Identifying disease-associated genes based on artificial intelligence

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Division of Biomedical Engineering
University of Saskatchewan
Saskatoon

By
Ping Luo

©Ping Luo, September, 2019, All rights reserved.

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Devision of Biomedical Engineering
Engineering Building
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5A9
Canada

OR

Dean of College of Graduate and Postdoctoral Studies
Room 116 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Identifying disease-gene associations can help improve the understanding of disease mechanisms, which has a variety of applications, such as early diagnosis and drug development. Although experimental techniques, such as linkage analysis, genome-wide association studies (GWAS), have identified a large number of associations, identifying disease genes is still challenging since experimental methods are usually time-consuming and expensive. To solve these issues, computational methods are proposed to predict disease-gene associations.

Based on the characteristics of existing computational algorithms in the literature, we can roughly divide them into three categories: network-based methods, machine learning-based methods, and other methods. No matter what models are used to predict disease genes, the proper integration of multi-level biological data is the key to improving prediction accuracy. This thesis addresses some limitations of the existing computational algorithms, and integrates multi-level data via artificial intelligence techniques. The thesis starts with a comprehensive review of computational methods, databases, and evaluation methods used in predicting disease-gene associations, followed by one network-based method and four machine learning-based methods.

The first chapter introduces the background information, objectives of the studies and structure of the thesis. After that, a comprehensive review is provided in the second chapter to discuss the existing algorithms as well as the databases and evaluation methods used in existing studies. Having the objectives and future directions, the thesis then presents five computational methods for predicting disease-gene associations.

The first method proposed in Chapter 3 considers the issue of non-disease gene selection. A shortest path-based strategy is used to select reliable non-disease genes from a disease gene network and a differential network. The selected genes are then used by a network-energy model to improve its performance. The second method proposed in Chapter 4 constructs sample-based networks for case samples and uses them to predict disease genes. This strategy improves the quality of protein-protein interaction (PPI) networks, which further improves the prediction accuracy. Chapter 5 presents a generic model which applies multimodal deep belief nets (DBN) to fuse different types of data. Network embeddings extracted from PPI networks and gene ontology (GO) data are fused with the multimodal DBN to obtain cross-modality representations. Chapter 6 presents another deep learning model which uses a convolutional neural network (CNN) to integrate gene similarities with other types of data. Finally, the fifth method proposed in Chapter 7 is a nonnegative matrix factorization (NMF)-based method. This method maps diseases and genes onto a lower-dimensional manifold, and the geodesic distance between diseases and genes are used to predict their associations. The method can predict disease genes even if the disease under consideration has no known associated genes.

In summary, this thesis has proposed several artificial intelligence-based computational algorithms to address the typical issues existing in computational algorithms. Experimental results have shown that the proposed methods can improve the accuracy of disease-gene prediction.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Fang-Xiang Wu, who offered me the opportunity to pursue the Ph.D. degree in his lab. Throughout my studies, he constantly provides me insightful suggestions, not only for research but also for my daily life. This work would not have been possible without his help.

I would also like to express my gratitude to other members of my advisory committee, Prof. Ian McQuillan, Prof. Mark Keil, and Prof. Francis Bui. I have received positive support and valuable suggestions from them during the Ph.D. program. Their courses also help me lay a solid foundation for my study.

In addition, I am lucky to learn and work with a group of great colleagues. I am particularly indebted to Yan Yan, Jinhong Shi, Yulian Ding, Pi-Jing Wei, Bolin Chen, Lin Wu, Fei Wang, Lingkai Tang, and Yichao Shen, for their help in both my life and research.

I would also like to thank my friends Wei Chen, Jinglin Gao, Qian Huang, Fan Zhang, Wen-Jing Zhang, Rebecca Mao, Tyler Zhang, Yaogeng Lei, Maodong Zhang, Lizhi Liu, Kang Jiang, and Qi Guo, for sharing their life experience with me, which has broadened my horizons.

Finally, I would like to thank all my families for their unconditional support and love. I am also grateful for the financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), University of Saskatchewan (UofS), the China Scholarship Council (CSC), and the family of Russell Haid.

This thesis is dedicated to my father Yuedong Luo, my mother Fang Wu, and my grandfather Jiazhang Wu, who encouged me all these years.

Contents

Permission to Use i

Abstract ii

Acknowledgements iii

Contents v

List of Tables ix

List of Figures x

List of Abbreviations xii

1 Introduction 1

1.1 Background . . . 1

1.2 Motivations and objectives . . . 1

1.3 Organization of the thesis . . . 3

2 Predicting disease-associated genes: computational methods, databases, and evaluations 4

2.1 Introduction . . . 5

2.2 Computational methods . . . 6

2.2.1 Network-based methods . . . 6

2.2.2 Machine learning-based methods . . . 10

2.2.3 Other methods . . . 14

2.3 Biological data . . . 15

2.3.1 Disease-gene associations . . . 15

2.3.2 PPI network . . . 17

2.3.3 Gene expression . . . 18

2.3.4 Mutation data . . . 19

2.3.5 Pathway . . . 20

2.3.6 Other types of data . . . 20

2.3.7 Data integration . . . 21

2.4 Evaluation methods . . . 23

2.4.1 Step 1: metrics . . . 23

2.4.2 Step 2: *de novo* study . . . 24

2.4.3 Other evaluation methods . . . 24
2.5 Perspectives and conclusions . . . 24

3 Disease gene prediction by integrating PPI networks, clinical RNA-Seq data and OMIM data 27

3.1 Introduction . . . 28
3.2 Methods and materials . . . 30
3.2.1 General model . . . 30
3.2.2 Differential network construction . . . 33
3.2.3 Non-disease genes . . . 35
3.2.4 Feature extraction . . . 37
3.2.5 Validation methods and evaluation criteria . . . 37
3.2.6 Data sources . . . 38
3.3 Results and discussion . . . 39
3.3.1 Threshold selection . . . 39
3.3.2 The results of AUC values . . . 40
3.3.3 Enrichment analysis . . . 40
3.3.4 Top 10 unknown genes . . . 44
3.4 Conclusion . . . 44

4 Ensemble disease gene prediction by clinical sample-based networks 48

4.1 Background . . . 49
4.2 Methods . . . 50
4.2.1 Sample-based networks . . . 50
4.2.2 Model design . . . 51
4.2.3 Network labeling and benchmark selection . . . 53
4.2.4 Ensemble prediction . . . 54
4.2.5 Datasets . . . 54
4.2.6 Evaluation metrics . . . 55
4.3 Results . . . 56
4.3.1 Clustering . . . 56
4.3.2 Sensitivity analysis . . . 57
4.3.3 Comparison . . . 58
4.3.4 *De novo* validation . . . 58
4.4 Discussion . . . 59
4.5 Conclusions . . . 60
4.6 Acknowledgements . . . 60

5 Enhancing the prediction of disease-gene associations with multimodal deep learning 64

5.1 Introduction . . . 65

5.2 Materials and methods . . . 67

5.2.1 RBM . . . 67

5.2.2 Multimodal DBN . . . 68

5.2.3 Raw feature extraction . . . 68

5.2.4 Evaluation metrics . . . 71

5.2.5 Hyperparameters . . . 72

5.2.6 Data sources . . . 72

5.3 Results . . . 73

5.3.1 Overall performance . . . 73

5.3.2 Comparison with other algorithms . . . 73

5.3.3 Prediction of new disease-gene associations . . . 74

5.4 Conclusion . . . 74

6 deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks 79

6.1 Introduction . . . 80

6.2 Material and methods . . . 81

6.2.1 General model . . . 81

6.2.2 Network-based convolution . . . 82

6.2.3 Mutation-based features . . . 82

6.2.4 Data sources . . . 83

6.2.5 Evaluation metrics . . . 84

6.2.6 Implementation . . . 85

6.3 Results . . . 85

6.3.1 Hyperparameters . . . 85

6.3.2 Cross-validation . . . 86

6.3.3 *De novo* study . . . 86

6.4 Conclusion . . . 87

7 Identifying disease-gene associations with graph-regularized manifold learning 98

7.1 Introduction . . . 99

7.2 Methods and material . . . 100

7.2.1 General model . . . 100

7.2.2 Similarity network . . . 103

7.2.3 Prior information . . . 104

7.2.4 Data sources . . . 104
7.2.5 Evaluation metrics . . . 105
7.3 Results . . . 106
7.3.1 Model parameters . . . 106
7.3.2 Cross-validation . . . 106
7.3.3 *De novo* study . . . 106
7.4 Conclusion . . . 107
7.5 Acknowledgements . . . 107

8 Summary and future work **110**
8.1 Summary . . . 110
8.2 Future work . . . 111

References **112**

Appendix A List of Publications **144**

Appendix B Copyright Permissions **146**

List of Tables

2.1 Some commonly used databases of disease-gene associations 17
2.2 Some commonly used databases of PPI networks 18
2.3 Some commonly used databases of Pathways 20

3.1 Enriched KEGG pathways of candidate genes in the BC dataset 43
3.2 Enriched KEGG pathways of candidate genes in the TC dataset 44
3.3 Enriched KEGG pathways of candidate genes in the AD dataset 46
3.4 Top 10 unknown genes 47

4.1 Sensitivity analysis. The resulted AUC values obtained with different combinations of hyperparameters for BC. 58
4.2 Sensitivity analysis. The resulted AUC values obtained with different combinations of hyperparameters for TC. 59
4.3 Sensitivity analysis. The resulted AUC values obtained with different combinations of hyperparameters for AD. 60
4.4 Top 10 unknown genes 63

5.1 Top-10 associations predicted by dgMDL, Known-GENE and PCFM 77
5.2 Top-10 susceptible lung cancer-associated genes 78

6.1 12 features extracted from mutation data. 87
6.2 Top 10 predictions of deepDriver 88
6.3 Top 10 predictions of 20/20+ 89
6.4 Top 10 predictions of SVM 90
6.5 Top 10 predictions of OncodriveCLUST 91

7.1 Top 10 predictions for lung cancer and bladder cancer 109

List of Figures

2.1 Classification of existing computational methods for disease gene prediction. 6

2.2 Schematic example of a random walk. The network contains 50 nodes (genes), in which 15 of them are disease-associated. Their corresponding entries in P_0 are equal to 1. The random walk is performed with a restart probability of $r = 0.5$, and it reaches a stead state when $t = 17$. . . . 8

2.3 Classic pipeline of supervised machine learning-based methods. 12

2.4 Eight types of evidence valuable for disease gene prediction. The five types of evidence in the left blue circle characterize the functional similarity of genes, and the two types of evidence in the right yellow circle contain disease-associated information. Gene expression in the middle contain both types of information. 16

3.1 The work flow of dgSeq. (a)–(b). Clinical RNA-Seq data of the case and control subjects; (c). Differential network constructed by (a) and (b); (d) disease-gene network constructed by OMIM data; (e). Labels of all the genes determined by (c), (d) and benchmark disease genes; (f). PPI network; (g). Feature matrix extracted from (c) and (f); (h). Logistic regression model trained with (e) and (g); (i). The calculated probabilities of all the genes being labeled as 1 (disease gene). 31

3.2 Scatter plot of the log-log degree distribution of G_{dif} for BC 34

3.3 Sample disease gene network 35

3.4 Sensitivity analysis of threshold k 39

3.5 The ROC curves of three algorithms in predicting BC-related genes 40

3.6 The ROC curves of three algorithms in predicting TC-related genes 41

3.7 The ROC curves of three algorithms in predicting AD-related genes 42

4.1 Work flow of the algorithm. 50

4.2 Hierarchical clustering dendrogram for BC. 56

4.3 Hierarchical clustering dendrogram for TC. 56

4.4 Hierarchical clustering dendrogram for AD. 57

4.5 ROC curves for BC. 61

4.6 ROC curves for TC. 61

4.7 ROC curves for AD. 62

5.1 Schematic example of an RBM. 67

5.2 Schematic example of a multimodal DBN for disease gene prediction. 75

5.3 AUC of dgMDL in different layers 76

5.4 ROC curves of the three algorithms 76

6.1 Schematic 1-D CNN. In this study, each CONV layer is followed by a pooling layer and the CONV-POOL pattern is repeated for several times. The final structure of the model is determined by grid search. 92
6.2 Construction of ϕ_i. Given the feature vectors of g_i and its k nearest neighbors $g_{s1}, g_{s2}, \ldots, g_{sk}$, a feature matrix ϕ_i is constructed by arranging the $2k$ vectors into a $2k \times n_f$ matrix, which is then used in the convolution. 93
6.3 ROC curves of the three algorithms obtained on the dataset of BRCA. The red, green and magenta lines depict the ROC curves of deepDriver, 20/20+ and SVM, respectively. The AUC value of deepDriver is 0.984, which is at least 15.1% higher than that of the other two algorithms. 93
6.4 ROC curves of the three algorithms obtained on the dataset of COAD. The red, green and magenta lines depict the ROC curves of deepDriver, 20/20+ and SVM, respectively. The AUC value of deepDriver is 0.976, which is at least 25.5% higher than that of the other two algorithms. 94
6.5 ROC curves of the three algorithms obtained on the dataset of LUAD. The red, green and magenta lines depict the ROC curves of deepDriver, 20/20+ and SVM, respectively. The AUC value of deepDriver is 0.998, which is at least 24.9% higher than that of the other two algorithms. 95
6.6 Learning curve for BRCA. 95
6.7 Learning curve for COAD. 96
6.8 Learning curve for LUAD. 96
6.9 ROC curves of deepDriver obtained from the second sets of driver genes. 97

7.1 ROC curves of the three competing algorithms on multiple-gene diseases. 108
7.2 ROC curves of the three competing algorithms on single-gene diseases. 108

List of Abbreviations

AD	Alzheimer's Disease
AI	Artificial Intelligence
AUC	Area Under the ROC Curve
AUPR	Area Under the Precision-Recall Curve
BBRBM	Binary-Binary RBM
BC	Breast Cancer
BRCA	Breast Invasive Carcinoma
CDF	Cumulative Distribution Function
CGC	Cancer Gene Census category
CNN	Convolutional Neural Network
COAD	Colon Adenocarcinoma
DAG	Directed Acyclic Graphs
DBN	Deep Belief Net
DGN	Disease Gene Network
eQTL	expression Quantitative Trait Loci
FI	Functional Interaction
FN	False Negative
FP	False Positive
FPKM	Fragments Per Kilobase Million
FPR	False Positive Rate
GBRBM	Gaussian-Binary RBM
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GIN	Gene Interaction Network
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Associations Studies
HPO	Human Phenotype Ontology
KNN	k Nearest Neighbors
lncRNA	long non-coding RNA
LOOCV	Leave-One-Out Cross Validation
LUAD	Lung Adenocarcinoma
MI	Mutual Information

NMF	Nonnegative Matrix Factorization
OMIM	Online Mendelian Inheritance in Man
PCC	Pearson Correlation Coefficient
PN	Potential Negative
PPI	Protein-Protein Interaction
PWEA	Pathway Enrichment Analysis
RBM	Restricted Boltzmann Machine
RN	Reliable Negative
RNAi	RNA interference
ROC	Receiver Operating Characteristic
RPKM	Reads Per Kilobase Million
RWR	Random Walk with Restart
RWRH	RWR on Heterogeneous
SGD	Stochastic Gradient Descent
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
TC	Thyroid Cancer
TN	True Negative
TP	True Positive
TPM	Transcripts Per Kilobase Million
TPR	True Positive Rate
TS	Target Sequencing
TSG	Tumor Suppressor Gene
UQ-FPKM	Upper Quartile normalized FPKM
WXS	Whole Exome Sequencing

1

Introduction

1.1 Background

Complex diseases are caused by the malfunctioning of a group of genes, known as disease-associated genes, or simply disease genes. Identifying these genes is critical for scientists to decipher the mechanism of diseases, which is beneficial to disease diagnosis and drug development [1]. However, this issue is still challenging since experimentally identifying disease genes is time-consuming and expensive. On the one hand, scientists need to conduct a few experiments to determine whether a gene is disease-associated, which might require years of efforts [2]. On the other hand, experimental techniques such as genome-wide association studies (GWAS) usually identify hundreds of candidates, and scientists have to determine the priority of validations to maximize the yield of their experiments. Therefore, computational methods which prioritize disease genes are valuable for disease-gene identification.

Currently, many algorithms have been proposed to predict disease genes. Despite their success, different methods all have their pros and cons. Artificial Intelligence (AI) tries to harness the power of techniques like machine learning and deep learning to solve problems by analyzing and learning from vast datasets at speeds and capacities not possible for humans alone. Its flexibility in data integration is also valuable for disease-gene prediction since the key to accurate prediction is to properly fuse multi-levels of biological data. Thus, this thesis mainly focuses on machine learning-based methods, especially deep learning models, which can characterize the non-linear relationships among different types of data.

In our studies, we first develop algorithms to solve common issues existing in developing machine learning-based methods, such as the selection of negative data and the quality control of protein-protein interaction (PPI) networks. Then, several deep learning models are applied to fuse multi-level of biological data and extract cross-modality features. These features characterize both linear and non-linear relationships among different modalities, which could advance the prediction of disease-gene associations.

1.2 Motivations and objectives

The overall objective of our studies is to combine multiple types of data with deep learning models to improve the accuracy of disease-gene prediction. Before applying deep learning models to this area, a few

issues should be addressed.

First, considering that we use supervised models to predict disease-gene associations, both positive and negative instances are required to train the models. However, disease-gene prediction is a positive-unlabeled learning problem, in which only positive instances (disease genes) are available [3]. A set of negative instances (non-disease genes) have to be defined before training the models. Thus, developing a strategy to select negative data is fundamental to our research.

Second, PPI networks are widely used in existing algorithms since they reveal the functional similarities of proteins, which are critical for disease-gene prediction. However, PPI networks obtained from online databases contain many false positives and false negatives [4], and directly using them in the algorithm would limit the accuracy of the prediction. Meanwhile, protein interactions are tissue-specific and dynamic, universal static networks downloaded from public databases cannot reveal true protein interactions in the samples. Therefore, a strategy should be developed to purify the static PPI networks and improve their quality.

After solving these two issues, the next step is to properly fuse multiple types of data to achieve accurate prediction. Specifically, deep learning models which use nonlinear activation functions are applied in our studies to capture the nonlinear relationships among various types of data. Multimodal deep belief nets (DBNs), a fundamental deep learning architecture which has been successfully applied to fuse image and text data [5], is first applied to learn latent representations from different types of biological data. In addition, convolutional neural networks (CNNs), which use the same filter for similar inputs, could also be applied for representation learning since they allow to leverage the functional similarities of genes.

Finally, since most supervised algorithms require known disease genes as positive data to train the model, they cannot be applied to diseases with only a few associated genes or no known disease genes. However, the nonnegative matrix factorization (NMF)-based methods are not limited by this problem, and a few NMF-based models have successfully been used to predict disease genes [6, 7]. Unfortunately, existing methods still perform badly for diseases with no known associated gene. A better method should be proposed to improve their accuracy. In the meantime, it is interesting to compare NMF-based methods with deep learning-based methods to find out their specialties.

Based on these motivations, this study has the following objectives:

Objective 1: Review existing computational algorithms for predicting disease-gene associations.

Objective 2: Develop a new strategy to select non-disease genes and combine it with network energy-based model to predict disease genes.

Objective 3: Develop a method to improve the quality of PPI networks and apply it to predict disease genes.

Objective 4: Use a multimodal DBN to integrate different types of data and extract cross-modality features to predict disease genes.

Objective 5: Apply the CNN model to integrate different types of data based on gene similarities.

Objective 6: Present an NMF-based method which can accurately predict disease genes for diseases both with many known associated genes and no known associated genes. Then compare it with the deep learning-based methods.

1.3 Organization of the thesis

This is a manuscript-style thesis. The main content is presented in the form of published or submitted manuscripts that I have written during my Ph.D. study. An introduction is given at the beginning of each chapter to describe the connection of the manuscript to the context of the thesis. All manuscripts have been reformatted to maintain consistency. The reference lists of all publications have been unified, and there is only one bibliography at the end of the thesis.

The remainder of the thesis is organized as follows. Chapter 2 reviews the existing computational methods, databases, and evaluation methods used in disease-gene prediction. Chapter 3 proposes a method to select non-disease genes from OMIM data and clinical expression data and combines it with network energy-based model to predict disease genes. Chapter 4 proposes a strategy to improve the quality of PPI networks by constructing sample-based networks and use them with an ensemble strategy to predict disease genes. Chapter 5 presents a deep learning-based algorithm which applies multimodal DBN to predict disease genes. Chapter 6 proposes another deep learning-based method which uses CNN model to leverage functional similarity data. Chapter 7 presents an NMF-based method which uses manifold learning to predict disease genes. Prior information is added to the association matrix to improve the accuracy of the algorithm in predicting associated genes for diseases with no known disease genes. Finally, Chapter 8 summarizes the work presented in this thesis and discusses several future directions for this research. The copyright permissions of the manuscripts are included in Appendix B.

2

Predicting disease-associated genes: computational methods, databases, and evaluations

Prepared as: Ping Luo, Bo-Lin Chen, Bo Liao, and Fang-Xiang Wu. Predicting disease-associated genes: computational methods, databases, and evaluations. WIREs: Data Mining and Knowledge Discovery, under revision, 2019. PL reviewed the existing literature, and FXW supervised the study. PL and FXW wrote the manuscript. All authors read, revised and approved the final version of the manuscript.

This chapter presents a literature review about computational algorithms, databases and evaluation methods used in predicting disease genes. The review classifies existing algorithms into three categories: network-based methods, machine learning-based methods and other methods. The pros and cons of different types of methods are discussed, as well as several perspectives to improve them. Commonly used databases and evaluation methods are also discussed so that researchers can easily develop their own algorithms. Objective 1 of the thesis is fulfilled in this chapter.

Abstract

Complex diseases are associated with a set of genes (called disease genes), the identification of which can help scientists uncover the mechanisms of diseases and develop new drugs and treatment strategies. Due to the huge cost and time of experimental identification techniques, many computational algorithms have been proposed to predict disease genes. Although several review publications in recent years have discussed many computational methods, some of them focus on cancer driver genes while others focus on biomolecular networks, which only cover a specific aspect of existing methods. In this review, we summarize existing methods and classify them into three categories based on their characteristics. Then, the state-of-the-art algorithms, biological data and evaluation methods used in the computational prediction are discussed. Finally, we highlight the limitations of existing methods and point out some future directions for improving these algorithms. This review could help investigators understand the principles of existing methods, and thus develop new methods to advance the computational prediction of disease genes.

2.1 Introduction

Deciphering the associations between diseases and genes is critical for us to understand the modular nature of complex diseases, which has many applications, such as diagnosis, treatment, and prevention of diseases. Usually, genes whose malfunctioning causes diseases are known as disease-associated genes, or simply disease genes. A few experimental techniques can be used to identify these genes, such as linkage analysis [8], genome-wide associations studies (GWAS) [9], and RNA interference (RNAi) [10]. Among these techniques, linkage analysis and GWAS are most frequently used. The former is successful in identifying genes associated with Mendelian diseases (single gene diseases), while the latter is superior to the former in predicting genes associated with complex diseases (non-Mendelian diseases) [11]. Despite their achievements, these techniques usually select genetic loci corresponding to hundreds of candidate genes, whose further validation is time-consuming and expensive. Thus, many computational algorithms have been proposed to predict or prioritize disease genes so that scientists can optimize the in-depth experimental validation and maximize the yield of their experiments.

An intuitive strategy for computational methods is to analyze the results of GWAS and predict disease genes from the previously mentioned hundreds of candidates. This results in a group of post-GWAS analysis algorithms. However, not all disease-associated variants can be identified by GWAS [12], and GWAS data are not always available for all kinds of diseases. Thus, computational methods have also used many other types of data, such as protein-protein interaction (PPI) networks, gene expression profiles, pathways, gene ontology (GO) terms, to predict disease genes. The authors of [2] have classified various existing types of data into five categories. Among them, the mutation data are most promising. Newly developed algorithms for cancer driver gene prediction also tend to focus on analyzing somatic mutations rather than other types of data. Unfortunately, large scale mutation data are usually unavailable for diseases other than cancer. Therefore, in this review, we focus on computational algorithms and data sources that can be used to predict associated genes for all kinds of diseases. Algorithms specifically designed for predicting cancer driver genes can be found in [13].

Based on the principles used in the classification, existing methods can be divided into various categories. In [11], the authors have classified computational methods based on three criteria: "type of evidence", "scope of application" and "type of prediction". With these criteria scientists can quickly find the methods they need based on their objectives and data at hand. However, methods in each category might vary a lot in terms of their core models, which is inconvenient for researchers who want to improve current models and develop new algorithms. In this review, instead of using these criteria, we focus on the characteristics of computational algorithms and classify them into network-based methods, machine learning-based methods, and other methods. Figure 2.1 shows the details of the classification. We believe that such a classification can help researchers capture the core ideas behind existing algorithms, which might assist them in developing new effective algorithms.

