orchard et al. The mintact project intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–63, January 2014.

[79] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049, 2015.

[80] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[81] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 10 2014.

[82] Joanna S Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, 47(D1):D1038–D1043, 11 2018.

[83] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, 45(D1):D833–D839, 2017.

[84] Noa Rappaport, Michal Twik, Inbar Plaschkes, Ron Nudel, Tsippi Iny Stein, Jacob Levitt, Moran Gershoni, C. Paul Morrey, Marilyn Safran, and Doron Lancet. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, 45(D1):D877–D887, 11 2016.

[85] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018.

[86] Alba GutiErrez-Sacristan, Solene Grosdidier, Olga Valverde, Marta Torrens, Alex Bravo, Janet Pinero, Ferran Sanz, and Laura I. Furlong. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics*, 31(18):3075–3077, 05 2015.

[87] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Research*, 47(D1):D948–D954, 09 2018.

[88] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1797–1806. ACM, 2017.

[89] Binbin Hu, Yuan Fang, and Chuan Shi. Adversarial learning on heterogeneous information networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 120–129. ACM, 2019.

[90] Qu Meng, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. An attention-based collaboration framework for multi-view network representation learning. In *CIKM 2017 - Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, International Conference on Information and Knowledge Management, Proceedings, pages 1767–1776. ACM, 2017.

[91] Yu Shi, Fangqiu Han, Xinwei He, Xinran He, Carl Yang, Jie Luo, and Jiawei Han. mvn2vec: Preservation and collaboration in multi-view network embedding. *arXiv preprint arXiv:1801.06597*, 2018.

[92] Jingchao Ni, Shiyu Chang, Xiao Liu, Wei Cheng, Haifeng Chen, Dongkuan Xu, and Xiang Zhang. Co-regularized deep multi-network embedding. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 469–478. ACM, 2018.

[93] Sezin Kircali Ata, Yuan Fang, Min Wu, Jiaqi Shi, Chee Keong Kwoh, and Xiaoli Li. Multi-view collaborative network embedding. *CoRR*, abs/2005.08189, 2020.

[94] Marc A. van Driel, Jorn Bruggeman, Gert Vriend, Han G. Brunner, and Jack A. M. Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–542, 2006.

[95] L. Li, L. Wu, H. Zhang, and F. Wu. A fast algorithm for nonnegative matrix factorization and its convergence. *IEEE Transactions on Neural Networks and Learning Systems*, 25(10):1855–1863, Oct 2014.

[96] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

[97] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4941–4949. IEEE Computer Society, 2017.

[98] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[99] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):84–96, 2017.

[100] Ming Hou, Brahim Chaib-Draa, Chao Li, and Qibin Zhao. Generative adversarial positive-unlabeled learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 2255–2261. AAAI Press, 2018.

[101] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1695–1704. IEEE, 2019.

[102] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[103] Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Peng He, Paul Weng, Han Gao, and Guihai Chen. Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In *The World Wide Web Conference*, WWW '19, page 2091–2102. ACM, 2019.

[104] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. KGAT: knowledge graph attention network for recommendation. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 950–958. ACM, 2019.

[105] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 07 2018.

[106] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2):325–340, 2018.

[107] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics, 2016.

[108] Ye Yuan and Ziv Bar-Joseph. Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, 116(52):27151–27158, 2019.

[109] Giovanni Iacono, Ramon Massoni-Badosa, and Holger Heyn. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome biology*, 20(1):110, 2019.

[110] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

[111] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[112] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, Mar 2019.

[113] Xiangyu Li, Weizheng Chen, Yang Chen, Xuegong Zhang, Jin Gu, and Michael Q. Zhang. Network embedding-based representation learning for single cell RNA-seq data. *Nucleic Acids Research*, 45(19):e166–e166, 08 2017.

[114] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *KDD*, pages 968–977, 2019.

[115] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2739–2745. International Joint Conferences on Artificial Intelligence Organization, 2020.