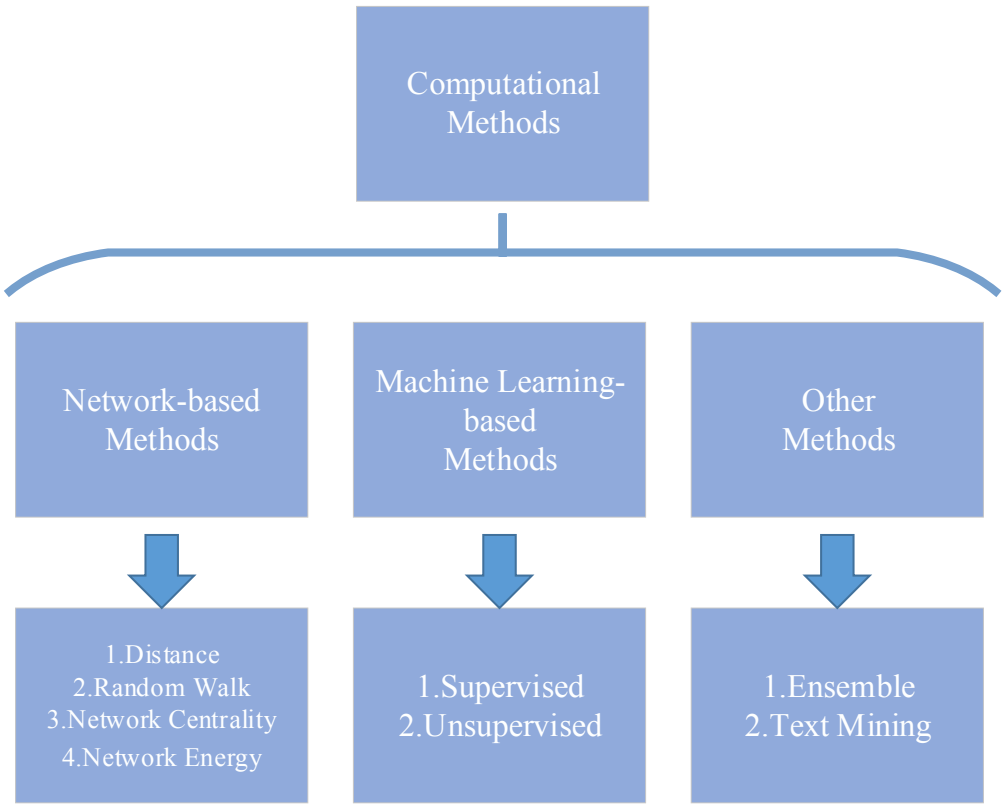

Figure 2.1: Classification of existing computational methods for disease gene prediction.

In the rest of this review, we first discuss existing computational disease gene prediction methods. Then, a few types of frequently used data are described in Section 2.3, as well as the strategies developed for analyzing them. After that, Section 2.4 introduces some evaluation methods. Finally, we conclude with a discussion of the limitations of existing methods and perspectives for developing new algorithms in the future.

2.2 Computational methods

2.2.1 Network-based methods

Based on the 'guilt by association' assumption, genes associated with each other may have similar functions [14]. Therefore, various types of biomolecular networks which characterize associations among genes have been used to predict disease genes. Very briefly, network-based algorithms can be divided into four groups: distance-based methods, random walk-based methods, network energy-based methods, and network centrality-based methods.

Distance-based methods

Distance-based methods were the first to be developed to predict disease genes. These methods use the length of the shortest path (distance) in biomolecular networks to determine if a gene is disease-associated. Unknown genes (genes that have not been identified as being associated with a certain disease) with a distance smaller than a threshold are predicted as disease-associated. For instance, George et al. developed a method known as CPS which predicted a gene as disease-associated if it was in a disease-related pathway and close to the known disease genes [15]. Snel et al. combined PPI network with disease-associated loci to predict a gene as disease-associated if it was located within these loci, and its neighbors contained known disease genes [16]. Franke et al. used the distance to calculate similarity scores with the Gaussian kernel, and unknown genes with higher similarity scores were predicted as disease-associated [17].

These methods only use local topological structures, and their accuracy is limited. Random walk-based methods have shown that the global topological structure is more valuable for disease gene prediction than local information [18]. Nevertheless, the distance-related evidence is still valuable and has been used to extract features with many machine learning-based methods [19].

Random walk-based methods

To improve the distance-based methods, the random walk is applied to predict disease genes. Even since the first random walk with restart (RWR) algorithm was proposed in [18], it has become one of the state-of-the-art algorithms for disease gene prediction. As a kind of information flow-based algorithm, random walk-based methods propagate the prior information from each node to its nearby nodes in an iterative manner for a predefined number of steps or until convergence. The final value of a node is influenced by the values of its direct neighbors, which in turn are affected by their neighbors. This value also represents the probability of each node being associated with the disease under consideration. Due to its superiority, random walk-based methods have been applied in many other areas as well, including gene function prediction [20] and drug target prediction [21].

Given a disease d, let P_0 denote the prior information where $P_0(i) = 1$ represents gene i is known to be associated with d and $P_0(i) = 0$ otherwise. The random walk can be performed by the following equation

$$P_{t+1} = WP_t = W^t P_0, \ (t \geq 0) \tag{2.1}$$

where W is the column normalized adjacency matrix of the PPI network. If W is a stochastic matrix, this process is equivalent to a random walk on the network. Furthermore, if we allow the random walk to restart in every step with a probability r, we can obtain the RWR algorithm as follows:

$$P_{t+1} = (1 - r)WP_t + rP_0 \tag{2.2}$$

where P_t in the steady state would contain the probability of each gene being disease-associated.

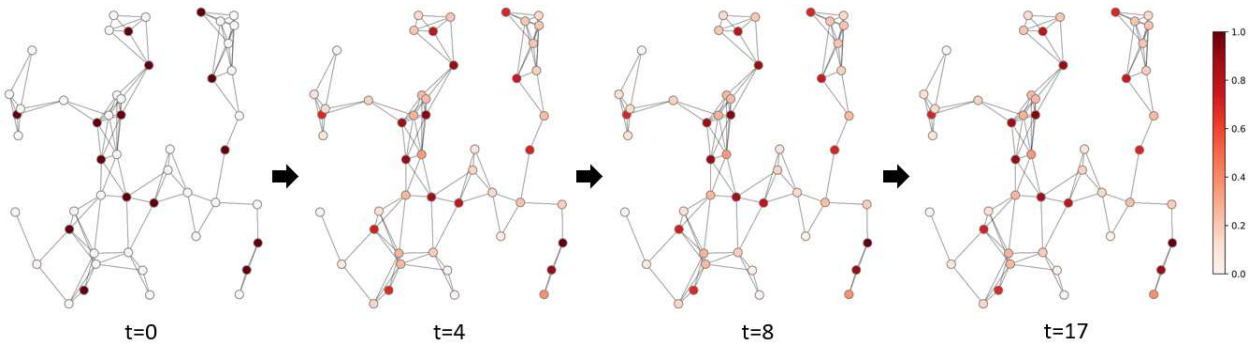

Figure 2.2: Schematic example of a random walk. The network contains 50 nodes (genes), in which 15 of them are disease-associated. Their corresponding entries in P_0 are equal to 1. The random walk is performed with a restart probability of $r = 0.5$, and it reaches a stead state when $t = 17$.

Figure 2.2 illustrates the process of a random walk which is performed on a randomly generated network with 50 nodes, 15 of which are disease-associated. With a restart probability $r = 0.5$, the random walk reaches the steady state ($P_{t+1} - P_t < 10^{-6}$) after 17 rounds of iterations. From the color of the nodes we can find that the prior information is propagated to the other 35 nodes during the iteration process.

Köhler et al. first used RWR to predict disease genes in [18]. They also proposed a diffusion kernel method which was the continuous-time analog of RWR. The diffusion kernel of a network was defined by $K = e^{-\beta L}$, where $L = D - A$ was the Laplacian matrix, D was a diagonal matrix containing the degrees of the nodes, A was the adjacency matrix of the network. P was then computed by $P = KP_0$.

RWR and diffusion kernel captured the global topological properties of the network and were superior to the distance-based algorithms. However, Köhler et al. only performed RWR on top of a PPI network, and P_0 only contained information of known disease-gene associations. To improve the accuracy of the prediction, many researchers started to combine RWR with other types of data.

One strategy is to enhance P_0 with information obtained from other types of data. For instance, PRINCE used a logistic function to calculate P_0 based on the similarity of diseases [22]. If a gene was associated with a disease which was similar to the disease under consideration, its prior probability would be close to 1. Another strategy is to enrich the PPI network with extra information. For instance, Erten et al. and Le et al. weighted the PPI network with reliabilities calculated from other data sources [23, 24]. Magger et al. built a tissue-specific network and performed PRINCE on it [25].

These methods improved the accuracy of disease-gene prediction. However, they only used PPI networks. To further leverage the potential of random walks, researchers begun to use heterogeneous networks to improve the prediction accuracy.

Heterogeneous networks are networks with multiple types of nodes and edges. The first RWR on heterogeneous (RWRH) network algorithm was proposed by Li et al. and the network was constructed by combining PPI network, disease gene associations and disease similarity network [26]. RWRH performed much better than the RWR algorithm, and many studies were then conducted to improve it. For instance,

Luo et al. constructed the heterogeneous network based on a curated PPI network and improved the prediction accuracy [27]. Jiang constructed three disease similarity networks and nine gene similarity networks and combined them into 27 heterogeneous networks. Then, RWR was applied on these networks respectively, and a weighted Fisher's method was used to combine all the propagated values to prioritize disease genes [28]. Valdeolivas et al. constructed a multiplex heterogeneous network in which the same nodes in different heterogeneous networks are connected with each other. This strategy allowed the random walk to transit between different networks, which significantly improved the accuracy of the prediction [29].

Note that heterogeneous networks allow different types of nodes, researchers can also integrate other omics data to construct the network. For instance, Lei et al. constructed a triple heterogeneous network by incorporating long non-coding RNA (lncRNA) into the network. Specifically, lncRNA-lncRNA similarity network, gene-lncRNA associations and lncRNA-disease associations were fused to the traditional heterogeneous networks [30].

Network centrality-based methods

Centrality characterizes the importance of nodes and edges in a network, which has been widely used in social network analysis [31, 32] and essential protein prediction [33, 34, 35]. However, genes that transcribed to essential proteins are usually not disease-associated [36], and directly applied network centrality to predict disease genes is difficult.

The most successful centralities used in disease gene prediction are feedback centralities, such as Katz centrality and PageRank. These centralities of one node depends on the centralities of its neighbors, which further depends on the centralities of their neighbors. The process of feedback is similar to RWR. In fact, RWR is also known as personalized PageRank [37]. Therefore, methods that use feedback centrality also use similar strategies as random walk-based methods to predict disease genes. For instance, Singh-Blom et al. built a heterogeneous network and used Katz centrality to characterize the possibilities of every disease-gene pairs being associated [38]. Ganegoda et al. also used Katz centrality in their study [39], except that they replaced the PPI network by a tissue-specific network.

Centrality-based methods improve the disease-gene prediction to some extent. However, compared to directly being used to predict disease genes, Centralities are more likely to be used to extract topological features in machine learning-based methods [38, 40, 41, 42, 43].

Network energy-based methods

The network energy-based method was first proposed in [44]. The authors formulated the disease gene prediction problem as a network labeling problem in which disease genes were labeled as 1 while non-disease genes (genes not associated with the disease) were labeled as 0.

Given h genes in a biomolecular network, a set of binary labels $\mathbf{x} = (x_1, x_2, \ldots, x_h)$ of these genes, $(x_i \in \{0, 1\})$, is known as a configuration of the biomolecular network, and the set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{2^h}\}$ of

all possible configurations is a random field. The probability of the configuration $\mathbf{x}$ of a random field X can be calculated by the Boltzmann distribution [45]

$$P(\mathbf{x}) = \frac{1}{Y} \cdot \exp(-\kappa H(\mathbf{x})) \tag{2.3}$$

where $H(\mathbf{x})$ is the Hamiltonian (energy) of the configuration $\mathbf{x}$, κ is a parameter, and Y is called the partition function and defined as $Y = \sum_{\mathbf{x} \in X} \exp(-\kappa H(\mathbf{x}))$. From this equation we can find that the true configuration of the network should have the maximum probability and thus the minimum energy.

Let $x_{[-i]}$ be the set of binary labels of all genes except for gene i in the network. Adopting the Ising model to calculate the Hamiltonian, the following equation can be obtained

$$P(x_i = 1 | x_{[-i]}, \tilde{\theta}) = \frac{\mathbf{e}^{\alpha + \beta N_{i0} + \gamma N_{i1}}}{\mathbf{e}^{\alpha + \beta N_{i0} + \gamma N_{i1}} + 1} \tag{2.4}$$

where $\tilde{\theta} = (\alpha, \beta, \gamma)$ are model parameters. N_{i0} and N_{i1} are the numbers of neighbors with labels 0 and 1 of gene i, respectively. Details of the derivation can be found in [46]. Parameter $\tilde{\theta}$ can be estimated by maximizing the posterior distribution of $P(x_1, \ldots, x_n | x_{n+1}, \ldots, x_{n+m})$ where $x_{n+1}, \ldots, x_{n+m}$ are the labels of the m known disease genes. Furthermore, if M networks are available, (2.4) can be generalized as follows

$$P(x_i = 1 | x_{[-i]}, \theta) = \frac{exp(V(i))}{exp(V(i)) + 1} \tag{2.5}$$

where

$$V(i) = \alpha + \sum_{m=1}^{M} [\beta^m \cdot N_{i0}^m + \gamma^m \cdot N_{i1}^m],$$

$\mu = (\alpha, \beta^m, \gamma^m)$ $(m = 1, \ldots, M)$ are model parameters. N_{i0}^m and N_{i1}^m are the numbers of neighbors with labels 0 and 1 of gene i in the m-th network, respectively.

Network energy-based model can be used to integrate multiple biomolecular networks to predict disease genes. In [46], the authors integrated five biomolecular networks and estimated θ with Gibbs sampling. Later on, they combined graph kernel with their model and proposed a kernel-based algorithm. This strategy allowed the algorithm to use the information between genes and their indirect neighbors since kernel brought similarity information into the network [47]. Having noticed that Eq. (2.4) followed a logistic model where $\phi_i = (1, N_{i0}, N_{i1})$ was the feature vector and $\tilde{\theta}$ was the weight parameter. Parameter $\tilde{\theta}$ could be estimated by a convex optimization problem which was much faster than the original Gibbs-based methods. N_{i0} and N_{i1} could also be extended to extract more similar features. Thus, Chen et al. proposed a fast algorithm based on this new strategy and extracted additional features N_{i0}^l and N_{i1}^l which were the number of the l-order neighbors of gene i [48]. This algorithm improved their original running time by more than 20 folds and the performance for more than 10% in terms of the area under the receiver operating characteristic curve.

2.2.2 Machine learning-based methods

Machine learning-based methods formulate disease gene prediction as a binary classification problem where disease genes are predicted as 1 and non-disease genes are predicted as 0. Based on whether known disease

genes are used in the prediction, machine learning-based methods can be divided into unsupervised methods and supervised methods.

Unsupervised methods

Unsupervised methods identify patterns in dataset without knowing the labels of instances. For disease gene prediction, most unsupervised methods use clustering algorithms to predict disease-associated modules from biomolecular networks. These modules are subnetworks consisting of genes associated with the diseases. The first widely used algorithm, Weighted Gene Correlation Network Analysis (WGCNA), used hierarchical clustering on top of co-expression networks to search disease-associated modules [49]. WGCNA provided an easy-to-use R package, and many studies have used it to identify disease-associated genes [50, 51]. In addition to co-expression networks, researchers also combined co-expression and PPI networks to find disease modules. For instance, Wu et al. weighted a protein functional interaction network with Pearson correlation coefficient (PCC) and applied Markov clustering on this network to predict disease genes [52]. This method allowed the clustering algorithm to leverage the topological properties within the PPI networks.

Although successful, these algorithms mainly used co-expression data to find disease-associated modules, and only a small amount of genes ($< 20\%$) in the detected modules are disease-associated [53]. To improve the efficiency, differential co-expression (also known as "guilt by rewiring") [54] was applied to predict disease modules.

Usually, genes that are differentially co-expressed in different groups of samples are more likely to be disease-associated [55, 56]. Methods based on this assumption construct differential networks in which edges are weighted by the variation of co-expression and apply clustering algorithms on these networks to predict disease genes. For instance, DiffCoEx built an adjacency change matrix based on the co-expression networks of two conditions (case and control). Then, the topological overlap was used to construct a dissimilarity network, and hierarchical clustering was used to find out modules that were differentially co-expressed with the same sets of genes [57]. Another example is EW_dmGWAS, which weighted the edges of a PPI network with differential co-expression and nodes of the network with P-values obtained from GWAS data. A seed-growth approach was then used to find subnetworks with the locally maximum proportion of low-P-values and highly rewired edges [58].

Unsupervised methods can be applied to predict disease genes even if a disease has no known associated genes. Since known disease-gene associations are not used in the prediction, their accuracy is usually worse than supervised methods. However, unsupervised methods are still widely used in analyzing biological data, especially gene expression data. One reason is that these algorithms are user-friendly, and do not require complicated data preprocessing. Another reason is that large numbers of gene expression data are available for applying unsupervised methods. Biclustering algorithms and differential co-expression analysis have successfully identified many tissue-specific and cell-type-specific disease-related modules.

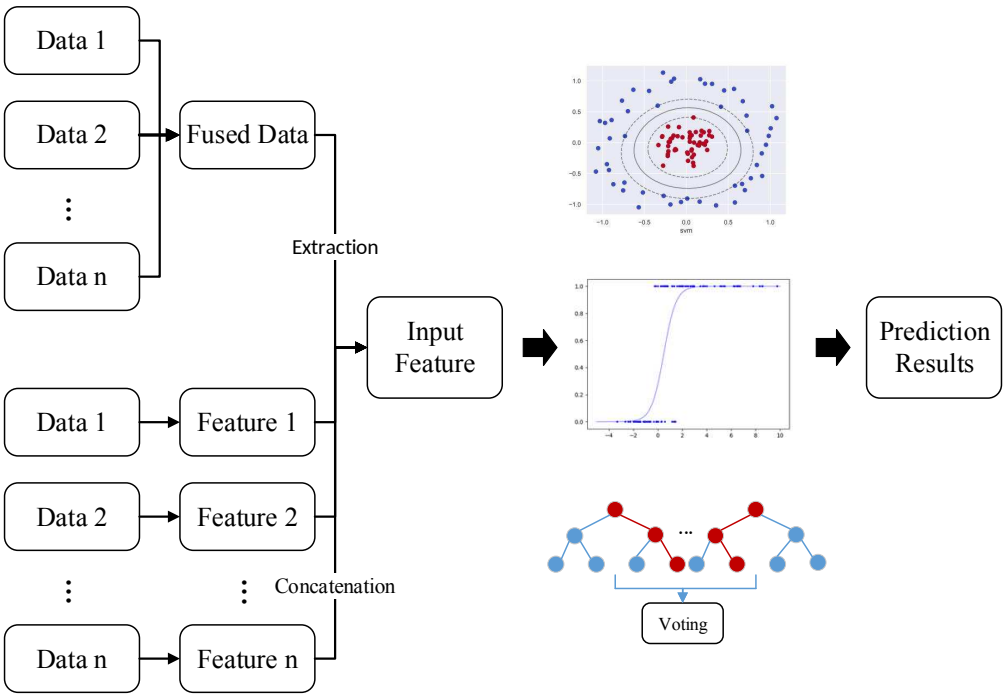

Figure 2.3: Classic pipeline of supervised machine learning-based methods.

Supervised methods

Supervised methods train a classifier/regressor based on the known disease-gene associations. Figure 2.3 shows the pipeline of most supervised methods. First, the features are either extracted from a fused dataset (e.g. PPI network weighted by gene expression or GWAS data) or concatenated from feature subsets extracted from each type of data. Then, a classifier/regressor, such as biased SVM, logistic regressor and random forest, is trained with these features and used for future prediction. Based on the key ideas of the methods, supervised methods can be further divided to feature-based, deep learning-based and matrix factorization-based methods.

A. Feature-based

Feature-based methods focus on extracting features which characterize the functional similarities of genes. Appropriate data integration is necessary for obtaining discriminative features.

As depicted in Figure 2.3, the features can be extracted from fused data or concatenated from individual features extracted from each type of data. No matter which strategy is chosen, most feature-based methods linearly combine different types of data. For instance, Wu et al. proposed CIPHER, which extracted shortest-path based features from a heterogeneous network [19]. ProDiGe used kernels to calculate gene similarities from each type of data and combined them together by the weighted average of these similarity profiles [3]. PUDI directly concatenated features extracted based on GO, protein domain and a PPI network [41].

Note that PPI networks were used to extract features in these algorithms. However, neither length of the

shortest path or degree of the node can capture the topological properties of the entire network. To improve these strategies, researchers started to use multiple types of centralities to extract both local and global topological properties. For instance, Ramadan et al. extracted 13 topological features, most of which were centrality indices, and trained a decision tree to predict disease genes [42]. Luo et al. used closeness centrality and edge clustering coefficient to capture both global and local topological structures from a sample-specific network [43] [1].

Methods using centrality-based features perform better than earlier developed algorithms; however, centrality indices are not the best approach for learning network representations. Considering that many algorithms have been developed to learn network embeddings, researchers started to use these algorithms to extract features for disease gene prediction. For instance, Ata et al. proposed an algorithm (Metagraph+) which constructed a metagraph by combining PPI networks and keywords that describe the mechanisms of proteins [59]. In the metagraph, each protein was connected with other proteins and the keywords that describe itself. A matagraph embedding learning algorithm, SymISO [60], was used to extract representations for each protein (gene). Moreover, researchers can also concatenate network embeddings with other types of features to improve prediction accuracy. In [61], Ata et al. proposed N2VKO, which directly extracted features from PPI networks by node2vec [62], and concatenated these embeddings with features extracted based on UniProt annotations. They tested the features with several classifiers and demonstrated that N2VKO performed better than many classic algorithms, such as RWRH, ProDiGe and Metagraph+.

Note that most algorithms extract features from PPI networks using all the known disease-gene associations. However, Know-GENE proposed by Zhou et al. showed that features extracted with a subset of disease genes were more discriminative than those extracted using all the disease genes [63]. These subset of genes, defined as "Core genes", were those residing in the largest interacting cluster formed by all the disease genes. Researchers could use this strategy when developing new algorithms.

B. Deep learning-based

Similar to feature-based methods, deep learning-based methods also focus on feature extraction. However, the non-linear activation functions in deep learning models enable an algorithm to learn the non-linear relationships between different types of data, which is different from traditional feature-based methods. For instance, Luo et al. proposed a method which used multimodal deep belief net (DBN) to combine features extracted from different modalities [64] [2]. Specifically, raw features learned based on gene ontology and PPI networks were fused by a multimodal DBN to learn cross-modality representations, which were further used to predict disease-gene associations. Their evaluation results showed that prediction using cross-modality features were more accurate than original raw features.

In addition to DBN, convolutional neural network (CNN) was also used to predict disease genes. Different

[1] [43] is Chapter 4 from this thesis
[2] [64] is Chapter 5 from this thesis

from image data, gene-related features do not contain spatial information. However, a model known as graph CNN can solve this problem. In this model, the convolution is performed by aggregating information from the neighbors of each node in a graph [65]. This graph can model any biomolecular networks that reveal the functional similarity of genes. In [66], the authors used graph CNN to learn representations from a heterogeneous network. Specifically, raw features extracted for diseases and genes were integrated into the heterogeneous and further learned by graph CNN. Their evaluation results showed that the method performed much better than most classic algorithms, such as inductive matrix completion (IMC) [6] and Katz [38].

C. Matrix factorization-based

Apart from feature-based methods, nonnegative matrix factorization (NMF) has also been used to predict disease genes. Unlike most supervised methods, NMF-based methods can predict disease genes even if the disease under consideration has no known associated genes. Given a disease-gene association matrix A, in which $A(i,j) = 1$ if disease i and gene j are associated and $A(i,j) = 0$ otherwise. The general idea is to find a low-rank matrix $P = WH^T$, where $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{n\times k}$ are of rank $k \ll \min(m,n)$, so that $P \approx A$. This problem can be solved by NMF, and matrix P contains the probabilities of every disease-gene pairs being associated. During the factorization, additional information can be integrated into the objective function so that the factorization is in concert with other biological evidence. A typical algorithm was the IMC proposed by Natarajan and Dhillon. The algorithm leveraged gene-related and disease-related features during the factorization, making it possible to integrate multiple types of data [6]. Another example was the probability-based collaborative filtering model (PCFM) proposed by Zeng et al. which used alternating least squares to solve the factorization [7]. In PCFM, disease similarities and gene similarities were used to regularize the objective function, which was a common strategy used by many NMF-based studies.

In addition to classic NMF algorithms, manifold learning can also be used in disease gene prediction. Different from NMF where entries in P denote the probabilities of diseases and genes being associated, manifold learning mapped the diseases and genes onto a lower dimensional manifold and used the geodesic distance between diseases and genes to predict their associations [67, 68] [3]. Known disease-gene associations are used in the mapping based on the assumption that distance between a disease and its associated genes should be shorter than other non-disease genes. Thus, disease-gene pairs with smaller distance on the lower dimensional manifold are more likely to be associated.

2.2.3 Other methods

Other than the methods discussed in Sections 2.2.2 and 2.2.1, there are also many algorithms that cannot be classified into the previous two categories. One of them is ensemble-based methods which calculate a few ranked lists based on different types of data and outputs the final prediction by fusing all the ranked lists. Different models can be applied when the ranking is calculated on each type of data. A typical

[3] [68] is Chapter 7 from this thesis

example is Endeavor, one of the most famous ensemble-based methods, which can analyze 75 data sources and predict associated genes for six species. During its prediction, rankings obtained from each data source were combined by order statistics [69, 70]. Another example was DADA, which proposed five statistical adjustment methods to generate five raw rankings and used the best ranking of each gene as its final ranking [71].

Another type of method is text-based, which used text mining to analyze existing literature and predict disease genes. For different methods, the model might be completely different. For instance, DISEASES calculated a score for each disease-gene pair according to their co-occurrence within the same abstracts and the same sentences [72]. DigSee extracted ten linguistically motivated features and used a Bayesian classifier to identify disease-gene association evidence [73, 74]. Although text-based methods cannot provide *de novo* prediction (only documented genes can be predicted), their results are useful for disease gene databases, and many databases have used text-based methods to collect disease-gene associations [72, 75, 76].

2.3 Biological data

As discussed in Introduction, various types of data have been used to predict disease genes. Figure 2.4 shows several types of widely used evidence in existing algorithms. Some of them characterize the functional similarity of genes, while others directly provide disease-associated information.

2.3.1 Disease-gene associations

The known disease-gene associations are the most significant. These data can be used to predict new disease genes and evaluate the prediction accuracy. Table 2.1 lists six databases that collect disease-gene associations. The first five databases (OMIM, DisGeNET, GAD, CTD, and PsyGeNET) contain genes associated with all kinds of diseases while the last one (COSMIC) mainly collects cancer-associated genes. Among these databases, OMIM is the most frequently used one in the algorithms we reviewed. However, OMIM focuses on Mendelian disorders, and genes associated with complex diseases are not comprehensively collected. An alternative is DisGeNET, which collects disease-gene associations from seven data sources, including CTD and PsyGeNET. This might be a better choice for studies on complex diseases.

In addition, although many algorithms are not specifically designed for cancer, researchers prefer to use cancer to validate their methods. One of the reasons is that many wet-lab studies have been conducted on various types of cancer, which is convenient for *de novo* validation. Another reason is that different types of cancer are well studied, and the number of known cancer-related genes is larger than other diseases, which is beneficial for training the model, especially for supervised learning methods. For methods that use cancer to evaluate their performance, the Cancer Gene Census project in COSMIC contains the most comprehensive cancer-related genes. Thus, algorithms should use their association data for the prediction.

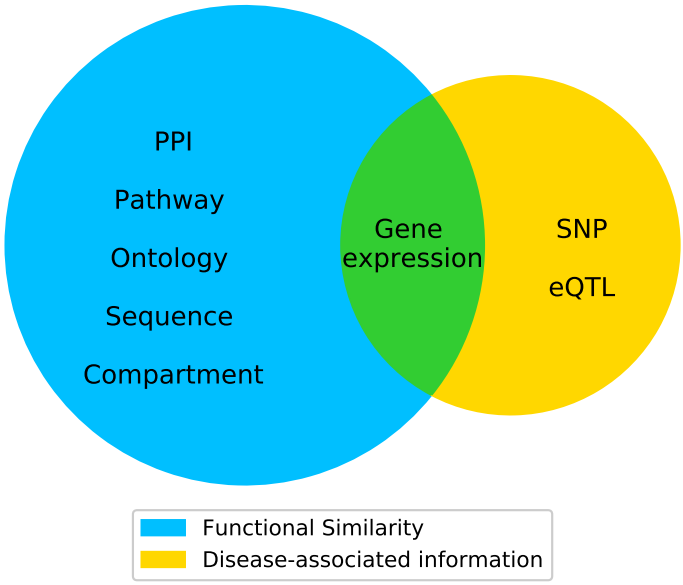

Figure 2.4: Eight types of evidence valuable for disease gene prediction. The five types of evidence in the left blue circle characterize the functional similarity of genes, and the two types of evidence in the right yellow circle contain disease-associated information. Gene expression in the middle contain both types of information.

Non-disease genes

Along with disease-gene associations, many algorithms require non-disease genes to train their models or evaluate their performance. Unfortunately, no databases contain non-disease genes, and computational methods have to select a set of unknown genes as non-disease genes. The simplest strategy is to randomly select a group of unknown genes as non-disease genes. The number of the selected genes varies depending on the model used for prediction. For weighted models (e.g. biased SVM or weighted random forest), the number of the non-disease genes is usually five to twenty fold higher than the known disease genes. For other models, the size of the positive and negative data should be similar to avoid imbalanced samples. Although the selected genes might be unidentified disease genes, the probability would be small since the number of all disease genes is much less than that of unknown genes. Furthermore, the bootstrap aggregating which improves the stability and accuracy of computational prediction [77] can be used by selecting several groups of non-disease genes and performing the prediction in an ensemble manner.

Additionally, instead of randomly selecting a group of unknown genes, for supervised machine learning-based methods, a better choice is to first define a set of highly potential non-disease genes [reliable negatives (RNs)], then select a subset of genes from RN as negatives. For instance, Luo et al. calculated the similarities of various diseases and used the associated genes of one disease as non-disease genes of the other one for

Table 2.1: Some commonly used databases of disease-gene associations

Name	URL	Ref.	Latest Update
OMIM	`https://www.omim.org`	[80]	01/01/2019
DisGeNET	`http://www.disgenet.org`	[81, 76]	01/14/2019
GAD	`https://geneticassociationdb.nih.gov`	[82]	08/18/2014
CTD	`http://ctdbase.org`	[83]	01/13/2019
PsyGeNET	`http://www.psygenet.org`	[84, 85]	09/02/2016
COSMIC	`https://cancer.sanger.ac.uk`	[86, 87]	11/13/2018

two dissimilar diseases [78] [4]. This strategy collected a set of RN which enabled the training of a successful model. However, note that the reasons of different genes not associated with a disease are different. Some of these genes might be passenger genes while others might have completely no function for disease genesis. A set of well defined non-disease genes might be linearly separable from the disease genes, resulting in high prediction accuracy. However, this accuracy does not guarantee the successful prediction of all unknown genes. To solve this issue, Yang et al. proposed a strategy in PUDI which classified unknown genes into four categories: likely positives, likely negatives, reliable negatives, and weak negatives [41]. This strategy allowed the model to be trained with different types of unknown genes, which would enhance its ability in predicting new disease genes.

For network-based and other methods, the artificial linkage interval was frequently used to select non-disease genes [3, 79]. This strategy selects 99 genes that surround a disease gene on the chromosome as non-disease genes. During the prediction, each known disease gene and its 99 closest genes are regarded as unknown. Since similar genes tend to cluster in chromosomal neighborhoods, if an algorithm is more accurate in identifying disease genes from its neighbor genes, its performance should be better than the competing algorithms.

2.3.2 PPI network

PPI networks are the most widely used data in all three types of algorithms. Usually, proteins are mapped to their corresponding genes to form a gene interaction network. In this section, we will use genes and proteins interchangeably. Right now, many databases are available for researchers to download PPI networks. Table 2.2 summarizes several databases and the date of their latest updates. The first four databases (BioGRID, HPRD, MINT, and DIP) collect PPIs from published literature while the rest ten databases also collect PPIs from other databases. Among these databases, INstruct curates the interactions and constructs a 3D PPI network.

From Table 2.2, we can find that more than half of the databases have not been updated for at least one

[4] [78] is Chapter 3 from this thesis

Table 2.2: Some commonly used databases of PPI networks

Name	URL	Ref.	Latest Update
BioGRID	https://thebiogrid.org	[88]	12/25/2018
HPRD	http://www.hprd.org	[89]	04/13/2010
MINT	https://mint.bio.uniroma2.it	[90]	09/01/2013
DIP	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	[91]	02/05/2017
INstruct	http://instruct.yulab.org	[92]	04/15/2013
STRING	https://string-db.org	[93]	01/19/2019
InWeb_IM	http://www.intomics.com/inbio/map	[94]	09/12/2016
IntAct	https://www.ebi.ac.uk/intact	[95]	12/01/2018
PINA	http://omics.bjcancer.org/pina	[96]	05/21/2014
HIPPIE	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie	[97]	07/18/2017
HINT	http://hint.yulab.org	[98]	Version 4
iRefIndex	http://irefindex.org/wiki/index.php?title=iRefIndex	[99]	01/22/2018
Mentha	https://mentha.uniroma2.it	[100]	01/28/2019
I2D	http://ophid.utoronto.ca/ophidv2.204	[101]	07/10/2015

year. Meanwhile, since protein interaction is tissue-specific and dynamic, PPI networks downloaded from these databases contain many false positives [4], which significantly affects the accuracy of the prediction, especially for network-based methods. Many approaches have been proposed to improve the quality of the downloaded PPI networks. One of them is to filter out PPIs with low confidence scores. A few databases, such as STRING and InWeb_IM, offer confidence scores which represent the reliability of the PPI. PPIs with low confidence scores are usually collected from animal experiment or computational prediction, which can be removed if users need a high-quality network.

2.3.3 Gene expression

Gene expression profiles are the second most popular data for disease gene prediction. Before the widespread use of next-generation sequencing technologies, gene expression data measured by microarray or RNA-seq are the most accessible data to enhance the prediction of disease genes. Although gene expression cannot directly provide association information, the expression pattern of genes in different groups of samples helps us identify disease genes.

Gene expression data can be used to extract features, weight the PPI networks, and build tissue-specific networks. We can also use them to build independent networks, such as co-expression networks [48], reg-

ulatory networks and differential co-expression networks [78] [5]. The datasets of gene expression can be obtained from many platforms, such as the Gene Expression Omnibus (GEO) [102, 103] and Genomic Data Commons (GDC) Data Portal (previously known as TCGA) [104]. Similar to PPI networks, gene expression data suffer from quality issues. No matter how the expression levels are measured (microarray or RNA-seq), preprocessing should be performed to remove low-quality samples and non-expressed genes. Moreover, the data have to be normalized before cross sample analysis, especially for RNA-seq data, which are usually not normalized. Many algorithms have been proposed to normalize the raw RNA-seq count data, and details of their comparison can be found in [105].

Among all the applications of gene expression, co-expression analysis is the most frequently used one. Co-expression patterns, especially differential co-expression, reveal disease-associated properties, and many algorithms have used them to predict disease genes [106]. However, both co-expression and differential co-expression characterize only a part of the disease gene-related information. Neither of them can solve the problem alone, and gene expression data should be integrated with other types of data when designing a computational method.

2.3.4 Mutation data

Unlike other types of data which mainly provide functional similarity information, mutation data contain the association information between diseases and mutations, which has accelerated the identification of disease genes. Currently, mutation data are obtained from sequencing studies, such as GWAS, Whole Exome Sequencing (WXS) and Target Sequencing (TS).

In typical GWAS, patients and normal controls are genotyped to identify disease-associated single nucleotide polymorphisms (SNPs). For each SNP, a P-value calculated from statistical tests is used to represent its likelihood of being disease-associated. These SNPs can be further mapped to their corresponding genes to generate a group of candidate disease genes. Generally, an SNP is mapped to a gene if it is located within the gene sequence or 20 kb upstream or downstream. If multiple SNPs are mapped to the same gene, the most significant SNP (the one with the smallest P-value) would be chosen. Post-GWAS algorithms then combine these candidate disease genes and their mapped P-values with other types of data, such as PPI network, to select a subset of genes as disease genes.

Another type of data that can be obtained from GWAS are the expression quantitative trait loci (eQTL). Mutations in these loci can modulate the expression of genes, which might lead to diseases. Based on the distance between the loci and the genes, eQTL can be divided to cis (close to a gene) and trans (distal to a gene or on different chromosomes). Studies have shown that trans eQTL are more important than cis eQTL for their influence on gene expression [107, 108]. Since eQTL can provide additional disease-related information, algorithms have combined SNPs and eQTL to improve the accuracy of disease gene prediction

[5] [78] is Chapter 3 from this thesis

Table 2.3: Some commonly used databases of Pathways

Name	URL	Ref.	Latest Update
KEGG	`https://www.genome.jp/kegg`	[119, 120, 121]	01/01/2019
Reactome	`https://reactome.org`	[122]	12/13/2018
WikiPathways	`https://www.wikipathways.org`	[123]	01/01/2019
Pathway Commons	`https://www.pathwaycommons.org`	[124]	01/28/2019

[109, 110].

Currently, several databases collecting GWAS data are available, such as GWAS catalog [111] and dbSNP [112]. The most popular eQTL database is GTEx Portal. Researchers can also obtain eQTL data of specific tissues from references [113, 114].

In WXS and TS studies, various types of somatic mutations [single nucleotide variants (SNVs), insertions or deletions (indels), etc.] are identified, and computational algorithms can predict disease genes by analyzing their frequency or functional impact. However, unlike GWAS which has been conducted for more than ten years. WXS relies on the next generation sequencing technologies, and mutation data measured from these studies are not always available for all kinds of diseases. Thus, somatic mutation data are not further discussed in this review. Studies about analyzing somatic mutation data can be found in [13].

2.3.5 Pathway

It is well known that disease genes of the same or similar diseases may exist in the same biological module, such as protein complexes [115], and pathways [116]. Therefore, pathways are also valuable for disease gene prediction. Table 2.3 lists a few pathway databases that are commonly used in computational algorithms.

In earlier stages, some computational algorithms regarded unknown genes in the same pathway with the known disease genes as candidates and identified real disease genes from them [15]. Later on, Chen et al. extracted features from a pathway co-exist network in which two genes were connected if they belonged to the same pathway [117, 118]. Currently, pathways are more commonly used to validate the *do novo* prediction of the algorithms. The corresponding methods are discussed in Section 2.4.3.

2.3.6 Other types of data

Apart from the widely used data discussed in the previous sections, many other types of data are also useful in predicting disease genes. One of them is ontology data, which includes gene ontology [125, 126] and phenotype ontology [127, 128]. Ontology terms can be described by directed acyclic graphs (DAG) where nodes represent terms while edges represent semantic relations. The major application of gene (phenotype)

ontology in disease gene prediction is to calculate the similarities among genes (diseases) [129, 64] [6]. A few algorithms have been proposed to calculate the semantic similarities based on ontology data [130, 131]. We can also extract features from gene ontology data and train machine learning models with them [41, 132]. Moreover, databases like MeSH, also contains disease terms in the form of DAG which can be used to compute disease similarities [133].

Another type of data is subcellular localization, which represents where a protein resides in cells. There are 11 compartments, and two proteins may not interact with each other if they are not localized in the same compartment. Databases such as COMPARTMENTS [134] and LOCATE [135] contain experimentally validated subcellular localization data. Based on our experience, directly using data obtained from these databases and removing protein interactions if two proteins are not in the same compartments might not improve the prediction accuracy, since subcellular localization information is still being identified. A better choice is to weight the PPI networks with the currently available localization data using the strategy proposed in [136, 79].

Finally, protein sequence and domain information are related to the protein functions, which is valuable for disease gene prediction. However, due to their complexity, only a few algorithms have used them in the prediction [137, 41].

2.3.7 Data integration

Each type of data discussed in previous sections characterizes its unique biological properties. To improve the prediction accuracy, it is necessary to properly integrate different types of data so that their shortcomings are compensated. Currently, network-based strategies are most commonly used in integrating multiple types of data.

Network-based

A. Single network-based

Single network-based strategies focus on weighting the PPI network with additional information. Generally, pair-wise measurements, such as correlation coefficients, subcellular localization and gene similarities, can be used to weight the edges of the network, while single gene-related information, such as average expression level, P-values mapped from GWAS data, can be used to weight the nodes of the network.

Additionally, PPI networks can be curated to improve its accuracy. A typical example is to construct tissue-specific networks based on gene expression data. The key idea is to remove a protein interaction if one of the two interacting genes is not expressed in the corresponding tissue. However, it is difficult to determine whether a gene is expressed or not. Earlier method regarded a gene as expressed if its expression level was higher than a threshold [138]. Later on, Ganegoda et al. calculated the PCC of all interacting genes and

[6] [64] is Chapter 5 from this thesis

removed an edge if its corresponding PCC was lower than a threshold [39]. Ni et al. also used PCC of the gene expression to construct a tissue-specific PPI network, except that they constructed the network using the k-nearest-neighbor strategy where each gene in the PPI network was connected with its top k "nearest" genes based on the PCC [139].

Although co-expression-based methods are better than those with a unified threshold, neither of them are optimal. To make the obtained network more informative for disease gene prediction, Luo et al. proposed an algorithm to construct sample-specific networks [43] [7]. Specifically, a unique network was generated for each disease (case) sample. A gene i was considered expressed in a case sample if its expression level was higher than $\lambda \cdot \text{mean}(cntl[i])$, which was the mean expression levels of i in the control samples. λ was then chosen by the grid search based on the performance of the algorithm. Although this strategy might remove several true positive PPIs from the network, the remaining PPIs were closely related to the disease, which should improve the prediction of disease genes. Results of the experiments showed that prediction based on this strategy was more accurate than some previous ones, such as the subcellular localization-based method [79].

Finally, researchers can also build functional interaction (FI) networks to replace the PPI network. In an FI network, genes are connected if their functional similarities are above a defined threshold. Thus, the connected genes do not need to physically interact. The functional similarities can be calculated from protein interaction, gene expression, pathway and many other types of data [140].

B. Multiple network-based

Multiple network-based strategies construct a few biomolecular networks and combine them together. For instance, a heterogeneous network is constructed by disease similarity network, PPI network and bipartite network which represents disease-gene associations. Since disease similarity is used in a heterogeneous network, evidence for estimating disease similarities can all be used to construct heterogeneous networks. Meanwhile, heterogeneous networks can also be used to integrate multi-omics data. For instance, Lei et al. constructed a triple heterogeneous network where lncRNA–lncRNA similarity network, gene–lncRNA associations and lncRNA–disease association were fused into the network [30].

Feature-based

Feature-based strategies are used in machine learning-based methods. Raw features extracted from each type of data can be fused by deep learning models such as DBN and graph CNN. Details of these models are discussed in Section 2.2.2. Although the learning process is in a black box, the fused representations improved the prediction accuracy according to existing studies.

[7] [43] is Chapter 4 from this thesis

2.4 Evaluation methods

Many strategies have been used to evaluate the performance of computational algorithms; however, no gold standard has been proposed. In this section, we present a two-step approach for researchers to evaluate their own algorithms. The two steps consist of the minimum required strategies that should be used for evaluation. In the meantime, we also introduce several other evaluation methods, and researchers are encouraged to use them to further evaluate their algorithms.

2.4.1 Step 1: metrics

The first step to evaluate an algorithm is to compare it with other state-of-the-art methods using different metrics. Since generating a probability (or a score) for each gene, rather than a binary label, is more helpful for scientists to select potential disease genes, metrics such as the area under receiver operating characteristic (ROC) curve (AUC) and the area under Precision–Recall (PR) curve (AUPR) are recommended to compare different algorithms. ROC curve plots the true positive rate (TPR) versus the false positive rate (FPR) at different thresholds, and PR curve plots precision against recall (also known as TPR) at different thresholds. A larger area under the curve represents better overall performance. Due to the small size of input data, computational methods usually use cross-validation to obtain the prediction results. 5-fold cross-validation and leave-one-out are two commonly used approaches.

To calculate precision, recall and FPR, known disease genes are regarded as positives, while non-disease genes are regarded as negatives. Given a threshold, these metrics can be calculated by

$$\begin{aligned} \text{precision} &= \frac{\#\text{ of TPs}}{\#\text{ of TPs} + \#\text{ of FPs}} \\ \text{recall} &= \frac{\#\text{ of TPs}}{\#\text{ of TPs} + \#\text{ of FNs}} \\ \text{FPR} &= \frac{\#\text{ of FPs}}{\#\text{ of FPs} + \#\text{ of TNs}} \end{aligned} \tag{2.6}$$

where a TP (true positive) is a known disease gene predicted as positive; a FP (false positive) is a non-disease gene predicted as positive; a TN (true negative) is a non-disease gene predicted as negative; a FN (false negative) is a disease gene predicted as negative.

Usually, AUC is suitable for balanced datasets while AUPR is more useful for imbalanced datasets [11]. Considering that algorithms with high recall rate are more valuable (correctly predict true disease genes is more useful), another metric known as "recall rates at different thresholds" can be used in concert with AUC or AUPR to demonstrate the superiority of the algorithm. An algorithm with higher recall rate within the top k (typically, $k = 100$) genes is superior to its competing algorithms.

Other than AUC and AUPR, researchers can also rank all the genes based on their probabilities and calculate the "cumulative distribution function (CDF) of the rank". CDF characterizes the number of disease genes that are ranked in the top k genes of the predicted list as a function of k. Similar to "recall

rates at different thresholds", CDF of the rank also measures the algorithm's ability in predicting true disease genes with different thresholds.

In summary, researchers should choose one of the three metrics (AUC, AUPR, and CDF) to evaluate their methods.

2.4.2 Step 2: *de novo* study

As discussed in Section 2.3.1, a high accuracy might be the results of well selected non-disease genes. To further demonstrate the performance of the algorithm, a *de novo* study should be conducted as the second step to evaluate the proposed algorithm. Specifically, researchers should use their algorithms and search the predicted disease genes against those known from existing literature. Usually, disease genes are not collected into databases unless their associations have been proved by multiple studies. Thus, it is possible that a gene has been identified as a disease gene by a few studies, but still has not been collected by the benchmark dataset. If most of the genes in the top k (usually $k = 10$) predictions have been experimentally identified as disease-associated, the algorithm would be a valuable one.

2.4.3 Other evaluation methods

The two-step approach provides the typically required evaluation methods to demonstrate the performance of an algorithm. In many studies, researchers tend to use additional methods to further evaluate the performance of their algorithms. For instance, pathway enrichment analysis (PWEA) identifies statistically significant gene sets which represent functions, mechanisms, processes, etc. Given a set of predicted disease genes, PWEA uses statistical tests to verify if a pathway is over-represented among input gene sets compared to the whole genome [141]. Results of the analysis are a list of pathways, each of which with a P-value represents its significance. The more significantly enriched disease-related pathways are found in the results, the better the algorithm performs. Another method is to use databases, such as DisGeNET, to research PubMed IDs of newly published articles that report the predicted disease-gene associations.

2.5 Perspectives and conclusions

In this review, we have discussed several types of computational methods for predicting disease genes. Based on their characteristics, we roughly divide them into three groups: network-based methods, machine learning-based methods, and other methods. For each type of method, we discussed those that are valuable for developing new algorithms. Note that a thorough comparison is not provided in this review since different methods use different types of data, and a method with lower accuracy could still be valuable for developing new algorithms. For instance, many NMF-based methods outperformed IMC in their evaluations. However, with new disease and gene features, a modified IMC still performed much better than most NMF-based methods [142]. Therefore, when developing new algorithms, researchers should incorporate different

strategies properly and leverage the advantages of various types of methods. Despite the good performance, existing algorithms might be improved in several ways to allow more accurate prediction.

For network-based methods, the state-of-the-art algorithms are those that use a random walk to predict disease genes. The contribution of a path to the prediction score decreases exponentially with its length, which might not be the best option for prediction. If we set different weights for different path lengths, the random walk might be more controllable which might generate better predictions. Another issue is that many algorithms are biased towards hub genes. Using P-value instead of original prediction scores could solve this problem [129, 143]. Moreover, the performance of network-based methods depends highly on the quality of the network. Although using multiple networks or heterogeneous networks can improve prediction accuracy, most approaches still focus on genomics data. Researchers should develop more architectures that can use multi-omics data to enhance the prediction.

For machine learning-based methods, the selection of negative data is a critical issue. A good non-disease gene selection strategy might find a group of highly possible negative data, resulted in high AUC and AUPR scores. However, the accuracy of the model in predicting new disease genes may not be satisfactory, since passenger genes are not like those selected non-disease genes, which can be easily separated from known disease genes. The two-step strategy which filters out those highly possible non-disease genes in the first step might help researchers design a good model [117, 118]. Furthermore, a bootstrap strategy which allows the model to be trained with multiple types of negative data should also improve the discriminative power of the algorithms. In addition to negative data, machine learning-based algorithms might also be biased toward hub genes. This issue is mainly raised by the features extracted from biomolecular networks. Combining traditional features with novel representations, such as mutation-based features [144, 145] or graph embeddings [146] should solve this issue. Last but not least, more and more studies have used deep learning to solve biomedical issues [147]. Deep learning models can directly learn features from raw sequence data and expression data, which might provide more valuable representations than traditional hand-craft features. Meanwhile, newly developed models such as graph CNN also provide a new way for data integration, and combining graph CNN with heterogeneous networks would allow us to learn representations from multi-omics data.

Another problem lies in the need for a high-quality disease-gene association database for complex diseases. Although many databases have been released, researchers tend to have divergence in determining whether a gene is disease-associated or not. Association data downloaded from different databases vary a lot for some complex diseases. Thus, an advanced database for complex diseases would be extremely valuable for computational algorithms.

Finally, computational algorithms are developed to assist the experimental identification of disease genes; however, most of them have not been used in wet-lab studies. The main reason is that most algorithms have not been implemented as user-friendly software tools. Even if the authors have provided the source codes, users still need to preprocess their data so that the algorithms can be performed. This issue is extremely

critical when the algorithms require multiple types of data. Thus, implementing these algorithms to a web tool or user-friendly software package would significantly improve their practicability in disease gene discovery. For instance, Endeavour provided a web tool for users to predict disease genes using multiple types of data [70]. Studies conducted for the identification of Autism-associated genes [148] and Parkinson's disease-associated genes [149] have used Endeavour to prioritize candidate genes. Similarly, WGCNA has also been used to analyze expression data and identify disease-associated modules [50, 51]. Note that algorithms that have not been developed to an online tool still contribute a lot to the prediction of disease genes. These algorithms provide insights on how to analyze biological data and predict disease genes, which might be used in other tools and areas.

Acknowledgements

This work has been supported in part by Natural Science and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC).

3

Disease gene prediction by integrating PPI networks, clinical RNA-Seq data and OMIM data

Prepared as: Ping Luo, Li-Ping Tian, Jishou Ruan, and Fang-Xiang Wu. Disease gene prediction by integrating PPI networks, clinical RNA-seq data and OMIM data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(1):222-232, 2019. PL, LPT, JR and FXW discussed about the methods. PL implemented the algorithm, designed and performed the experiments. FXW supervised this study. PL and FXW wrote the manuscript. All authors read, revised and approved the final version of the manuscript.

As discussed in Chapter 1, disease-gene prediction is a positive-unlabeled learning problem, where only positive instances (disease genes) are available in benchmark datasets. To train and evaluate the models, non-disease genes have to be selected as negative instances. Most algorithms randomly select a group of unknown genes as non-disease genes, some of which might be real disease genes. A strategy should be proposed to select a set of highly possible non-disease genes.

In this chapter, a shortest path-based strategy is proposed to combine OMIM data and clinical gene expression data and select reliable non-disease genes. Applying these non-disease genes with energy-based model shows that these negative instances can significantly improve the prediction accuracy. This chapter fulfills Objective 2 of this thesis.

Abstract

Disease gene prediction is a challenging task that has a variety of applications such as early diagnosis and drug development. The existing machine learning methods suffer from the imbalanced sample issue because the number of known disease genes (positive samples) is much less than that of unknown genes which are typically considered to be negative samples. In addition, most methods have not utilized clinical data from patients with a specific disease to predict disease genes. In this study, we propose a disease gene prediction algorithm (called dgSeq) by combining protein-protein interaction (PPI) network, clinical RNA-Seq data, and Online Mendelian Inheritance in Man (OMIN) data. Our dgSeq constructs differential networks based on rewiring information calculated from clinical RNA-Seq data. To select balanced sets of non-disease genes (negative samples), a disease-gene network is also constructed from OMIM data. After features are extracted

from the PPI networks and differential networks, the logistic regression classifiers are trained. Our dgSeq obtains AUC values of 0.88, 0.83 and 0.80 for identifying breast cancer genes, thyroid cancer genes and Alzheimer's disease genes, respectively, which indicates its superiority to other three competing methods. Both gene set enrichment analysis and predicted results demonstrate that dgSeq can effectively predict new disease genes.

3.1 Introduction

Complex diseases are usually caused by the malfunction of a group of genes, known as disease-associated genes or disease genes. Identifying these genes is critical for understanding the mechanisms of diseases. Traditional methods such as GWAS and linkage analysis usually generate hundreds of candidate disease genes, making the further validation time-consuming and expensive [11]. As a result, many researchers have developed efficient computational methods to predict and prioritize candidate disease genes to reduce the number of candidates while helping scientists optimize the in-depth wet lab validation.

Based on whether the algorithms require known disease-gene associations as input, existing algorithms can be divided into two categories: undifferentiated and differentiated [11]. Undifferentiated algorithms treat all the genes in the genome equally, and provide overall probabilities for all the genes involved in a disorder. For instance, dmGWAS [150] and EW_dmGWAS [58] first searched dense modules from a PPI network weighted by GWAS and gene expression data. Then, genes in the top 1% of the ranked modules were regarded as disease genes. MetaRankder 2.0 prioritized candidate genes by integrating five kinds of heterogeneous data [151]. Text mining algorithms such as MeSHOP [152] and Genie [153] searched candidates from biomedical literature and generated a list of ranked genes for a given specific disease. Those methods are useful especially for disorders that have no known disease genes.

Differentiated algorithms analyze the known disease genes along with other biological data, and provide more valuable information than undifferentiated algorithms. Many machine-learning-based and statistics-based algorithms are differentiated. An example is the popular tool Endeavor, which prioritized candidate genes according to the relationships between user submitted training genes and candidate genes in various kinds of biological data. Another example is the network energy-based algorithms proposed by Chen et al. [48, 117, 118], where genes are classified as being disease-associated or not according to the posterior probabilities calculated by a formula derived from the Boltzmann distribution and their defined network energy function.

Differentiated algorithms, especially machine learning-based algorithms, have gained success in the prediction of various kinds of disease genes. However, since the number of known disease genes (positive samples) is far less than that of unknown genes (negative samples), most machine learning-based algorithms have to face an imbalanced classification problem. Moreover, no databases contain non-disease genes for a specific disease. Thus, training an accurate machine learning model for predicting disease genes is usually difficult.

To solve this imbalanced classification issue, one possible solution is to divide the individual diseases into disease classes. Since the number of genes in each class is much more than that of a specific disease, training a model with disease classes is possible. Algorithms such as the RWR of Köher et al. [18] and the MRF method of Chen et al. [47] used this strategy. Another approach is to narrow down the non-disease genes space. For instance, in their "two step" method [117], Chen et al. first removed the genes with low relevance to the disease under consideration, then predicted disease genes from the remaining genes. This strategy successfully predicted cancer-related genes; however, negative samples in the training set were still five times higher than positive samples. As well, the method required individual diseases to be laboriously classified into disease classes. Although Goh et al. classified disease genes into 22 classes in [154], their dataset was out of date. Databases such as Online Mendelian Inheritance in Man (OMIN) [155] are updated daily and require time and expertise to be classified.

In this study, we propose a strategy to reduce the number of non-disease genes. Unlike existing methods, our strategy selects non-disease genes for a specific disease rather than a disease class. The obtained non-disease genes are directly selected from the latest OMIM dataset. No classification is needed in the process, thus improving the accuracy and efficiency of the algorithm. Additionally, since the number of non-disease genes in the training set is similar to the number of disease genes, our model avoids the imbalanced classification problem.

In addition to non-disease gene selection strategy, we integrate PPI networks and RNA-Seq Data to improve the prediction accuracy. Previous research has shown that the integration of PPI networks and gene expression data is valuable for predicting essential proteins and protein function [156, 157]. In this work, we integrate these two types of data by extending our previous study on 'guilt by rewiring' [158]. A network is considered to be rewired if its edges (wires) are changed during a specific process. This phenomenon is observed in many biomolecular networks, such as regulatory networks, protein-protein interaction (PPI) networks and co-expression networks. Previous studies have shown that network rewiring is an important implication for analyzing biological data. For instance, Hu et al. showed that besides differential expression, rewiring information was very useful for analyzing gene expression data [159]. In our study, the rewiring of co-expression network from control to case subjects is used to predict disease genes. In [160], Hou et al. demonstrated that co-expression between disease genes were more frequently rewired than a random pair of genes. The major reason behind this phenomenon is that disease genes are usually extensively expressed in case subjects compared to control subjects (differential expression), while non-diseases may maintain a stable expression level in different conditions. This difference raises the variance of the correlations between genes, which is reflected on the co-expression network.

In [158], the rewiring information calculated from gene expression under different conditions (case and control) was employed to weight the PPI network, and predicted disease genes through a logistic regression model trained on the features extracted from the weighted PPI network. This strategy is usually useful, except for non-disease genes with large degrees. Although the average weights (rewiring information) around

these genes are less than real disease genes, they may have similar features as disease genes because of their high degrees. To solve this problem, in this study, instead of weighting the PPI network with gene expression data, we used the rewiring information computed from expression data to build an independent scale-free differential network. This differential network, combined with the PPI network, is employed to extract features for predicting disease genes. The new features extracted from two networks can reveal the topological structure of PPI and the rewiring information of all genes at the same time, which can solve the problems in the previous model. The challenges of the new model is to build a valuable differential network, which is discussed in Section 3.2.2. It is noting that the expression data in [158] was measured by microarray, which has been replaced by RNA-Seq by many databases because of its limitations. Thus, in this study, the rewiring information is calculated from clinical RNA-Seq data instead. Experiments performed on Breast Cancer (BC), Thyroid Cancer (TC) and Alzheimer's disease (AD) reveal that our new algorithm is superior to existing methods. An implementation of dgSeq is available at: `https://github.com/luoping1004/dgSeq`.

The rest of the paper is organized as follows. Section 3.2 describes the methods and materials used in the study. Section 3.3 analyzes the experimental results of the algorithm and compares dgSeq with three other competing algorithms. Section 3.4 draws some conclusions.

3.2 Methods and materials

The work flow of the algorithm is depicted in Fig. 3.1. First, the RNA-Seq data from case subjects (a) and control subjects (b) are used to build a differential network by the strategy proposed in Section 3.2.2 (c). Then, a disease-gene network (d) is constructed with OMIM data, and labels of all the genes (e) are determined by (c), (d) and benchmark disease genes. These labels are used to label the PPI network (f) and differential network (c). After that, features of the known disease genes and non-disease genes are extracted from (c) and (f). A logistic regression model (h) is trained by the extracted feature matrix (g) and its corresponding labels (e). Finally, the probability of a gene being labeled as 1 (disease gene) (i) is calculated in each round of the cross validation. Details of the algorithm are discussed in the following subsections 3.2.1–3.2.4. Subsections 3.2.5 and 3.2.6 explain the validation methods and data sources, respectively.

3.2.1 General model

Identifying disease genes from a biomolecular network can be formulated as a network labeling problem in which disease genes are labeled as 1 while non-disease genes are labeled as 0. Let $g_1, g_2, \ldots, g_h$ represent all the h genes in the network. A set of binary labels $x = (x_1, x_2, \ldots, x_h)$ of these h genes is known as a configuration of the biomolecular network, and the set of all possible configurations X is a random field. The probability distribution of the configuration x of a random field X can be calculated by Boltzmann distribution [45]

$$P(x) = \frac{1}{Y} \cdot \exp(-\kappa H(x)) \tag{3.1}$$

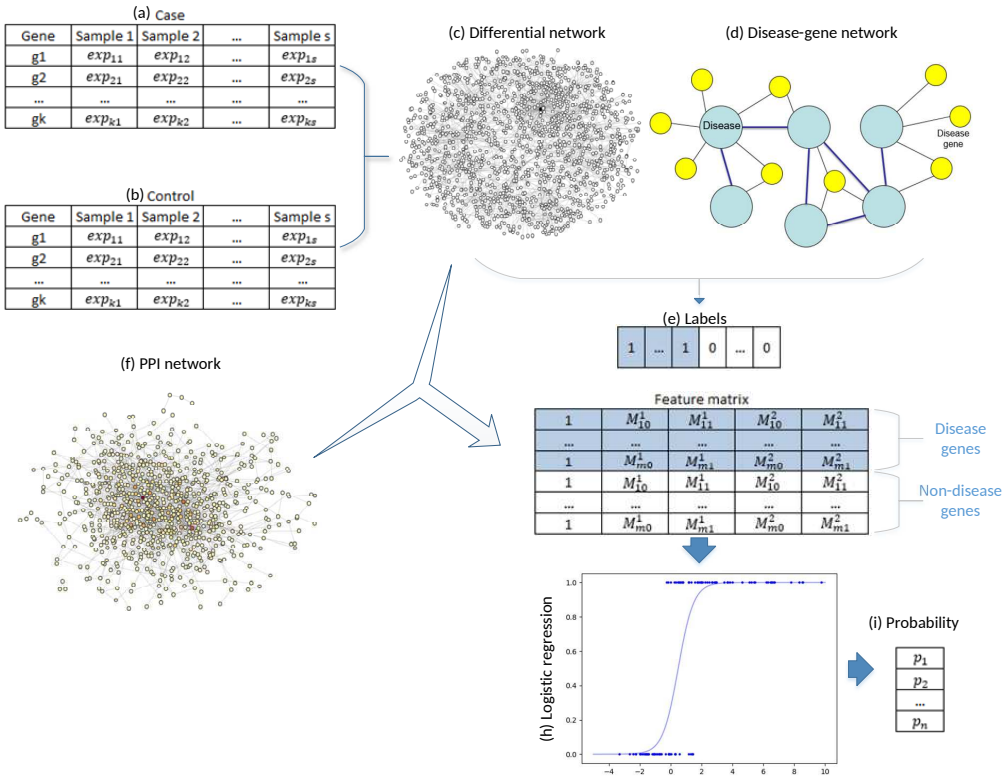

Figure 3.1: The work flow of dgSeq. (a)–(b). Clinical RNA-Seq data of the case and control subjects; (c). Differential network constructed by (a) and (b); (d) disease-gene network constructed by OMIM data; (e). Labels of all the genes determined by (c), (d) and benchmark disease genes; (f). PPI network; (g). Feature matrix extracted from (c) and (f); (h). Logistic regression model trained with (e) and (g); (i). The calculated probabilities of all the genes being labeled as 1 (disease gene).

where $H(x)$ is the Hamiltonian (energy) of the configuration x, κ is a positive constant parameter, and Y is called the partition function and defined as $Y = \sum_{x \in X} \exp(\kappa H(x))$.

Let $x_{[-i]}$ be the binary labels of all nodes except for node i in a network. Then, knowing the labels (disease or non-disease) of other genes (that is, $x_{[-i]}$), the probability that gene i is a disease gene is a conditional probability $P(x_i = 1|x_{[-i]})$. By Bayes' rule we have

$$P(x_i = 1|x_{[-i]}) = \frac{P(x_i = 1, x_{[-i]})}{P(x_i = 1, x_{[-i]}) + P(x_i = 0, x_{[-i]})} \tag{3.2}$$

In (3.2), $(x_i = 1, x_{[-i]})$ is the configuration that gene i is a disease gene while $(x_i = 0, x_{[-i]})$ is the configuration that gene i is a non-disease gene. The probability of both configurations can be calculated by (3.1).

By adopting the Ising model to calculate the Hamiltonian in [46, 47, 48], the probability $P(x_i = 1|x_{[-i]})$ in (3.2) can be parameterized as follows

$$P(x_i = 1|x_{[-i]}, \tilde{\mu}) = \frac{\mathbf{e}^{\alpha+\beta M_{i0}+\gamma M_{i1}}}{\mathbf{e}^{\alpha+\beta M_{i0}+\gamma M_{i1}} + 1} \tag{3.3}$$

where $\tilde{\mu} = (\alpha, \beta, \gamma)$ are model parameters. M_{i0} and M_{i1} are the numbers of neighbors of node i with label

0 and 1, respectively. Furthermore, if Z networks are available for determining disease genes, (3.3) can be generalized as follows [46]

$$P(x_i = 1|x_{[-i]}, \mu) = \frac{exp(V(i))}{exp(V(i)) + 1} \tag{3.4}$$

where

$$V(i) = \alpha + \sum_{z=1}^{Z}[\beta^z \cdot M_{i0}^z + \gamma^z \cdot M_{i1}^z],$$

$\mu = (\alpha, \beta^z, \gamma^z)$ $(z = 1, \ldots, Z)$ are model parameters. M_{i0}^z and M_{i1}^z are the number of neighbors of node i with labeled 0 and 1 in the z-th network, respectively. Clearly, (3.4) follows a logistic model

$$P(x_i = 1|x_{[-i]}, \mu) = \frac{exp(\mu^T \varphi_i)}{exp(\mu^T \varphi_i) + 1} \tag{3.5}$$

where

$$\begin{aligned} \varphi_i &= (1, M_{i0}^1, M_{i1}^1, \ldots, M_{i0}^z, M_{i1}^z)^T \\ \mu &= (\alpha, \beta^1, \gamma^1, \ldots, \beta^z, \gamma^z)^T. \end{aligned} \tag{3.6}$$

We can also have

$$\begin{aligned} P(x_i = 0|x_{[-i]}, \mu) &= 1 - P(x_i = 1|x_{[-i]}, \mu) \\ &= \frac{1}{exp(\mu^T \varphi_i) + 1} \end{aligned} \tag{3.7}$$

which computes the probability of a gene being labeled as 0.

Given a known configuration, the parameters in μ can be estimated by the following likelihood function

$$\hat{\mu} = \arg\max_{\mu}(\prod_{i=1}^{h} P(x_i|x_{[-i]}, \mu)) \tag{3.8}$$

However, since the number of disease genes is far less than that of unknown genes, the parameter estimated by (3.8) is inaccurate because of the sample imbalanced problem. To address this problem, we employ the under sampling strategy which uses (3.9) to replace (3.8) in this study.

$$\hat{\mu} = \arg\max_{\mu}(\prod_{i=1}^{2m} P(x_i|x_{[-i]}, \mu)) \tag{3.9}$$

For the disease d under consideration, m is the number of known disease genes associated with d. S_{ndg} contains a set of non-disease genes. We perform the under sampling to randomly select m non-disease genes from S_{ndg}, and (3.9) estimates μ based on the features of the m disease genes and m non-disease genes. The under sampling is performed 100 times and each time computes a μ which is then used to compute the probabilities of unknown genes being labeled as 1. Finally, for each unknown gene, its average probability of being labeled as 1 in the 100 runs is regarded as its probability of being disease-associated. The algorithm for defining non-disease genes with d is discussed in Subsection 3.2.3.

Maximizing the likelihood function in (3.9) is equivalent to maximizing the log likelihood function in (3.10) as follows

$$\hat{\mu} = \arg\max_{\mu} \mathcal{L}(\mu) \tag{3.10}$$

where

$$\mathcal{L}(\mu) = \sum_{i=1}^{2m} \ln(P(x_i|x_{[-i]}, \mu)) \tag{3.11}$$

Substituting (3.5) and (3.7) into (3.11) yields

$$\mathcal{L}(\mu) = \sum_{i=1}^{2m} \{x_i \mu^T \varphi_i - \ln[1 + exp(\mu^T \varphi_i)]\} \tag{3.12}$$

Since (3.12) is a concave function of μ [161], the optimization problem (3.10) can be solved by Python's library SciPy using the optimization function *minimize()* through searching the minimum solution of the convex function $-\mathcal{L}(\mu)$ [162].

3.2.2 Differential network construction

Not only are the expressions of disease genes in case subjects are significantly different from those in control subjects, but also their correlations in case subjects should also be significantly different from those in control subjects (rewiring). Actually, it is believed that a pair of disease genes are more frequently rewired than a random pair of genes in the genome. To take the rewiring information into account, in this study the differential networks are constructed with clinical RNA-Seq data.

Given two genes g_i and g_j with their corresponding v-dimensional expression values $(g_{i1}, g_{i2}, \ldots, g_{iv})$ and $(g_{j1}, g_{j2}, \ldots, g_{jv})$, respectively, their Pearson correlation coefficient (PCC) can be calculated as follows

$$r(g_i, g_j) = \frac{\sum_{q=1}^{v}(g_{iq} - \bar{g}_i)(g_{jq} - \bar{g}_j)}{\sqrt{\sum_{q=1}^{v}(g_{iq} - \bar{g}_i)^2}\sqrt{\sum_{q=1}^{v}(g_{jq} - \bar{g}_j)^2}} \tag{3.13}$$

where $\bar{g}_i$ and $\bar{g}_j$ are the mean of the expression values of g_i and g_j, respectively.

For a pair of genes g_i and g_j, let r_{ij}^{case} denotes the PCC between the expression of genes g_i and g_j in case subjects, and r_{ij}^{cntl} denotes their PCC in control subjects. Instead of Fisher's test of difference used in our previous study [158], we directly compute an absolute value

$$p_{ij} = |r_{ij}^{case} - r_{ij}^{cntl}| \tag{3.14}$$

for all pairs of genes i and j $(i, j \in [1, h], i \neq j)$, and obtain a correlation difference matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1h} \\ p_{21} & p_{22} & \cdots & p_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ p_{h1} & p_{h2} & \cdots & p_{hh} \end{bmatrix}$$

which contains the rewiring information between the case and control subjects.

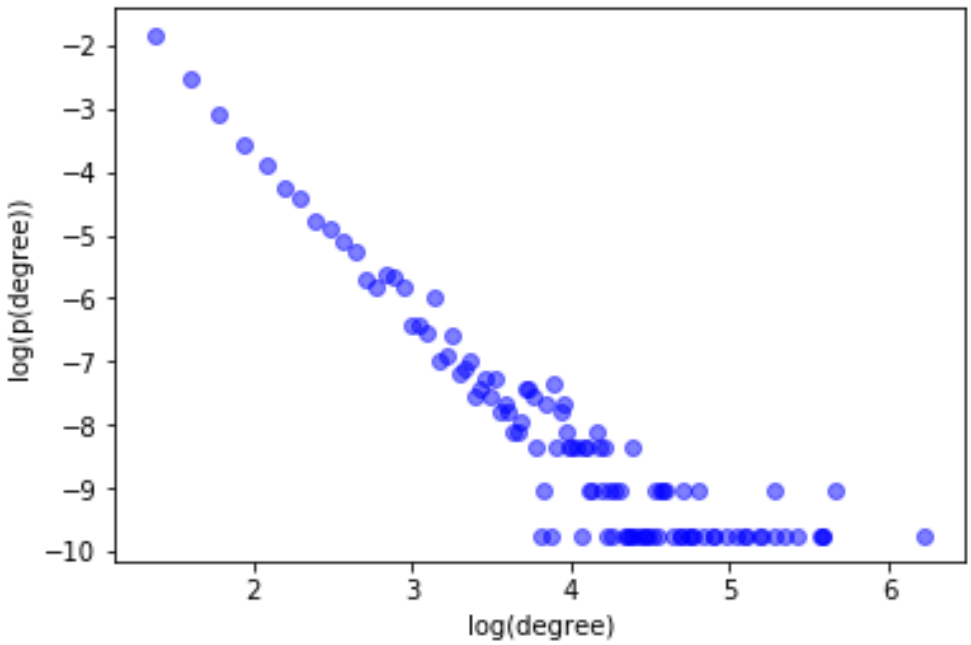

Figure 3.2: Scatter plot of the log-log degree distribution of G_{dif} for BC .

Although matrix P could be used as an adjacency matrix to construct a differential network, this network would contain too much noise since the process of producing clinical RNA-Seq data has various noisy resources. To filter out the noisy edges which typically correspond to a small value in matrix P, we use the k nearest neighbors (K-NN) algorithm [163] to determine whether an edge should be kept. Specifically, for gene i, the largest k entries in the i-th row of the matrix P are kept while others are set to be zeros. Let p_{ij} represent one of the k kept entries in row i. Then an edge is added between i and j for each of the k entries. Finally, a differential network G_{dif} is built by adding edges for all the largest k entries in every row of the matrix P.

For different values of k, the degree distribution of the constructed differential network is different. Then, it is nontrivial to choose a reasonable k for constructing a meaningful differential network. Since many biological networks are scale-free, such as the PPI network used in our study, it is believed that a scale-free differential network would be more reasonable. We use different values of k to construct the networks, and find that the differential networks are scale-free for all values of k from 3 to 9. However, the large value of k may include more noisy edges while the small value may exclude more informative edges. In this study, we use $k = 5$ in the experiments. More details about this threshold is discussed in the **Results**. Fig. 3.2 shows the scatter plot of the log–log degree distribution for G_{dif} with $k = 5$ constructed based on BC's clinical RNA-Seq data. Apparently, G_{dif} is scale-free. The other differential network for TC is similar to Fig. 3.2.

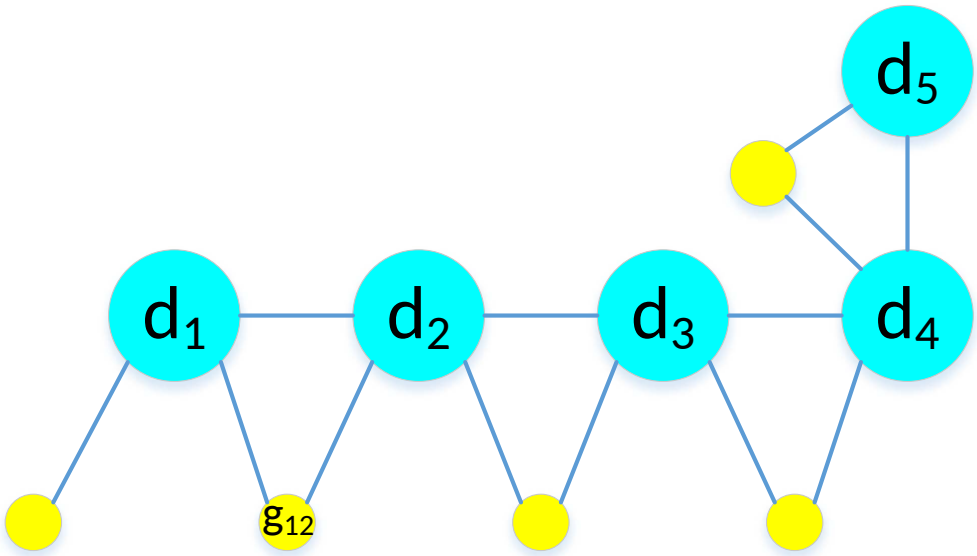

Figure 3.3: Sample disease gene network

3.2.3 Non-disease genes

Although no databases contain disease specific non-disease genes, we adopt a strategy to determine non-disease genes. The strategy includes three steps. First, we build a disease-gene network (DGN) from OMIM data and select a group of genes from DGN as potential non-disease genes. Second, another group of non-disease genes are collected from the differential network G_{dif}. Third, two groups of candidate non-disease genes are combined to form the final set of non-disease genes.

Select non-disease genes from DGN

A disease-gene network is built to select the initial group of non-disease genes. In this network, each node represents either a disease or a disease-associated gene. Every disease node is connected with its associated gene nodes, and two disease nodes are connected if they share at least one same disease-associated genes. To illustrate, Fig. 3.3 depicts a sub network of DGN with 5 diseases $(d_1, d_2, d_3, d_4, d_5)$ and 5 disease genes. d_1 and d_2, d_2 and d_3, d_3 and d_4, d_4 and d_5 are connected because they share at least one disease-associated gene.

The length of the shortest path between two diseases in the DGN represents their relationships. Length of 1 (d_1 and d_2) indicates two diseases have at least one same disease-associated gene. Length of 2 (d_1 and d_3) indicates that there is some other disease to which both two diseases are connected. Length of 3 (d_1 and d_4) indicates that the neighbors of two diseases are connected. If two diseases are neighbors, they may have similar mechanisms. Thus, d_1 and d_3 may have similar mechanisms because both of them are connected with d_2. d_1 and d_4 may also have similar mechanisms because d_2 and d_3 are connected with each other. Moreover, d_1 and d_4 are less likely to have similar mechanism compared to d_1 and d_2. Therefore, the longer

of the shortest path between two diseases in DGN, the less possible they have similar mechanisms. If two diseases d_i and d_j have completely different mechanisms, their associated disease genes should have different properties, which means the disease genes of one disease d_i can be regarded as the non-disease genes of another disease d_j.

To determine whether the distance between two diseases is enough to show that they have different mechanisms, we need to set a threshold for the length of the shortest path. If we set the threshold to 4, in a special situation, g_{12} would be a non-disease gene of d_5 because the length of the shortest path from d_1 to d_5 is equal to 4. However, g_{12} is also a disease gene of d_2 and the distance between d_2 and d_5 is only 3, which means g_{12} should not be a non-disease gene of d_5.

To address this problem, let d_i denotes the disease under consideration, $G(d_{[-i]})$ denotes the set of genes in DGN not associated with it. Instead of computing the length of the shortest path between each disease and d_i, we compute the length of the shortest path (η) between each gene g_k in $G(d_{[-i]})$ and d_i. If $\eta \geq \Gamma_1$, we consider g_k as a potential non-disease gene for d_i. Γ_1 is a predefined threshold, which is set as 5 in this study. If there is no path between g_k and d_i, we also select g_k as a non-disease genes. These selected candidate non-disease genes form a set S_1.

Select non-disease genes from G_{dif}

After selecting non-disease genes from the DGN, we find out that the same strategy can also be employed on the differential network.

Since disease genes are more frequently rewired, the value p_{ji} calculated by (3.14) corresponding to disease gene i is more likely to be larger than those values corresponding to non-disease genes in the j-th row. Therefore, compared with a non-disease gene, p_{ji} has more chance to be in the largest k entries of the j–th row. In another word, a disease gene g_i is more likely to be connected with other genes in G_{dif} than a non-disease gene. Let $G(d_i)$ denotes the set of m disease genes associated with d_i, compared to a non-disease gene, a potential disease gene g_k should be closer to the known disease genes in G_{dif}. Thus, if the smallest distance of all the shortest paths between gene g_k and the genes in $G(d_i)$ is larger than or equal to a predefined threshold Γ_2, we consider g_k as a non-disease gene. In this study, Γ_2 is set as 4, and S_2 is used to denote the set that contains all the non-disease genes selected from G_{dif}.

Generate non-disease gene set

Once we obtain S_1 and S_2 from DGN and G_{dif}, respectively, genes in the union of the two sets ($S_{non} = S_1 \cup S_2$) are regarded as non-disease genes, and labeled as 0. These genes, along with the m known disease genes, are used to train the models (3.5) or (3.7). Genes contained in the intersection of the two sets ($S_{ndg} = S_1 \cap S_2$) are considered to be non-disease genes with the highest possibilities. We randomly select m genes from S_{ndg} as the benchmark non-disease genes. These genes along with the m known disease genes are used in the cross-validation.

3.2.4 Feature extraction

Considering that we have two networks (PPI and G_{dif}) in the study, feature vector (3.6) is specified as follows

$$\varphi_i = (1, M_{i0}^1, M_{i1}^1, M_{i0}^2, M_{i1}^2)^T \tag{3.15}$$

where M_{i0}^1 and M_{i1}^1 (M_{i0}^2 and M_{i1}^2) represent the numbers of neighbors of g_i which are labeled as 0 and 1 in the PPI network (G_{dif}), respectively.

To extract features, we need to assign labels for all the genes in the two networks. As discussed in the previous sections, disease genes are labeled as 1 while non-disease genes in S_{non} are labeled as 0. The remaining genes are treated as unknown. As unknown genes could be disease genes with a small possibility, in this study 0.01% of them (BC: 119, TC: 126, AD: 100) are randomly labeled as 1 while the others being labeled as 0. Note that the information of unknown genes with such a labeling is used only for extracting the features of known disease genes and non-disease genes in S_{non}, and yet the features of unknown genes are not extracted and used for estimating the parameters in (3.9).

3.2.5 Validation methods and evaluation criteria

To investigate its performance, we perform the under sampling which randomly selects m non-disease genes from S_{ndg} 100 times, and run our algorithm 100 times respectively with the set of m disease genes and one of 100 sets of m non-disease genes. For each pair of benchmark genes, the leave-one-out cross validation (LOOCV) is employed to validate the algorithm. In each round of the LOOCV, one of the benchmark genes (the validation gene) is regarded as unknown and labeled randomly as all the other unknown genes. This strategy allows us to hide the information of the validation gene from the training genes. Then, parameters in μ (Eq. 3.12) is calculated based on the features of the training genes extracted from the new labels, and the probability of the validation gene being labeled as 1 is computed by (3.5).

The area under the receiver operating characteristic (ROC) curve (AUC) is employed as one of the evaluation criteria. The ROC curve plots the true positive rate (TPR) verse the false positive rate (FPR) at various thresholds. The TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{3.16}$$

$$FPR = \frac{FP}{TN + FP} \tag{3.17}$$

where TP, FP, TN, and FN are the numbers of true positive, false positive, true negative, and false negative, respectively. In this study, a true positive is a disease gene identified as a disease gene, a false positive is a non-disease gene identified as a disease gene, a true negative is a non-disease gene identified as a non-disease gene, and a false negative is a disease gene identified as a non-disease gene.

The ROC curve features the TPR on the Y axis, and the FPR on the X axis. This makes the top left corner of the plot an ideal point, with a FPR of 0 and TPR of 1, and it also means that a method with a

larger AUC performs better. In this study, we obtain 100 AUC values from 100 runs, the average of which is used as the AUC of the algorithm.

In terms of AUC, we compare dgSeq with the "two-step" (2Step) and "Rebalancing" (Re-Balanced) algorithms [117, 118]. These two algorithms outperformed their competing methods in the identification of cancer-related genes.

To further evaluate our algorithm, we compare dgSeq with Endeavor through gene set enrichment analysis (GSEA). In a previous study that compared eight public available web-based tools, Endeavor was one of the two best algorithms when all performance measures were considered [164]. In this study, we first use the latest version of Endeavor to rank all the unknown genes. Then, for dgSeq, we also rank all the unknown genes according to their probabilities of being labeled as 1. In each round of the 100 runs, we compute the probability of each unknown gene being labeled as 1 by (3.5). The corresponding parameter vector μ^T is calculated by (3.12) with the features of the $2m$ benchmark genes in each round. Finally, for each unknown gene g_i, we obtain 100 probabilities, the average of which is regarded as the probability of gene g_i being labeled as 1. The top 100 genes in the two lists are then analyzed and compared in terms of GSEA using WebGestalt [165, 166, 167].

Finally, deSeq is also evaluated on predicting disease genes associated with Alzheimer's disease.

3.2.6 Data sources

The BC-associated and TC-associated genes are collected from the Cancer Gene Census category (CGC, http://cancer.sanger.ac.uk/census#) [87]. 35 BC-associated and 34 TC-associated genes are chosen as the benchmark disease genes. The AD-associated genes are collected from MalaCards: The human disease database (http://www.malacards.org/), which contains 182 ranked genes for AD. We select the top 50 as AD-associated genes. Among these 50 genes, 43 of them appear in the PPI network, and are used as the benchmark.

The cancer case and control gene expression data are downloaded from the Genomic Data Commons (GDC) [104]. GDC measures the data by RNA-Seq technique, and provides three types of values: Fragments Per Kilobase of transcript per Million mapped reads (FPKM), Upper Quartile normalized FPKM (UQ-FPKM) and the raw mapping count. To facilitate cross-sample comparison, we use the UQ-FPKM values in the study. In total, the data sets contain 1222 case subjects and 113 control subjects for BC, and 502 case subjects and 58 controls subjects for TC. The AD RNA-Seq data are downloaded from Gene Expression Omnibus (GSE53697) [168], which contains the raw mapping count files of 9 case subjects and 8 control subjects. We normalize the data with DESeq2 [169], because DESeq2 was proved to be one of the best algorithms for RNA-Seq data normalization [105]. During the preprocessing, genes not in the PPI network or not expressed (expression values are 0) are removed from the data sets.

The PPI network is obtained from the InWeb_InBioMap database (version 2016_09_12) [170], which consists of 17,653 nodes and 625,641 interactions aggregated from eight source databases. We map proteins

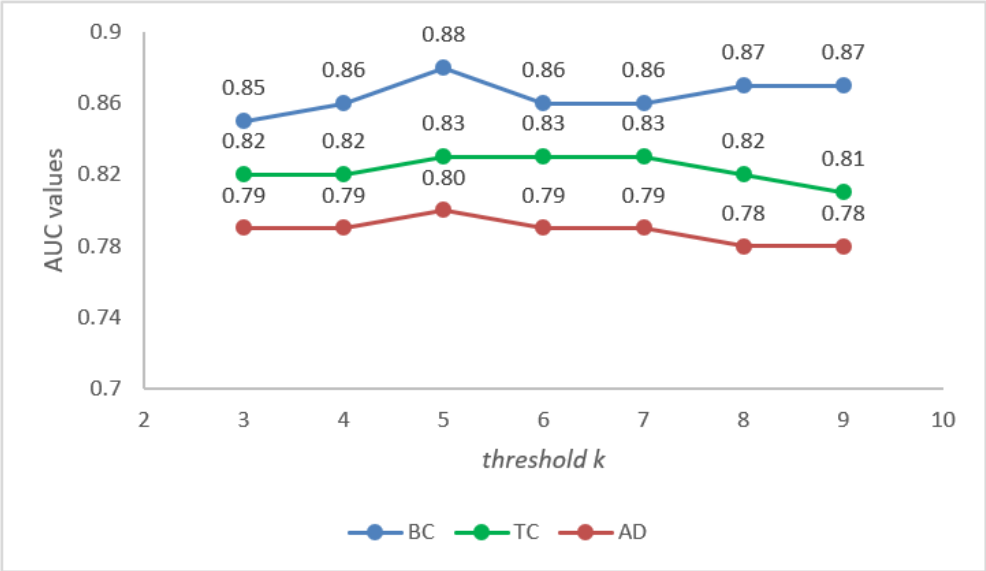

Figure 3.4: Sensitivity analysis of threshold k

in the network to their corresponding genes, and remove those genes that have no expression data from the network. To simplify the network, proteins correspond to multiple genes are also removed from the PPI network. As a result, the final PPI networks contain 16,945 nodes and 589,234 edges for BC, 16,837 nodes and 587,537 edges for TC, 15056 nodes and 520,211 edges for AD.

The disease-gene association data used to build DGN are obtained from the OMIM database (Feb 17, 2017) [155]. The original data set consists of 4450 diseases and 3402 disease genes when we only consider the diseases with known molecular basis. Then, the disease genes not in the PPI network are removed from the data set, and finally the data set contains 3221 diseases and 3187 disease genes

3.3 Results and discussion

3.3.1 Threshold selection

In this study, we define three thresholds: Γ_1, Γ_2 and k. The first two thresholds are used to determine whether a gene is a non-disease gene. We empirically set them as 5 and 4, respectively. The third threshold k determines the minimum number of neighbors around each node in G_{dif}. This value controls the number of edges in G_{dif}, which will further affect the selection of non-disease genes. We choose k from 3 to 9, and perform a sensitivity analysis. Fig. 3.4 depicts the results of the analysis for all three disease data. We can see that the AUC of dgSeq is varying with respect to the threshold k. We choose $k = 5$ in the study because the algorithm performs best with this threshold in terms of AUC.

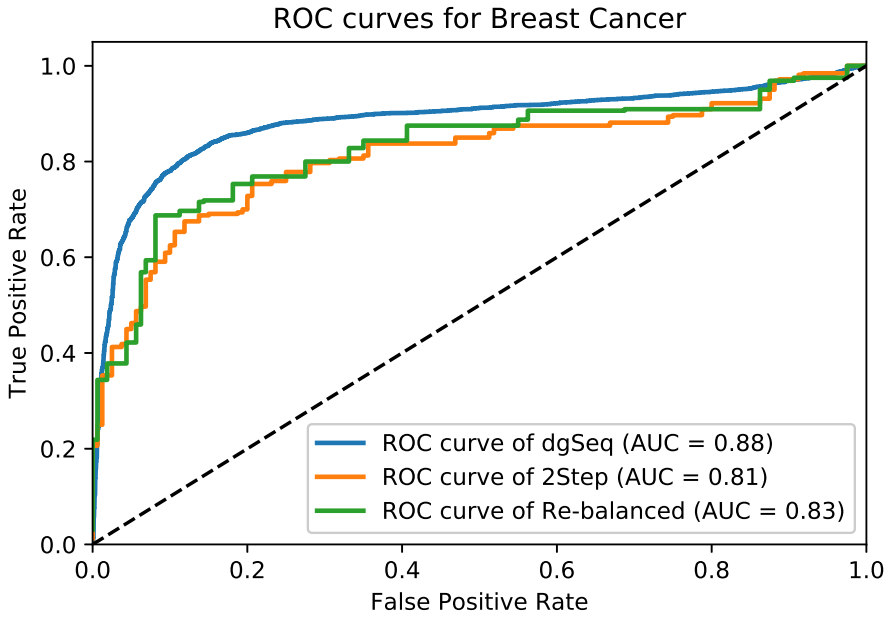

Figure 3.5: The ROC curves of three algorithms in predicting BC-related genes

3.3.2 The results of AUC values

Fig. 3.5, 3.6 and 3.7 show the ROC curve of the three algorithms in predicting BC-associated, TC-associated and AD-associated genes with the same sets of benchmark genes, respectively. For BC, dgSeq obtains an average AUC value of 0.88, whereas the two competing (2Step and Re-balanced) methods only achieve 0.81 and 0.83, respectively. For TC, the AUC values of the two competing algorithms are smaller than 0.80, which is less than 0.83 from our dgSeq. For AD, dgSeq obtains an average AUC values of 0.80 while the AUC values of the two competing algorithms are around 0.5. It is worth noting that the two competing methods were developed to predict cancer-associated genes, which is the reason why their performance in AD is almost like random. However, their principle allows them to predict non-cancer disease genes such as AD. In a word, dgSeq outperforms the two competing methods in the experiments.

3.3.3 Enrichment analysis

To further evaluate our algorithm, we rank the unknown genes with dgSeq and perform GSEA on the top 100 genes regarded as potential new disease genes. We also perform the same analysis on the top 100 ranked genes reported by Endeavor. The size of gene universe in GSEA is 26,533. The enriched pathways are ranked by their corresponding P-values in ascending order, and the top 10 enriched pathways are listed in Tables 3.1, 3.2, 3.3 for BC, TC and AD, respectively.

Among the ten pathways in Table 3.1, candidate BC disease genes reported by dgSeq are enriched in six cancer-related pathways: the “Thyroid hormone signaling pathway”, “TGF-beta signaling pathway”,

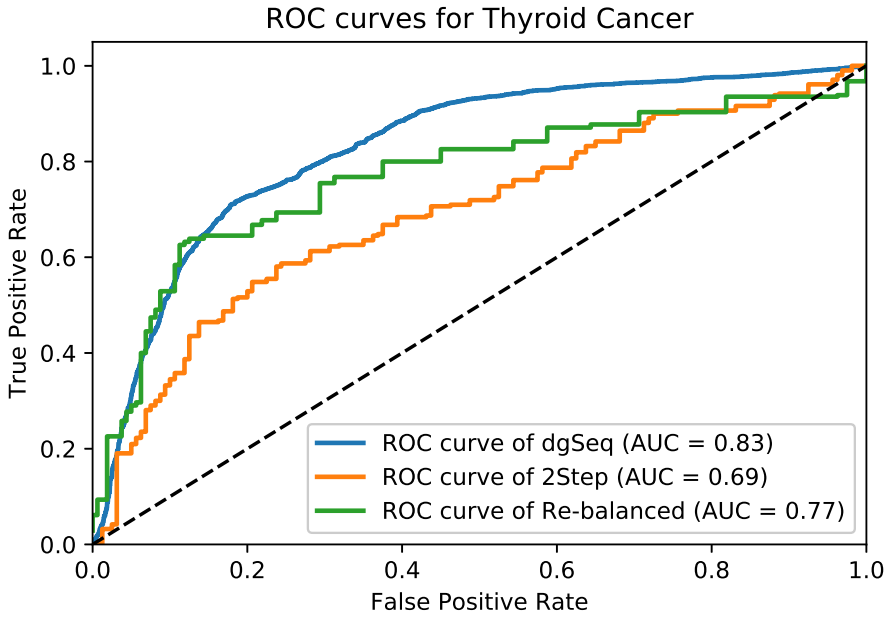

Figure 3.6: The ROC curves of three algorithms in predicting TC-related genes

"Epstein-Barr virus infection", "Pathway in cancer", "Breast cancer" and "Hepatitis B". The "Thyroid hormone signaling pathway" contains multiple thyroid hormone receptor (TR) isoforms, the mutation of which may lead to various kinds of cancers, such as thyroid cancer and breast cancer [171, 172]. "TGF-beta signaling pathway" is related to breast cancer because TGF-β1 was found to be linked with increased tumor progression and cancer invasiveness in late stages of breast cancer. Several drugs against TGF-β1 have also been developed to treat breast cancer [173]. "Epstein-Barr virus infection" pathway affects the infection of Epstein-Barr virus (EBV), which has strong connection with breast cancer [174]. Research has shown that EBV may accelerate the development of malignant breast cancer [175]. "Pathways in cancer" has been targeted for many drugs, as well as "Breast cancer" pathway. "Hepatitis B" pathway controls another kind of virus infection which may cause breast cancer [176]. Interestingly, the "Longevity regulating pathway" is enriched with 5 genes. Although this pathway is not directly related with breast cancer, longevity is a well-known feature of cancer. Thus, we further analyze the 5 enriched genes and find that 4 of them (SIRT1, HDAC1, HDAC2, RPS6KB1) have been studied as BC-related genes in previous studies [177, 178, 179]. However, candidate genes reported by Endeavor are only enriched in the "FoxO signaling pathway" with a P-value of 9.98×10^{-1}. Although this pathway is related with breast cancer, the P-value of the analysis is much lager than the P-values of the pathways enriched by the top 100 genes reported by dgSeq.

From the results of GSEA for TC in Table 3.2, all the 10 pathways are cancer-related. Among these pathways, the "PI3K-Akt signaling pathway" plays a pivotal role in many key cellular processes, and thyroid cancer has been shown to be highly associated to this pathway in previous studies [180]. The "FoxO signaling pathway" is also correlated with many cancers, and one of the genes on this pathway (FOXO3) is reported

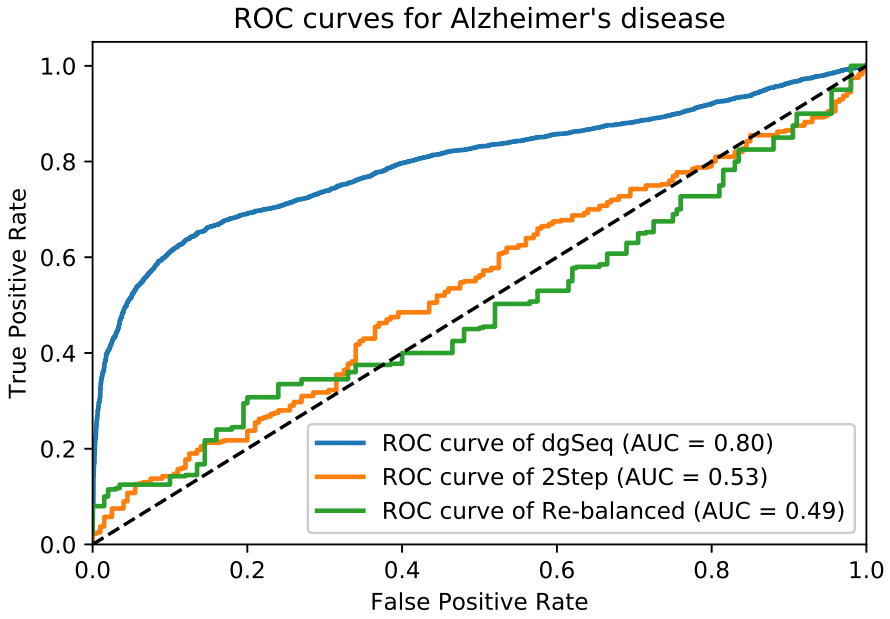

Figure 3.7: The ROC curves of three algorithms in predicting AD-related genes

to be a driver gene of thyroid cancer [181]. Similar to BC, the reported candidate genes are also enriched in the "Prostate cancer" pathway and "Pathways in cancer". Because thyroid cancer and prostate cancer also have the same disease genes (BRAF), genes enriched in the "Prostate cancer pathway" might also be related to thyroid cancer. "Viral carcinogenesis" pathway and "Epstein-Barr virus infection" pathway are virus-related pathways which are responsible for various cancers, including thyroid cancer [182, 183]. Melanoma is a severe tumor which has been proved to be related with thyroid cancer [184]. "Hippo signaling pathway" regulates organ size and tissue homeostasis. Its fundamental importance make its malfunction leading to many cancers, such as breast cancer [185]. Although thyroid cancer has not been proved to be related with "Hippo signaling pathway", its relationship with breast cancer make us believe that the genes enriched by "Hippo signaling pathway" might also be related to thyroid cancer. Similar to prostate cancer, lung cancer and thyroid cancer have two identical disease genes (STRN, KRAS), which make the genes enriched by "Non-small cell lung cancer" pathway have possibilities to be related with thyroid cancer. Likewise, candidate TC disease genes reported by Endeavor are only enriched by one pathway: "Colorectal cancer". The P-value of the analysis is 9.97×10^{-1}, which is still much larger than the average P-values of dgSeq's results.

According to the pathways in Table 3.3, candidate AD disease genes reported by dgSeq are enriched in seven AD-related pathways: "NOD-like receptor signaling pathway", "Neurotrophin signaling pathway", "GnRH signaling pathway", "Herpes simplex infection", "cAMP signaling pathway", "Inflammatory mediator regulation of TRP channels" and "cGMP-PKG signaling pathway". Amongt these seven pathways, "NOD-like receptor signaling pathway" contains NOD-Like receptors which have been demonstrated to be associated with many diseases, including AD [186]. Neurotrophins are small proteins critical for neuronal

Table 3.1: Enriched KEGG pathways of candidate genes in the BC dataset

Enriched KEGG pathway	Number of Genes	P-value
dgSeq		
Cell cycle	15	0
Thyroid hormone signaling pathway	10	9.96×10^{-3}
TGF-beta signaling pathway	9	7.91×10^{-3}
Epstein-Barr virus infection	12	2.2×10^{-2}
Pathways in cancer	18	2.2×10^{-2}
HTLV-I infection	9	3.82×10^{-2}
Longevity regulating pathway	5	5.54×10^{-2}
Breast cancer	11	7.28×10^{-2}
Hepatitis B	10	1.48×10^{-1}
Adherens junction	5	1.15×10^{-1}
Endeavor		
FoxO signaling pathway	27	9.98×10^{-1}

growth, and Neurotrophin signaling via BDNF/TrkB-TK+ has strong connection with AD [187]. "GnRH signaling pathway" is related to AD because GnRH affect Alzheimer's disease through its marker Aβ protein [188]. In addition, "Herpes simplex infection" has been implicated as a main factor in AD [189]. "cGMP-PKG" and "cAMP/PKA" cooperate to control long-term memory which is affected by AD [190, 191]. Finally, TRP channels mediate physiological responses and research showed that analyzing the connections between TRP channels and AD may lead to new drugs [192]. Candidate genes reported by Endeavor are enriched in five AD-related pathways: "Alzheimer's disease", "Prion diseases" [193], "Chemokine signaling pathway" [194], "Phospholipase D signaling pathway" [195] and "VEGF signaling pathway" [196].

Although the number of enriched pathways of Endeavor is less than dgSeq, Endeavor performs better than dgSeq in terms of the P-value. This result may be caused by the following two reasons. First, only 9 AD case subjects are contained in the data sets. Then, some of the disease genes may not contribute to the disease in these subjects, and rewiring information obtained from these subjects are not comprehensive. Disease genes that are not active in the case subjects may be predicted as non-disease genes, which affect the overall performance of dgSeq. Second, unlike cancers, which have been found to be associated with vast amount of rewiring in co-expression networks [197, 198], the expression level of genes in AD is much lower than that of cancers, making dgSeq hard to capture valuable rewiring information from PCC. One possible solution is to replace PCC with mutual information (MI) when computing the dependence between two genes as MI can measure linear and nonlinear dependence at the same time, while PCC only measures linear dependence. Thus, in the future we may use MI instead of PCC to improve the performance of dgSeq.

Table 3.2: Enriched KEGG pathways of candidate genes in the TC dataset

Enriched KEGG pathway	Number of Genes	P-value
dgSeq		
FoxO signaling pathway	12	0
PI3K-Akt signaling pathway	23	0
Prostate cancer	13	0
Pathways in cancer	21	1.92×10^{-3}
Viral carcinogenesis	23	5.65×10^{-3}
Epstein-Barr virus infection	21	1.09×10^{-2}
Melanoma	9	1.3×10^{-2}
Hippo signaling pathway	10	1.36×10^{-2}
Oocyte meiosis	14	1.57×10^{-3}
Non-small cell lung cancer	9	1.69×10^{-2}
Endeavor		
Colorectal cancer	17	9.97×10^{-1}

3.3.4 Top 10 unknown genes

Table 3.4 lists the top 10 unknown genes in the ranked lists of BC and TC, respectively. We search these top 10 unknown genes online, and find that most of them have been studied as disease genes in previous research. For those not verified as disease genes, we leave their functions blank. From Table 3.4, we can see that 8 out of 10 genes obtained from the BC data set, 6 out of 10 genes obtained from the TC data set and 6 out of 10 genes obtained from AD data set are potential disease genes which have been studied in literature. This analysis reveals that the results of our algorithm are in concert with other existing studies, suggesting that dgSeq is a valuable computational method for discovering new disease genes.

3.4 Conclusion

In this study, we have presented a disease gene prediction method which combines PPI network, clinical RNA-seq data and OMIM data. The method first constructs a differential network based on rewiring information computed from case and control clinical RNA-Seq data. A DGN is constructed based on OMIM database. Then, the set of non-disease genes is selected from the DGN and the differential network according to the shortest path theory. Finally, features of these non-disease genes and known disease genes are extracted from the PPI network and the differential network, and used to train a logistic classifier, which is then employed to predict disease genes.

Evaluations conducted on data sets of two cancers and Alzheimer's disease reveal that our algorithm

is overall more effective than previous methods. Further analysis on the top predicted disease genes have also proved that dgSeq is powerful for predicting new disease genes. In the future, we would integrate more omics data into the disease gene prediction method and improve the performance of dgSeq in predicting new disease genes. We can also replace PCC with MI and extend the strategy for capturing network rewiring information in different types of biomolecular networks to enhance dgSeq's performance in various types of diseases.

Acknowledgements

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC), China Scholarship Council (CSC) and by the National Natural Science Foundation of China under Grant No. 61571052 and No. 61772552.

Table 3.3: Enriched KEGG pathways of candidate genes in the AD dataset

Enriched KEGG pathway	Number of Genes	P-value
dgSeq		
NOD-like receptor signaling pathway	7	1.21×10^{-2}
Neurotrophin signaling pathway	6	4.82×10^{-2}
HTLV-I infection	8	1.27×10^{-1}
Influenza A	5	1.33×10^{-1}
GnRH signaling pathway	6	3.05×10^{-1}
Herpes simplex infection	6	3.41×10^{-1}
cAMP signaling pathway	9	3.92×10^{-1}
Inflammatory mediator regulation of TRP channels	6	3.98×10^{-1}
Epstein-Barr virus infection	11	4.41×10^{-1}
cGMP-PKG signaling pathway	6	4.43×10^{-1}
Endeavor		
Th1 and Th2 cell differentiation	11	1.96×10^{-3}
Alzheimer's disease	11	2.09×10^{-3}
Prion diseases	5	1.91×10^{-2}
Influenza A	7	1.98×10^{-2}
Chemokine signaling pathway	10	2.49×10^{-2}
Phospholipase D signaling pathway	6	2.66×10^{-2}
T cell receptor signaling pathway	12	2.68×10^{-2}
Leishmaniasis	5	3.37×10^{-2}
VEGF signaling pathway	10	3.57×10^{-2}
Endocrine resistance	19	3.91×10^{-2}

Table 3.4: Top 10 unknown genes

Gene Name	Function	Reference
BC		
UBB	Potential disease gene	[199]
SKP2	Potential Oncogene	[200]
KAT5		
HDAC1	Potential disease gene	[178]
RARA	Potential therapeutic target	[201]
HDAC2	Potential disease gene	[178]
HDAC3	Potential disease gene	[178]
CDK8	Potential Biomarkers	[202]
MED1	Potential therapeutic target	[203]
SMARCC1		
TC		
HSP90AA1		
XPO1	Potential disease gene	[204]
YWHAB		
MDM2	Oncogene	[205]
MAX		
PPP2CA	Disease gene for many cancer	[206]
EGFR	Potential marker	[207]
GRB2	Potential disease gene	[208]
RB1	Potential disease gene	[209]
UBE2I		
AD		
RNF32		
MAST1	Potential disease gene	[210]
CSNK1A1		
HSPA5	Poteintial target	[211]
PPP5C	Potential disease gene	[212]
PPP1CA	Potential disease gene	[212]
CAMK2A	Disease gene	[213]
RBBP4	Potential disease gene	[214]
ATP5A1		
H2AFX		

4

Ensemble disease gene prediction by clinical sample-based networks

Prepared as: Ping Luo, Li-Ping Tian, Bo-lin Chen, Qianghua Xiao, and Fang-Xiang Wu. Ensemble disease gene prediction by clinical sample-based networks. BMC Bioinformatics, accepted, 2019. PL conducted the bioinformatics analysis, and FXW supervised the study. PL and FXW wrote the manuscript. All authors read, revised and approved the final version of the manuscript.

In addition to non-disease gene selection, another issue that limits the accuracy of computational prediction is the quality of the PPI networks. Since PPI is dynamic and tissue-specific, directly using static PPI networks might affect the performance of the algorithms. In this chapter, sample-based networks constructed based on the clinical gene expression data are proposed. These networks consist of those significant genes associated with the disease under consideration, which are more valuable than the original static networks. Meanwhile, an ensemble strategy is used to guarantee that all the disease genes could be predicted. This chapter fulfills Objective 3 of this thesis.

Abstract

Disease gene prediction is a critical and challenging task. Many computational methods have been developed to predict disease genes, and protein-protein interaction (PPI) network is widely used to predict disease genes. However, existing methods commonly use a universal static PPI network, which ignore the fact that PPIs are dynamic, and PPIs in various patients should also be different. To address these issues, we develop an ensemble algorithm to predict disease genes from clinical sample-based networks (EdgCSN). The algorithm first constructs single sample-based networks for each case sample of the disease under study. Then, these single sample-based networks are merged to several fused networks based on the clustering results of the samples. After that, logistic models are trained with centrality features extracted from the fused networks, and an ensemble strategy is used to predict the final probability of each gene being disease-associated. EdgCSN is evaluated on breast cancer (BC), thyroid cancer (TC) and Alzheimer's disease (AD) and obtains AUC values of 0.970, 0.971 and 0.966, respectively, which are much better than the competing algorithms. Subsequent *de novo* validations also demonstrate the ability of EdgCSN in predicting new disease genes.

4.1 Background

Disease gene prediction is a critical yet challenging task. It helps us understand the mechanisms of diseases, find therapeutic targets, and develop novel treatment strategies [215]. During the past decades, disease gene prediction has gained great development. Many computational algorithms have been developed to predict disease genes so that the cost and time for in-depth validation could be maximal reduced.

Among the various types of data that have been used to predict disease genes, protein-protein interactions (PPIs) are the most widely used evidence. On the one hand, interacting proteins (genes) usually have similar functions, which means algorithms can predict new disease genes based on their relationships with known disease genes in the PPI network. On the other hand, due to the network property of PPIs, most network analysis algorithms can be used to predict disease genes from PPI networks. For example, earlier methods, such as RWR, performed the random walk on PPI networks to predict disease genes [18]. Gillis et al. used degree centralities to rank all the genes [216].

However, PPIs are dynamic during the life time of cells, and not all PPIs exist in all the tissues. Static PPI networks downloaded from online databases contain lots of false positives which limit the performance of the methods that directly use them [217]. Thus, many studies integrate static PPI networks with disease-related data, such as GWAS and gene expression data, to improve the prediction accuracy [218, 58, 219]. This leads to two types of approaches. The first type of approach weights PPI networks with disease-related data, and predicts candidate genes from the weighted networks. For instance, Wang et al. searched dense modules from a PPI network weighted by gene expression and GWAS data [58]. Our previous study trained a regression model with features extracted from a PPI network weighted by differential co-expression [158]. The second type of approach constructs heterogeneous networks and combines them with PPI networks to enhance the prediction. For example, Chen *et al.* combined gene co-expression networks and pathway coexist networks with PPI networks to predict disease genes [117, 118]. Singh-Blom *et al.* trained a biased SVM with features extracted from phenotype-phenotype networks and PPI networks [220] to predict disease genes. Despite their success, the discussed approaches still use PPI networks with false positive interactions, which contain inaccurate topological structures. PPI networks downloaded from different databases might affect the prediction results.

To solve these issues, in our previous study, gene expression data of clinical samples have been used to construct sample-specific PPI networks [43]. Each single sample-based network only contains the significant PPIs associated with the disease under consideration, which reduces the false positive interactions. A network that fuses all the single sample-based networks was used to predict the disease-associated genes, so that disease genes that function in different patients could all be identified. In this study, to further extend our research, an ensemble algorithm that predicts disease genes from clinical sample-based networks (EdgCSN) is proposed. Meanwhile, Katz centrality is used instead of edge clustering coefficient to better extract local structural information from the sample-based networks.

4.2 Methods

Fig. 4.1 depicts the work flow of EdgCSN which is explained as follows. (a)-(b). A single sample-based network is constructed for each case sample by combining clinical samples and the universal static PPI network. (c). The case samples are clustered into a few groups and single sample-based networks of the samples in the same group are fused to one network. (d). A logistic model is trained by the centrality features extracted from each fused network, and the probability of each gene being disease-associated is predicted. (e). The maximum probability of a gene calculated from all the logistic models is regarded as its probability of being disease-associated. In the following subsections, details of the five steps in Fig. 4.1 are first discussed. Then, the data sources and evaluation metrics are explained.

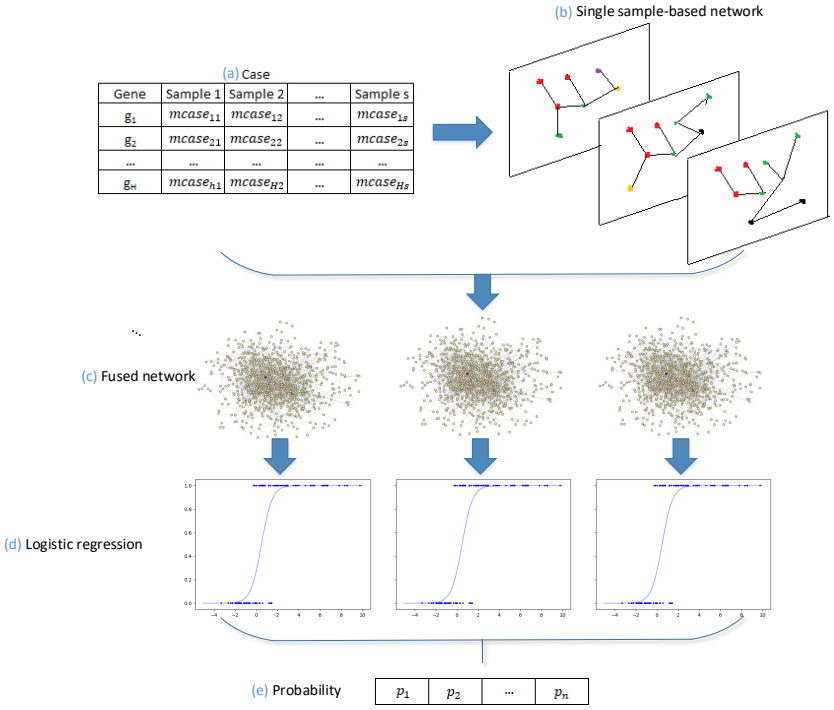

Figure 4.1: Work flow of the algorithm.

4.2.1 Sample-based networks

To obtain the most informative PPIs and remove the false positive ones, sample-based networks are used in this study instead of the universal static PPI networks. In addition, since the real caustic genes of different patients may not be the same, case samples are divided into different clusters so that patients with distinct conditions are analyzed separately. Specifically, three steps are performed to obtain the sample-based networks.

1. A single sample-based network is constructed for each case sample;

2. Case samples are classified into different clusters;

3. Networks of the samples in the same cluster are fused together.

For the first step, we assume that a PPI exists in a single sample-based network N_s only if the two interacted proteins are both activated in sample s. Concretely, a gene i in a case sample s is considered being activated if

$$\text{mcase}[i, s] \geq \lambda * mean(\text{mcntl}[i]) \tag{4.1}$$

where mcase$[i, s]$ is the expression value of gene i in sample s, and $mean$(mcntl$[i]$) is the mean expression value of gene i over all control samples. To construct N_s, every edge (i, j) in the static PPI network is validated and only the one with both i and j being activated is added to N_s. Then, S single sample-based networks are constructed for the S case samples.

For the second step, hierarchical clustering is used to classify case samples into different clusters. Given two samples s_1 and s_2, their pairwise distance is calculated by

$$dist(s_1, s_2) = 1 - \frac{(\mathbf{s}_1 - \bar{\mathbf{s}}_1) \cdot (\mathbf{s}_2 - \bar{\mathbf{s}}_2)}{\|\mathbf{s}_1 - \bar{\mathbf{s}}_1\|_2 \|\mathbf{s}_2 - \bar{\mathbf{s}}_2\|_2} \tag{4.2}$$

where $\mathbf{s}_1$ ($\mathbf{s}_2$) is a vector of expression values of genes in sample s_1 (s_2), and $\bar{\mathbf{s}}_1$ ($\bar{\mathbf{s}}_2$) is the corresponding average expression value. During the bottom-up process, the distance between two newly formed clusters u and v is computed as follows

$$Distance(u, v) = \max_{p \in u, q \in v} (dist(p, q)) \tag{4.3}$$

which is the maximum distance between samples in u and v. Let $dmax$ denote the maximum distance among clusters, $0.7 * dmax$ is used as the threshold to select clusters from the resulted dendrogram.

For the third step, assuming all the S samples are classified into l clusters and the t-th cluster contains S_t samples, we have $S = \sum_{t=1}^{l} S_t$. The objective is to fuse the networks of the samples in the same cluster into one network. Although many network fusion methods have been published [221], most of them cannot efficiently fuse complex PPI networks, especially when the number of networks to be fused is more than $1,000$. Thus, we propose a simple strategy which uses a threshold ϵ to determine whether an edge exists in the fused networks. An edge (i, j) is considered as significant only if it appears in at least ϵ single sample-based networks. Precisely, given a cluster with S_t samples, let f_{ij} be the number of times edge (i, j) appears in the S_t single sample-based networks. When $f_{ij} < \epsilon$, (i, j) is not included in the fused network, and when $f_{ij} \geq \epsilon$, (i, j) is in the fused network. Finally, l fused networks are obtained for the l clusters, respectively.

4.2.2 Model design

Given a biomolecular network, if disease genes are labeled as 1 and non-disease genes are labeled as 0, the disease gene prediction problem can then be formulated as a network labeling problem [46]. Let $\mathbf{x} =$

$(x_1, x_2, \ldots, x_H)$ denote a set of binary labels of all the H genes in the biomolecular network. $\mathbf{x}$ is known as the configuration of the network, and the set X of all possible configurations is a random field. Based on our previous studies [47, 118, 158], a generalized model was proposed in [43] which predicted the probability of a gene i being labeled as 1 by

$$P(x_i = 1|x_{[-i]}, \theta) = \frac{\exp(\theta\phi_i)}{1 + \exp(\theta\phi_i)} \tag{4.4}$$

where θ is a parameter vector and ϕ_i is the feature vector of gene i extracted from the biomolecular network labeled by a prior configuration $\mathbf{x}$.

In [43], ϕ_i is a 7-dimensional feature vector which consists of a dummy feature (1) and three pairs of 0-1 centrality features: 0-1 degree centrality, 0-1 closeness centrality and 0-1 edge clustering coefficient. These three 0-1 centrality indices have shown their ability in characterizing discriminative features for classifying disease and non-disease genes. However, edge clustering coefficient can only capture the structural information between genes and their direct neighbors, and the relations between genes and their k-th order ($k \geq 2$) neighbors cannot be obtained. Since proteins usually form a complex or functional module to achieve a specific function [217], the k-th order neighbors should also be considered when the local structural information is extracted. Previous study also showed that the indirect neighbors were useful for disease gene prediction [48]. Thus, we replace edge clustering coefficient by Katz centrality in this study to leverage the local structure information between nodes and their higher order neighbors.

Given a labeled network $N = (V, E)$, V is the set of nodes and E is the set of edges, the 0-1 degree centrality denoted by C_{i0}^d and C_{i1}^d are defined as follows

$$C_{i0}^d = \sum_{(i,j)\in E} (1 - x_j), \quad C_{i1}^d = \sum_{(i,j)\in E} x_j \tag{4.5}$$

The 0-1 closeness centrality denoted by C_{i0}^c and C_{i1}^c are defined as

$$\begin{aligned} C_{i0}^c &= \frac{1}{n_0 - 1} \sum_{j\in V, j\neq i} \frac{1}{dsp(i,j)}(1 - x_j), \\ C_{i1}^c &= \frac{1}{n_1 - 1} \sum_{j\in V, j\neq i} \frac{1}{dsp(i,j)} x_j \end{aligned} \tag{4.6}$$

where $dsp(i, j)$ is the length of the shortest path between node i and j, n_0 and n_1 are the number of nodes labeled as 0 and 1, respectively

Katz centrality measures the relative influence of a node in the network [222]. It is defined by

$$C_i = \sum_{k=0}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji} \tag{4.7}$$

where A is the adjacency matrix of the network, k is the length of the path between i and j, α is a damping factor penalizes the impact node j on i. The longer the path, the smaller the impact node j is on i.

When α is properly chosen, Eq. (4.7) will converge. However, when Katz centrality is used in this study, we care more about the information conveyed by paths with short distance (less than 5). Study in link

prediction also showed that $k = 3$ or $k = 4$ can yield satisfactory performance [223]. Thus, α and k are chosen by grid search without the proof of convergence.

In previous studies, Katz centrality calculated from heterogeneous networks had been used to prioritize disease genes [220]. However, results of directly using Katz centrality were not better than existing methods, such as RWR [18]. To make Katz centrality suitable for disease gene prediction, we define the 0-1 Katz centrality as follows:

$$\begin{aligned} C_{i0} &= \sum_{k=0}^{\infty}\sum_{j=1}^{n}\alpha^k(A^k)_{ji}(1-x_j), \\ C_{i1} &= \sum_{k=0}^{\infty}\sum_{j=1}^{n}\alpha^k(A^k)_{ji}x_j \end{aligned} \tag{4.8}$$

Similar to 0-1 degree and 0-1 closeness centrality, the 0-1 Katz centrality measures the importance of a gene among disease genes and non-disease genes, respectively, which is more appropriate for disease gene prediction. The new feature vector of each gene is then defined as

$$\phi_i = (1, C_{i0}^d, C_{i1}^d, C_{i0}^c, C_{i1}^c, C_{i0}, C_{i1}) \tag{4.9}$$

4.2.3 Network labeling and benchmark selection

As discussed in the previous section, biomolecular networks are needed to be labeled by a prior configuration so that disease genes can be predicted. In this study, we use the l fused networks to predict disease genes, which means the known disease genes in these networks are labeled as 1 while other genes are labeled as 0. Then, the feature vectors of all genes can be extracted by Eq. (4.9).

In addition, to train the logistic models used for prediction, we also need a set of non-disease genes, which are used as negative instances. Unfortunately, no databases contain non-disease genes. Therefore, our previous strategy proposed in [78] is used to select the non-disease genes used in the training.

In [78], a disease gene network (DGN) was constructed with the disease-gene association data downloaded from OMIM [80]. In the DGN, each node is either a disease or a disease-associated gene. Diseases are connected with their associated genes, and two diseases are connected if they share one or more associated genes. Thus, diseases that are close to each other in the DGN have more chances to share similar disease genes, which means they are more likely to have similar mechanisms. If the length of the shortest path between two diseases is larger than a threshold η, they might not have similar mechanisms, and the disease genes of one disease could be regarded as non-disease genes of the other disease. With this strategy, a group of non-disease genes are obtained for the disease under study, and only non-disease genes that exist in all the l fused networks are selected. $\eta = 5$ is chosen based on our previous experience.

Assuming m disease genes are known to be associated with the disease under study, we randomly select m genes from the set of non-disease genes, and these $2m$ genes form a set of gold standard genes. This process is performed 50 times and finally we obtain 50 sets of gold standard genes and regarded them as benchmarks.

4.2.4 Ensemble prediction

Given m disease genes and m non-disease genes, features of these genes extracted from the l fused networks are used to train l logistic models, respectively. Equation (4.4) is then used to predict the probability of each gene being disease-associated in each fused network.

For each gene, $l^{'}(1 \leq l^{'} \leq l)$ probabilities are calculated. Considering that the caustic genes of different samples might be different, the obtained probabilities only reveal the potential of the gene being disease-associated in the corresponding clusters. Thus, for each gene, the ensemble strategy chooses the maximum value of the $l^{'}$ probabilities as its probability of being disease-associated.

4.2.5 Datasets

In this study, datasets of breast cancer (BC), thyroid cancer (TC) and Alzheimer's disease (AD) are used to evaluate the algorithm. The BC-associated genes and TC-associated genes are obtained from the Cancer Gene Census category (http://cancer. sanger.ac.uk/census#) [87]. In total, 35 BC-associated genes and 33 TC-associated genes are used as the benchmarks. The AD-associated genes are obtained from MalaCards: The human disease database (http: //www.malacards.org/). The database contains 182 potential AD associated genes ranked by their probability of being AD-associated in descending order. 39 of the first 50 genes exist in the static PPI network are used as benchmarks.

The gene expression data of BC and TC are downloaded from NCI Genomic Data Commons (GDC) [104], which measures the data by RNA-Seq. We download the data normalized by FPKM (Fragments Per Kilobase Million) and transform them to TPM (Transcripts Per Kilobase Million) by the strategy proposed in [224]. The expression data of Alzheimer's disease (AD) are downloaded from Gene Expression Omnibus (GSE53697) [168], which are also measured by RNA-seq. The data normalized by RPKM (Reads Per Kilobase Million) are downloaded and transformed to TPM with the same strategy used for the data downloaded from GDC. TPM is chosen because it facilitates the comparison of the proportion of reads that are mapped to a gene in each sample and is usually better than FPKM and RPKM in cross-sample comparison, which helps us properly cluster all the samples. In total, the dataset of BC contains 1102 case samples and 113 control samples; the dataset of TC contains 502 case samples and 58 control samples; the dataset of AD contains 9 case samples and 8 control samples.

After downloading the gene expression data, four steps are performed to control the genes used in our study. (1). TPM values less than 1 are replaced by 0 because of the unreliability. (2). $\log_2(\text{TPM}+1)$ is used instead of the original TPM values. (3). Genes expressed in less than 10% of samples (case and control) are removed. (4). Genes not existing in the PPI network are removed. In total, 14436 genes, 13959 genes and 13370 genes are left for BC dataset, TC dataset and AD dataset, respectively.

The static PPI network is downloaded from the InWeb_InBioMap database (version 2016_09_12) [170]. The database consists of more than 600,000 protein interactions collected from eight source databases, which

insures that valuable protein interactions are not missed during the construction of the sample-based PPI networks. In this study, the proteins in the PPI network are mapped to their corresponding genes to form a gene-gene interaction network. In the paper, the term "PPI network" is used to represent the gene-gene interaction network because of simplicity.

4.2.6 Evaluation metrics

In this study, a disease gene is regarded as positive while a non-disease gene is regarded as negative. Given a threshold Γ, a gene i with a probability $p_i \geq \Gamma$ is predicted as positive, and otherwise it is predicted as negative. For all genes in the benchmark, the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are defined as follows

1. TP: a disease gene is predicted as a disease gene

2. FP: a non-disease gene is predicted as a disease gene

3. TN: a non-disease gene is predicted as a non-disease gene

4. FN: a disease gene is predicted as a non-disease gene

Then, we can calculate the true positive rate (TPR) and the false positive rate (FPR) of the prediction results by the following equations

$$TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{TN + FP} \tag{4.10}$$

To evaluate the algorithm, the receiver operating characteristic (ROC) curve is created by plotting the TPR against FPR with various Γ. The area under the ROC curve (AUC) is also used to evaluate the overall performance of the algorithm.

Since the number of genes used as benchmark is small, leave-one-out cross validation (LOOCV) is performed to calculate the probabilities of genes in the benchmark being disease-associated. With the 50 sets of gold standard genes, LOOCV is performed 50 times. In each round, the probabilities of the $2m$ genes being disease-associated are calculated, as well as the AUC value. The average AUC value is then used to evaluate the algorithm.

In addition, *de novo* validation is performed by ranking all the unknown genes in descending order by their average probabilities calculated by the models trained with the 50 sets of gold standard genes. The top 10 unknown genes are analyzed from published literature to illustrate the ability of EdgCSN in predicting new disease genes.

4.3 Results

4.3.1 Clustering

Figs. 4.2, 4.3 and 4.4 show the dendrograms of the hierarchical clustering. BC and TC samples are divided to three clusters and AD samples are divided to two clusters. Thus, three fused networks are constructed for BC and TC, respectively, and two fused networks are constructed for AD.

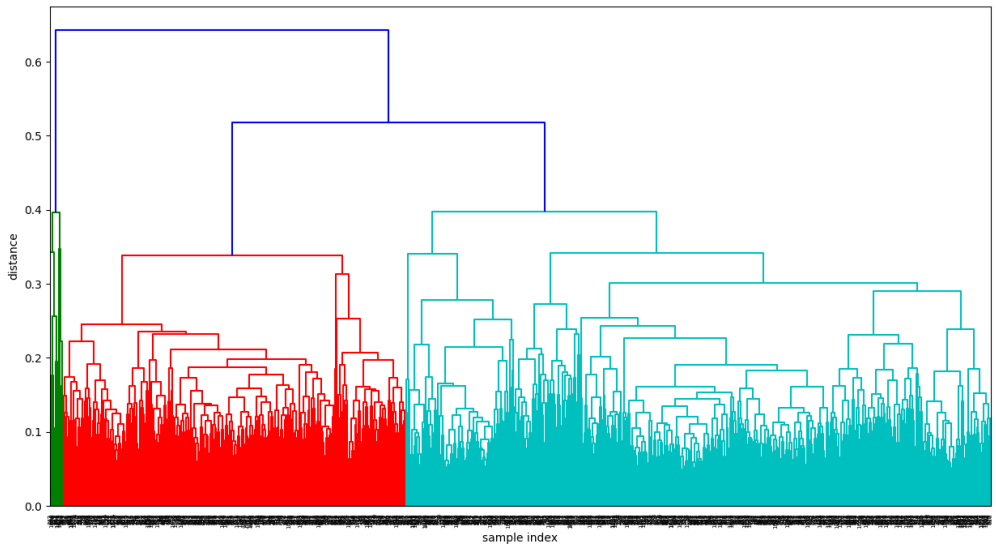

Figure 4.2: Hierarchical clustering dendrogram for BC.

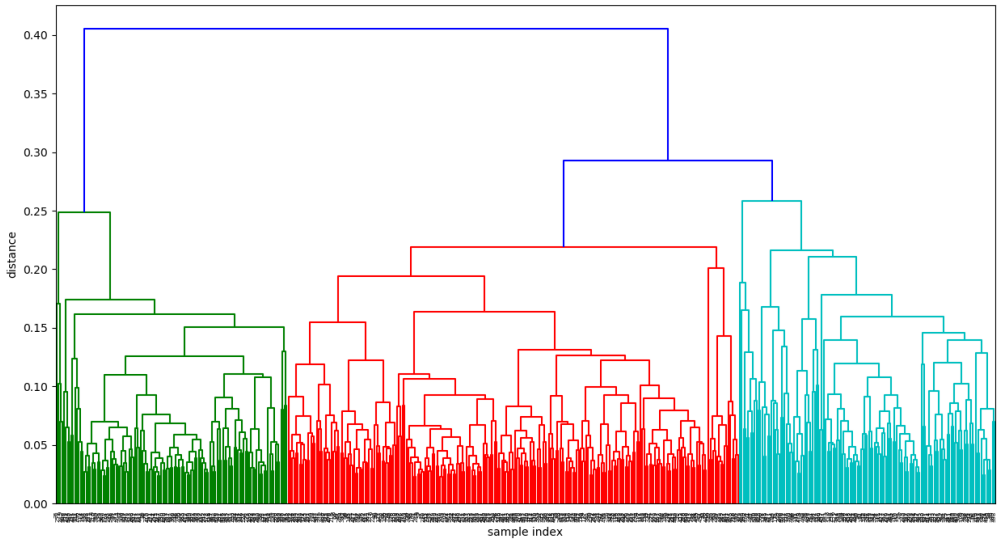

Figure 4.3: Hierarchical clustering dendrogram for TC.

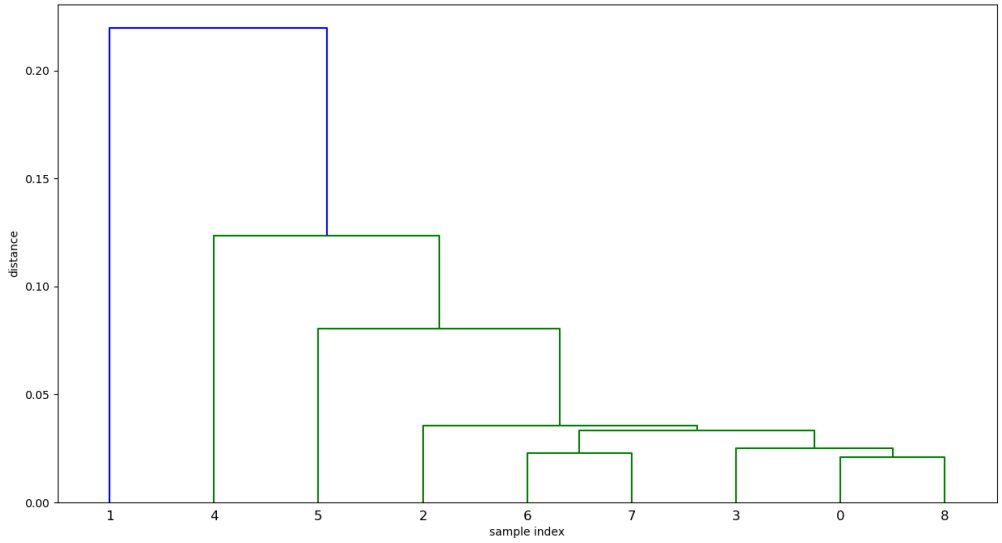

Figure 4.4: Hierarchical clustering dendrogram for AD.

4.3.2 Sensitivity analysis

The performance of our algorithm is affected by four hyperparameters: λ, ϵ, α and k. The first two control the resulted fused networks. Based on our previous study, edges that exist in more than three networks were significant [43]. Thus, $\epsilon = 3$ is empirically chosen in this study. As for λ, since the RNA-seq data are normalized by TPM rather than DESeq2 [169], λ is searched from a new set $\{1.0, 1.1, 1.2, 1.3, 1.5\}$, which is different from the one obtained in our previous study. The other two hyperparameters control the information extracted by Katz centrality. To obtain the appropriate hyperparameters, α is searched from $\{0.1, 0.2\}$, and k is searched from $\{1, 2, 3, 4\}$, respectively.

Tables 4.1, 4.2 and 4.3 show the results of the grid search for BC, TC and AD, respectively. EdgCSN performs best for BC when $\lambda = 1.1, \alpha = 0.2, k = 2$ with an AUC $= 0.970$; for TC when $\lambda = 1.11, \alpha = 0.1, k = 2$ with an AUC $= 0.971$; for AD when $\lambda = 1.0, \alpha = 0.2, k = 2$ with an AUC $= 0.966$. '-' denotes that more than 10% known disease genes are not contained in the fused networks constructed by the corresponding hyperparameters.

All the three experiments obtain their best AUC values when $k = 2$, and a smaller or higher k would significantly affect the performance of the algorithm. These results indicate that local structural information contained within the second order neighborhood is valuable for disease gene prediction. Other disease gene prediction algorithms that use topological structure of biomolecular networks could also further include these information to improve their prediction.

Table 4.1: Sensitivity analysis. The resulted AUC values obtained with different combinations of hyperparameters for BC.

		k			
λ	α	1	2	3	4
1.0	0.1	0.867	0.961	0.873	0.878
1.0	0.2	0.869	0.966	0.889	0.870
1.1	0.1	0.883	0.967	0.890	0.903
1.1	0.2	0.881	**0.970**	0.909	0.896
1.2	0.1	0.845	0.957	0.877	0.898
1.2	0.2	0.846	0.958	0.892	0.894
1.3	0.1	0.787	0.938	0.819	0.842
1.3	0.2	0.787	0.940	0.841	0.842
1.5	0.1	0.777	0.938	0.813	0.775
1.5	0.2	0.777	0.938	0.786	0.816

4.3.3 Comparison

EdgCSN is compared with three algorithms: the Re-balanced algorithm of Chen *et al.* [118], the AIDG algorithm of Tang *et al.* [79], and our previous algorithm dgCSN [43]. Re-balanced method combined multiple types of biomolecular networks to predict cancer-related genes, and AIDG used sub-cellular localization to purify universal PPI networks. These algorithms have been shown better than many classical methods, such as the RWR method [18], the DIR method [225] and the ToppNet [226].

The resulted ROC curves for BC, TC, and AD are depicted in Figs. 4.5, 4.6, 4.7, respectively. The AUC values of EdgCSN for BC, TC and AD are 0.970, 0.971 and 0.966, respectively, which are much better than those of the competing algorithms. For BC, our EdgCSN is 7% more accurate than the competing algorithms, and for TC and AD, EdgCSN is 20% more accurate than the other three algorithms.

4.3.4 *De novo* validation

To validate the performance of EdgCSN in predicting new disease genes, unknown genes are ranked in descending order by their average probabilities of being disease-associated predicted by the 50 sets of genes in the benchmark. The top 10 predictions are further searched in existing literature to find out if they are associated with the disease under study.

Table 4.4 shows the top 10 predictions of the three diseases. Functions of the genes that have not been studied in existing literature are left blank. Most of the genes have been analyzed as disease-associated in existing studies, especially for BC, where all the 10 genes have been studied in the existing literature. For TC, although only 5 of the 10 genes have been studied, 3 of the 5 genes that have not been studied

Table 4.2: Sensitivity analysis. The resulted AUC values obtained with different combinations of hyperparameters for TC.

		k			
λ	α	1	2	3	4
1.0	0.1	0.716	0.966	0.839	0.790
1.0	0.2	0.713	0.967	0.795	0.802
1.1	0.1	0.729	**0.971**	0.800	0.746
1.1	0.2	0.728	0.969	0.744	0.779
1.2	0.1	0.809	0.954	0.748	0.776
1.2	0.2	0.808	0.953	0.652	0.792
1.3	0.1	0.621	0.962	0.779	0.786
1.3	0.2	0.620	0.960	0.662	0.794
1.5	0.1	0.412	0.965	0.809	0.720
1.5	0.2	0.411	0.963	0.645	0.679

('CEP72', 'CEP131' and 'GPR83') belong to the Centrosomal Protein family and G Protein-coupled Receptor respectively. Many proteins belong to these families are closely related to cancers [227], which means 'CEP72', 'CEP131' and 'GPR83' might be predicted as being TC-associated in the future.

4.4 Discussion

Many algorithms have been proposed to predict disease genes, and most of them rely on PPI networks to achieve the prediction. However, PPI is dynamic and tissue-specific, static PPI networks downloaded from online databases contain many false positives, and directly using them would limit the accuracy of disease gene prediction. Moreover, for patients with a specific disease, their disease states might be driven by different subset of disease genes, and analyzing their data together might affect the identification of rarely mutated disease genes .

Therefore, in this study, an ensemble algorithm is proposed to predict disease genes from clinical sample-based networks. The algorithm first constructs single sample-based networks by combining clinical samples and a universal static PPI network. A group of networks which contain disease-related PPIs are generated. Then, case samples are divided into different clusters and networks belong to the samples in the same cluster are merged together. This step allows patients with similar causing genes to be analyzed together. After that, 0-1 centrality features extracted from the fused networks are used to train the logistic models that calculate the probability of each genes being disease-associated in each fused network. Finally, an ensemble strategy is performed by choosing the maximum probability obtained from different fused networks as the final probability of a gene being disease-associated.

Table 4.3: Sensitivity analysis. The resulted AUC values obtained with different combinations of hyperparameters for AD.

		k			
λ	α	1	2	3	4
1.0	0.1	0.808	0.964	0.809	0.763
1.0	0.2	0.809	**0.966**	0.764	0.705
1.1	0.1	0.665	0.956	0.757	0.685
1.1	0.2	0.665	0.957	0.596	0.636
1.2	0.1	0.564	0.938	0.809	0.605
1.2	0.2	0.563	0.939	0.608	0.596
1.3	0.1	0.508	0.914	0.810	0.674
1.3	0.2	0.508	0.914	0.608	0.614

In the experiments conducted on BC, TC and AD, our EdgCSN is much better than the competing algorithms in terms of AUC scores. Further analysis of the top 10 unknown genes also illustrate that EdgCSN is capable of predicting novel disease genes. Our study has provided insight into how clustering patient samples might improve the prediction of disease genes.

4.5 Conclusions

Our EdgCSN use ensemble learning to predict disease genes from clustered sample-based networks. In the future, the strategies used for clustering can be further improved. For instance, Eq. (4.2) uses the expression data of all the genes to calculate the pairwise distances, and the results might be dominated by non-disease genes. We could reduce the number of genes used for clustering and choose those differentially expressed genes or marker genes that are associated with a specific subtype. These subsets of genes should improve the clustering results as well as the final prediction.

4.6 Acknowledgements

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC), China Scholarship Council (CSC) and by the National Natural Science Foundation of China under Grant No. 61772552 and No. 61602386, and the Natural Science Foundation of Shaanxi Province under Grant No. 2017JQ6008.

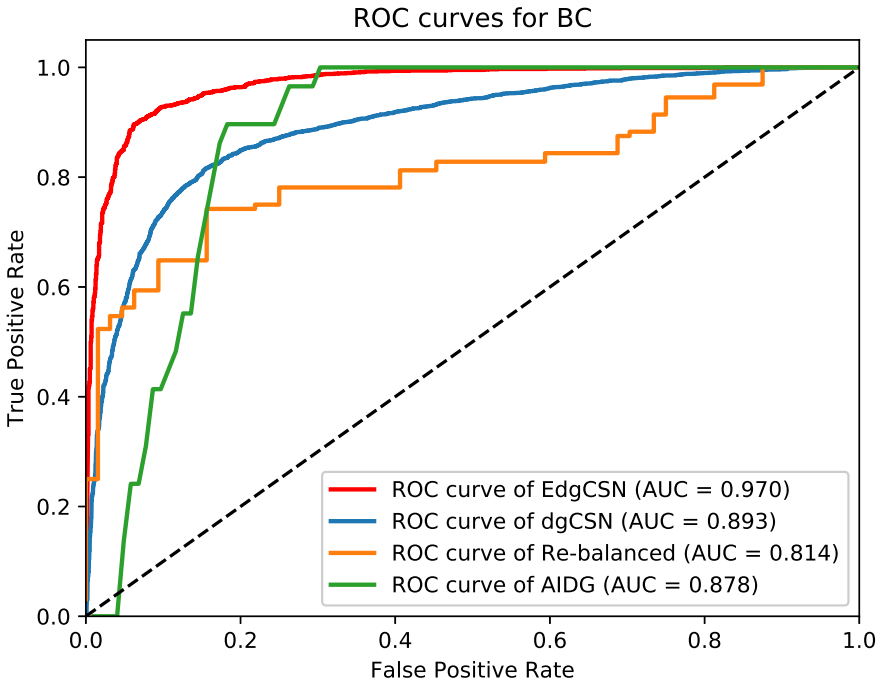

Figure 4.5: ROC curves for BC.

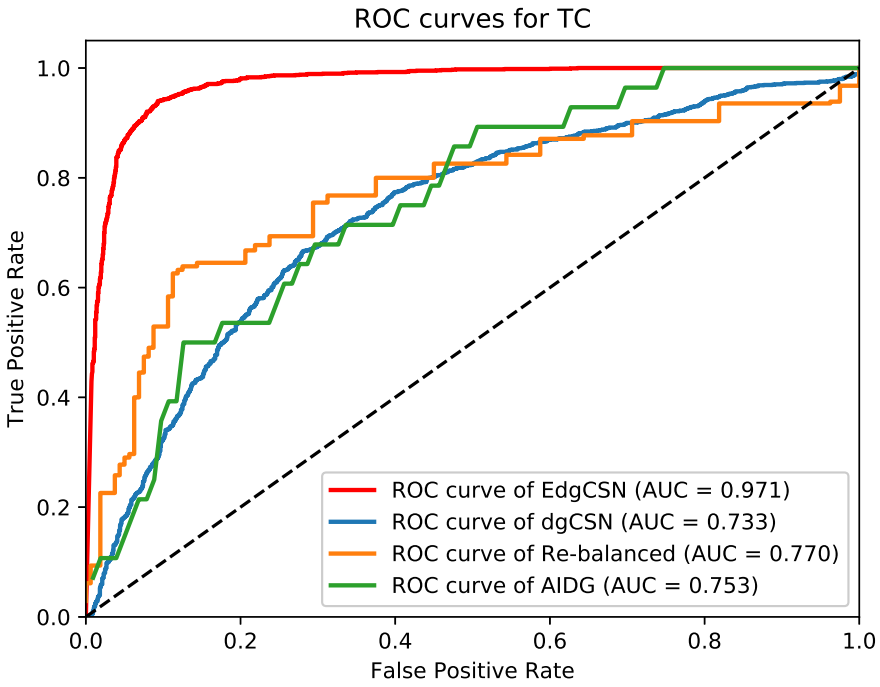

Figure 4.6: ROC curves for TC.

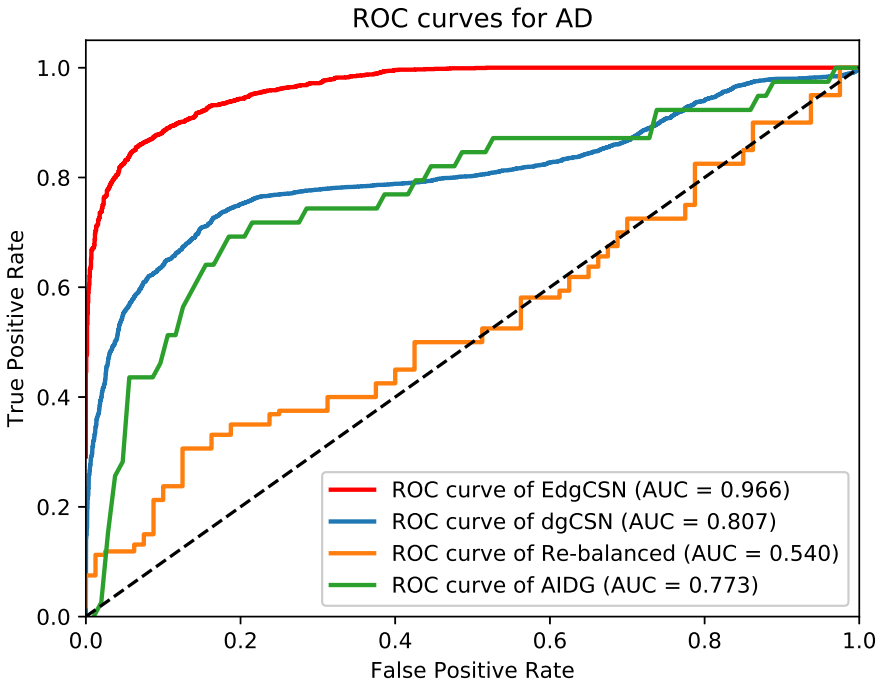

Figure 4.7: ROC curves for AD.

Table 4.4: Top 10 unknown genes

Gene Name	Function	Reference
BC		
CREBBP	Potential disease gene	[228]
NBN	Potential disease gene	[229]
PARP1	Potential biomarker	[230, 231]
NCOR2	Potential biomarker	[232]
RXRA	Potential therapeutic target	[233]
WRN	Potential disease gene	[234]
EXO1	Potential disease gene	[235]
NCOA3	Potential disease gene	[236]
RMI2	Potential disease gene	[237]
TOPBP1	Potential therapeutic target	[238]
TC		
HRAS	Potential disease gene	[239]
HAUS7		
CEP72		
GTF2I	Potential disease gene	[240]
BCLAF1	Potential disease gene	[241]
HAUS3		
FGFR1OP	Potential disease gene	[242, 243]
CEP131		
GPR83		
ALMS1	Potential disease gene	[244]
AD		
MAP2	Potential disease gene	[245]
DPYSL3		
ERRFI1	Potential disease gene	[246]
DAB2	Potential disease gene	[247]
AMPH	Potential disease gene	[248]
SYN1	Potential disease gene	[249]
SYT9	Potential disease gene	[250]
AXIN1		
PRNP	Potential disease gene	[251]
AAK1	Potential disease gene	[252]

5

Enhancing the prediction of disease-gene associations with multimodal deep learning

Prepared as: Ping Luo, Yuanyuan Li, Li-Ping Tian, and Fang-Xiang Wu. Enhancing the prediction of disease-gene associations with multimodal deep learning. Bioinformatics, in press, 2019. PL, YL, LPT and FXW discussed about the methods. PL implemented the algorithm, designed and performed the experiments. FXW supervised this study. PL and FXW wrote the manuscript. All authors read, revised and approved the final version of the manuscript.

The previous two chapters have proposed strategies to solve the two problems that exist in developing machine learning-based methods. In this chapter, a deep learning-based method is proposed to fuse multiple types of data. Specifically, multimodal DBN is used to combine raw features learned from PPI network and GO data. The model can learn both linear and nonlinear relationships within different types of data, and extract cross-modality features which are more valuable for disease-gene prediction. This chapter fulfills Objective 4 of this thesis.

Abstract

Motivation: Computationally predicting disease genes helps scientists optimize the in-depth experimental validation and accelerates the identification of real disease-associated genes. Modern high-throughput technologies have generated a vast amount of omics data, and integrating them is expected to improve the accuracy of computational prediction. As an integrative model, multimodal deep belief net (DBN) can capture cross-modality features from heterogeneous datasets to model a complex system. Studies have shown its power in image classification and tumor subtype prediction. However, multimodal DBN has not been used in predicting disease-gene associations.

Results: In this study, we propose a method to predict disease-gene associations by multimodal DBN (dg-MDL). Specifically, latent representations of protein-protein interaction networks and gene ontology terms are first learned by two DBNs independently. Then, a joint DBN is used to learn cross-modality representations from the two sub-models by taking the concatenation of their obtained latent representations as the multimodal input. Finally, disease-gene associations are predicted with the learned cross-modality

representations. The proposed method is compared with two state-of-the-art algorithms in terms of 5-fold cross-validation on a set of curated disease-gene associations. dgMDL achieves an AUC of 0.969 which is superior to the competing algorithms. Further analysis of the top-10 unknown disease-gene pairs also demonstrates the ability of dgMDL in predicting new disease-gene associations. The Supplementary data are available at `https://doi.org/10.1093/bioinformatics/btz155`.

5.1 Introduction

Ever since the discovery of the first disease gene in 1949 [2], thousands of genes have been identified to be disease-associated. Identifying disease-gene associations helps us decipher the mechanisms of diseases, find diagnostic markers and therapeutic targets, which further leads to new treatment strategies and drugs. High-throughput technologies usually predict a few hundreds of candidate genes, and validating all these candidates requires an extensive amount of cost and time. Thus, a commonly used approach is to first computationally predict/prioritize candidate genes associated with the diseases under consideration, then experimentally validate a subgroup of candidates based on the results of computational prediction so that the yield of the experiments can be greatly improved.

Currently, various types of data have been used to predict disease-gene associations, and protein-protein interaction (PPI) networks are the most widely used evidence. Previous algorithms tried to predict disease-gene associations by directly using the topological structure of PPI networks [18, 22]. However, universal PPI networks downloaded from online databases contain lots of false positives, and only using them cannot further improve the prediction accuracy. Thus, researchers tend to combine more types of data with PPI networks to predict disease-gene associations.

One strategy is to combine PPI networks with clinical data which capture the difference between patients (case) and normal people (control). This resulted in a group of GWAS-based methods [218, 253, 254] and gene expression (GE)-based methods [54, 58, 78]. GWAS-based methods first map the single-nucleotide polymorphisms and their corresponding P-values to the human genome. Then, the mapped P-values are combined with PPI networks and other evidence to predict disease-gene associations. GE-based methods analyze the expression level of each gene in case and control subjects and identify differentially expressed genes or rewired co-expressions, which are then combined with PPI networks to predict disease-gene associations.

Although algorithms based on clinical data are more accurate than the previous methods, their performance is still limited by the amount and quality of the data. For diseases not well studied, the amount of available data limits the performance of the algorithms. For other diseases like cancers, although projects such as TCGA [255] have generated a large amount of omics data, not all disease-gene associations can be successfully identified because of the following reasons. The tumorigenesis of most patients is associated with several frequently mutated genes, and clinical data-based algorithms can easily identify the associations between cancers and these genes. However, for other less mutated genes, the overwhelming abundance of

frequently mutated genes would make the computational model believe that the less mutated ones are not disease-associated. As a result, algorithms based on clinical data tend to generate results that do not include less mutated genes. Therefore, the key issue now is to identify those critical but less mutated genes [256].

To address the problems of existing methods, a generic model which combines different types of non-clinical data would be more valuable. On the one hand, this model predicts disease-gene associations using evidence that can reveal the intrinsic properties of diseases and genes, such as disease similarities, gene similarities, PPI networks, gene ontology (GO) terms, protein domains etc. Integrating such multiple types of information could complement the shortage of previous PPI-based algorithms. On the other hand, since clinical data is not used in the prediction, the results are less likely to be affected by the frequency of the disease-associated mutations.

Methods based on matrix factorization (MF) are generic models and can leverage the disease similarities and gene similarities to predict disease-gene associations [6, 257, 258]. However, MF-based algorithms usually need too much time to converge and most of them can only use limited types of data, which limits their performance. Since studies have shown that integrating multiple types of data could enhance the prediction of disease-gene associations [46, 48, 118, 70], a good generic model should be able to integrate multiple types of data with a unified framework so that the advantages of multi-view data can be properly utilized.

Currently, many algorithms have been proposed to integrate multi-view biological data. Among these algorithms, multimodal deep learning reveals great potential in capturing cross-modality features to uncover the mechanisms of biological systems [259]. Deep learning algorithms, such as deep belief net (DBN) [260], have been applied to drug repositioning [261] and cancer subtype prediction [262]. Although these studies have shown the abilities of deep learning in analyzing biological systems, no studies have used deep learning in disease gene prediction because of two reasons. First, if deep learning is used to predict the disease genes of a specific disease, the number of known disease genes would be too small to train a deep model. Second, if DBN is used to extract features from the biological data, Gaussian units have to be used in the visible layer so that the model can accept real-valued data. The corresponding restricted Boltzmann machine (RBM) in the DBN is a Gaussian-Binary RBM (GBRBM), which is hard to train [263, 264]. More attention is needed to choose appropriate hyperparameters.

To solve the above issues, in this study, instead of predicting associated genes for a specific disease, we build a generic model to predict disease-gene associations for all known diseases. This strategy greatly increases the number of positive samples, making it possible to train a deep network. Meanwhile, the Gaussian visible layer is used to learn latent features from original real-valued features. To leverage the advantage of deep learning in data fusion and improve prediction accuracy, multimodal DBN is used to fuse different modalities and obtain joint representations. Specifically, two sub-models are first trained based on PPI networks and GO terms, respectively. Then, a joint DBN is used to combine the two sub-models to learn cross-modality representations.

In the rest of the paper, Section 5.2 describes the details of the algorithm and the experiments. Section

5.3 discusses the results of the evaluation. Section 5.4 draws some conclusions.

5.2 Materials and methods

5.2.1 RBM

RBM is a graphical model which consists of a visible layer and a hidden layer. In this model, every unit in one layer is connected to every unit in another layer, and there are no within layer connections. Fig. 5.1 shows an example RBM with four visible units and five hidden units. RBM can characterize the distribution of input data, and the learned probabilities of hidden units can be used as features to characterize raw data. When data is binary, the corresponding RBM is a Binary-Binary RBM (BBRBM), and the probability distribution is defined by the following likelihood function:

$$P(v) = \sum_h P(v,h) = \sum_h \frac{e^{-E(v,h)}}{Z} \tag{5.1}$$

where $E(v,h) = -b^T v - c^T h - h^T W v$ is the energy function. $Z = \sum_v \sum_h e^{-E(v,h)}$ is known as the partition function. W is the weight matrix that connects visible and hidden units. b and c are the biases of visible and hidden layers, respectively.

RBM can be learned by using the stochastic gradient descent (SGD) on the empirical negative log-likelihood of training data, which results in the following gradients for a BBRBM [265]

$$-\frac{\partial \log p(v)}{\partial W_{ij}} = E_v[p(h_i|v) \cdot v_j] - v_j^{(i)} \cdot sigm(W_i \cdot v^{(i)} + c_i) \tag{5.2}$$

$$-\frac{\partial \log p(v)}{\partial c_i} = E_v[p(h_i|v)] - sigm(W_i \cdot v^{(i)}) \tag{5.3}$$

$$-\frac{\partial \log p(v)}{\partial b_j} = E_v[p(v_j|h)] - v_j^{(i)} \tag{5.4}$$

where $sigm$ denotes the sigmoid function $sigm(x) = 1/(1+\exp(-x))$. These equations compute the expectations over all possible configurations of input data, which is difficult. A feasible solution is to estimate the expectations with a fixed number of samples. Several sampling techniques have been developed to calculate the gradients [266, 267, 268]. In this study, we choose the contrast divergence (CD) because of its simplicity. Details of the algorithms can be found in [266].

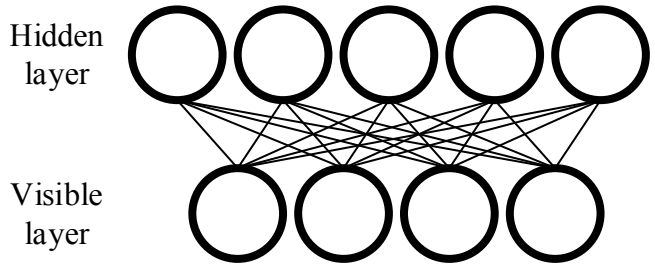

Figure 5.1: Schematic example of an RBM.

For GBRBM, the energy function becomes:

$$E(v,h) = \sum_{i \in vis} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in hid} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} \tag{5.5}$$

where σ_i is the standard deviation of the Gaussian noise for visible unit i. Since learning the variance is difficult with CD, we use the same strategy as in [269] which normalizes each feature to have zero mean and unit variance. The variance in Eq. (5.5) is then set to 1, and the resulted learning procedures remain the same except for that when CD is performed, the reconstructed value of a Gaussian visible unit changes from $sigm(W^T h + b)$ to $(W^T h + b)$.

5.2.2 Multimodal DBN

Multimodal DBN was originally proposed to learn joint representations from image and text data [5]. In this study, multimodal DBN is used to learn cross-modality features with raw features extracted based on PPI networks and GO terms. Fig. 5.2 gives a schematic multimodal DBN for predicting disease genes. The left and right subnetworks denote two DBNs which model PPI-based features and GO-based features, respectively. The top network is a DBN that models the joint distribution and a sigmoid activation function as the output layer for decision making.

According to [270], each DBN in Fig. 5.2 can be regarded as a stack of RBMs and trained in a greedy layer-wise manner. Starting from the visible layer, every pair of adjacent layers form an RBM, which can be trained by the approach discussed in Section 5.2.1. In this study, the visible layers in the two sub-models use Gaussian units, and the corresponding RBMs formed by v_p, h_p^1 and v_g, h_g^1 are GBRBM. All the rest RBMs formed by adjacent hidden layers are BBRBM. Once an RBM is trained, the activation probabilities of its hidden layer are used as the input data to train the next RBM, and the DBN can be trained in this layer-wise manner. After training the two sub-DBNs, their output (hidden probabilities of the top layers) are concatenated, and the resulted representations are used as the input to train the joint DBN.

The whole model is trained in an unsupervised way, and the resulted multimodal DBN can be further analyzed by many approaches. In this study, we add an output layer with a sigmoid function to predict the probability of each disease-gene pair being associated using the cross-modality representations learned by the joint DBN.

5.2.3 Raw feature extraction

The input data of the multimodal DBN is the raw features of disease-gene pairs. These features are extracted from disease similarity networks and gene similarity networks. Specifically, for each sub-model, a disease similarity network and a gene similarity network are first constructed. Then, features of diseases and genes are extracted from their corresponding similarity networks, respectively, by node2vec [62], which is an algorithm that can learn features for nodes in networks. This algorithm performs random walk on

a network and captures both local topological information and global structural equivalent properties to extract features. We choose node2vec because it can generate independent features which are suitable for the input of the multimodal DBN. In addition, experiments have shown that features obtained by node2vec are more informative than those of other algorithms in classification task [62].

The following two sections discuss the strategies used to construct similarity networks based on PPI networks and GO terms.

Similarity networks in PPI-based sub-model

In the PPI-based model, gene-gene interaction network mapped from the PPI network is regarded as the gene similarity network. This strategy is chosen because interacting proteins may have similar functions and protein interactions can reflect the functional similarities between the corresponding genes. Meanwhile, instead of constructing another gene similarity network, the topological structure of the PPI network is also valuable when extracting features with node2vec.

The disease similarity network N_d^{PPI} is constructed according to the disease module theory. A disease module in an interactome is a subgraph consisting of genes associated with the disease [271]. Let $M_1 = (V_1, E_1)$ denote the disease module of disease d_1 in the interactome (gene-gene interaction network). $V = \{g_{11}, g_{12}, \ldots, g_{1n_1}\}$ is a set of disease genes associated with d_1, and E_1 is a set consisting of their interactions. $M_2 = (V_2, E_2)$ is another disease module with similar definition. According to [272], the similarity between two disease modules M_1 and M_2 can be calculated as follows:

$$sim_{ppi}(M_1, M_2) = \frac{\sum_{1 \leq s \leq n_1} F_{M_2}(g_{1s}) + \sum_{1 \leq t \leq n_2} F_{M_1}(g_{2t})}{n_1 + n_2} \tag{5.6}$$

where $F_M(g) = avg(\sum_{g_i \in M} sim(g, g_i))$ measures the relations between gene g and disease module M, which is the sum of the transformed similarities between g and the genes in disease module M. Given two genes g_1 and g_2 in the PPI network, their transformed similarity is calculated by

$$sim(g_1, g_2) = \begin{cases} 1, & g_1 = g_2 \\ e^{-sp(g_1, g_2)}, & \text{otherwise} \end{cases}$$

where $sp(g_1, g_2)$ is the length of the shortest path between g_1 and g_2 in the PPI network. The larger the transformed similarity, the closer the relationship between g_1 and g_2.

After calculating the similarities between modules M_1 and M_2, the similarities between diseases d_1 and d_2 can be obtained by normalizing the module similarities as follows:

$$SIM_{ppi}^{d}(d_1, d_2) = \frac{2 * sim_{ppi}(M_1, M_2)}{sim_{ppi}(M_1, M_1) + sim_{ppi}(M_2, M_2)} \tag{5.7}$$

Finally, N_d^{PPI} is constructed by k nearest neighbors (KNN) algorithm [163]. Specifically, edges are added to N_d^{PPI} for each disease and its top-k most similar diseases obtained by Eq. (5.7). These edges are weighted by the similarity scores of their two connected diseases. In this study, $k = 10$ is chosen according to our previous experience [78].

Similarity networks in GO-based sub-model

Similar to the construction of N_d^{PPI}, the GO-based similarity networks are also built by KNN algorithm, except that the similarities between diseases and genes are calculated based on GO instead of PPI network.

GO database provides a set of vocabularies to describe gene products based on their functions in the cell. Three types of ontologies are defined in GO: biological process, cellular component and molecular function. All the GO terms exist as directed acyclic graphs (DAGs) where nodes represent terms while edges represent semantic relations. In this study, we use the approach developed by [130] to measure the semantic similarities of GO terms and genes.

Let $DAG_A = (T_A, E_A)$ represent GO term A, where T_A contains all the successor GO terms of A in the DAG, and E_A contains the semantic relations between A and other terms in T_A. Each term t in T_A has an S-value related to A:

$$\begin{cases} S_A(t) = 1, if\ t = A \\ S_A(t) = max\{w_e * S_A(t^{'})|t^{'} \in children\ of\ t\}, otherwise \end{cases} \tag{5.8}$$

where w_e is the weight of the edge (semantic relations) in the DAG. Two types of semantic relations are used in the DAG: 'is_a' and 'part_of', and the corresponding w_e is set as 0.8 and 0.6, respectively, as recommended in [130].

Given $DAG_A = (T_A, E_A)$ and $DAG_B = (T_B, E_B)$ for two GO terms A and B, the semantic similarity of these two terms is computed by:

$$SGO(A, B) = \frac{\sum_{t \in T_A \cap T_B}(S_A(t) + S_B(t))}{\sum_{t \in T_A} S_A(t) + \sum_{t \in T_B} S_B(t)} \tag{5.9}$$

The semantic similarity of one GO term $t^{'}$ and a set of GO terms $GO = \{t_1, t_2, \ldots, t_l\}$ is defined as:

$$sim_{go}(t^{'}, GO) = \max_{1 \le i \le l}(SGO(t^{'}, t_i)) \tag{5.10}$$

Then, the functional similarity of two genes g_1 and g_2, annotated by GO term set $GO_1 = \{t_{11}, t_{12}, \ldots, t_{1n_1}\}$ and $GO_2 = \{t_{21}, t_{22}, \ldots, t_{2n_2}\}$, is calculated by:

$$SIM_{go}^{g}(g_1, g_2) = \frac{\sum_{1 \le i \le n_1} sim_{go}(t_{1i}, GO_2) + \sum_{1 \le j \le n_2} sim_{go}(t_{2j}, GO_1)}{n_1 + n_2} \tag{5.11}$$

The similarity of two diseases d_1 and d_2, associated with two sets of genes $V_1 = \{g_{11}, g_{12}, \ldots, g_{1n_1}\}$, $V_2 = \{g_{21}, g_{22}, \ldots, g_{2n_2}\}$, is defined as:

$$SIM_{go}^{d}(d_1, d_2) = \frac{\sum_{1 \le i \le n_1} SG(g_{1i}, DG_2) + \sum_{1 \le j \le n_2} SG(g_{2j}, DG_1)}{n_1 + n_2} \tag{5.12}$$

where $SG(g^{'}, DG) = \max_{1 \le i \le l}(SIM_{go}^{g}(g^{'}, g_i))$.

Sub-model input construction

After obtaining the similarity networks, features are extracted by node2vec. Let ϕ_i^p denote the extracted feature vector of disease i, and φ_j^p denote the extracted feature vector of gene j in the PPI-based model. Their concatenation, $\psi_{ij}^p = (\phi_i^p, \varphi_j^p)$, is the feature vector of disease-gene pair (i, j) in the PPI-based model, which is then used as the input of the PPI-based sub-DBN. Similarity, ψ_{ij}^{go} is constructed and used as the input of the GO-based sub-DBN.

5.2.4 Evaluation metrics

The area under Receiver Operating Characteristics (ROC) curve (AUC) is used to evaluate the algorithms. ROC curve plots the true positive rate [TP/(TP+FN)] versus the false positive rate [FP/(FP+TN)] at different thresholds, and a larger AUC score represents better overall performance. In this study, a true positive (TP) is a known disease-gene association (positive sample) predicted as a disease-gene association, while a false positive (FP) is a non- disease-gene association (negative sample) predicted as a disease-gene association. A false negative (FN) is a positive sample predicted as negative while a true negative (TN) is a negative sample predicted as negative.

Considering that negative samples are not included in existing databases, we combine our previous study in [78] and the idea of reliable negatives in [41] to collect a subset of unknown samples as potential negative samples (PN). Taking the PPI-based model as an example, let ψ_{avg}^p denote the average feature vector of all positive samples. For each unknown sample u, we calculate the Euclidean distance d_u^p between u and ψ_{avg}^p. The average distance is then denoted as d_{avg}^p. If $d_u^p > d_{avg}^p$, sample u is considered as a reliable negative sample. With this approach, two sets of reliable negative samples are collected from the PPI-based model and GO-based model, respectively. disease-gene pairs in the intersection of the two sets are regarded as PN. In our experiment, 4432 samples (the same as the number of positive samples) are randomly selected from PN as negative samples and the dataset contains 8864 samples in total. This random selection is performed three times to generate three sets of data.

The proposed method is evaluated in three steps. First, the whole dataset is randomly split into three subsets: training set (80%), validation set (10%) and testing set (10%). The optimized hyperparameters are determined based on the average AUC obtained from 10 randomly split validation sets. The average AUC obtained from testing sets with the optimized hyperparameters is used to evaluate the overall performance of the model. Second, dgMDL is compared with two newly developed algorithms: PBCF [257] and Know-GENE [63] in 5-fold cross-validation. PBCF is an MF-based algorithm and Know-GENE uses the boosted regression to predict disease-gene associations. Both of them are generic models which use similar types of data as dgMDL does. For each set of data, the cross-validation is run for five times to remove the influence of the random splitting. Associations left for testing are not used to calculate disease similarities. Third, unknown disease-gene pairs are ranked by their probabilities of being associated predicted by dgMDL. The

top-10 pairs and top-10 unknown lung cancer-related genes are further studied in existing literature to evaluate the performance of dgMDL in predicting new disease-gene associations.

5.2.5 Hyperparameters

In this study, several hyperparameters affect the accuracy of the prediction. For the multimodal DBN, the numbers of hidden layers and the number of nodes in each hidden layer determine the architecture of the model. In our experiments, the model is found to be insensitive to the number of hidden nodes. Thus, we set the number of hidden nodes in the sub-modal and the joint-model to 256 and 512, respectively. In addition, since the performance of the model becomes stable when the numbers of hidden layers are larger than 2, we set the numbers of hidden layers to be 3 in both the sub-DBN and the joint-DBN.

Another three hyperparameters [learning rate (lr), batch size (bs) and number of epochs (ne)] determine whether the model is well trained. For lr, 0.01 is recommended for training BBRBM in [273]. In our study, we find that 0.01 is small enough to train the BBRBM. A smaller or adaptive lr barely changes the prediction accuracy. Thus, lr used for training BBRBM is set to 0.01. Meanwhile, it is recommended that lr used for training GBRBM should be one or two orders of magnitude smaller than that for BBRBM. Thus, we search lr of the GBRBM from {0.001, 0.0005, 0.0002, 0.0001}. For bs, a recommended value is usually equal to the number of the classes, and it would be better if each mini-batch contains at least one sample from each class. Considering that we only have two classes in this study and using a bs equals to two can hardly guarantee the recommendation, bs is searched from {2, 4, 8, 10}. For ne, we fix it to 30 because the performance of dgMDL becomes stable after being trained for 30 epochs. Table S1 in the Supplementary gives the average AUC obtained from the validation sets with different combinations of lr and bs. The optimized lr for the GBRBM and bs are 0.0005 and 4, respectively.

For node2vec, the hyperparameters include: dimension of features (d); return parameter (p); in-out parameter (q); number of walks (r); length of walk (l) and context size (k). The corresponding default values recommended in [62] are 128, 1, 1, 10, 80 and 10, respectively. Although these hyperparameters should be changed for networks with different numbers of nodes and edges, searching all of them with brute force would be time-consuming. In our study, we do test different combinations of d, p, q and l, but the results are all worse than the ones obtained with the default values. To determine the real optimized hyperparameters used in node2vec, one might need a large amount of time on the grid search, which is not the key issue of the deep learning model. Therefore, the default values of node2vec are used in our study.

5.2.6 Data sources

The disease-gene association data are downloaded from the Online Mendelian Inheritance in Man (OMIM) database [274]. The latest Morbid Map at OMIM contains nearly seventy-five hundred entries sorted alphabetically by disease names, thirty-nine hundred genes and more than sixty-one hundred diseases. Each entry represents an association between a gene and a disease. Different entries are labelled with different tags

['(3)', '[ ]' and '?'] indicating their reliabilities. To get the most reliable entries, in this study three steps are performed to preprocess the originally downloaded dataset. The first two steps are similar to the approach used in [154]. From the website of OMIM, diseases with tag '(3)' indicate that the molecular basis of these diseases is known, which means the associations are reliable. Entries with '[ ]' represent abnormal laboratory test values while entries with '?' represent provisional disease-gene associations. At the first step, entries with the tag '(3)' are selected while others are abandoned. At the second step, we classify these disease entries into distinct diseases by merging disease subtypes based on their given disorder names. For instance, 14 entries of '46XX sex reversal' are merged into disease '46XX sex reversal', and the 9 complementary terms of 'Renal cell carcinoma' are merged into 'Renal cell carcinoma'. During the classification, string match is first used to classify adjacent entries, and then the classified results are manually verified. At the third step, 475 diseases are removed because each of them is associated with only one gene which is not associated with any other diseases. As a result, we obtain the final dataset consisting of 4432 associations between 1154 diseases and 2909 genes. All these disease-gene associations are included in Supplementary Table S2.

The PPI network is obtained from the InWeb_InBioMap database (version 2016_09_12) [170], which consists of more than 600,000 interactions collected from eight databases. The proteins in the network are mapped to their corresponding genes to form a gene-gene interaction network. In total, there are 17429 genes in the network. GO data are downloaded from the GO database [125, 126]. For genes that have no ontology information, the values of their features in the GO-based model are all 0.

5.3 Results

5.3.1 Overall performance

Fig. 5.3 shows the average AUC obtained with the hidden representatives learned from different layers of the model. The raw feature vectors and the activation probabilities learned in each hidden layer are used to predict disease-gene associations in the testing set. The blue bars and purple bars show the AUC scores obtained from the PPI-based DBN and GO-based DBN, respectively. AUC scores obtained from the joint DBN are shown by the red bars. Clearly, the accuracy of the prediction improves when the model is continuously trained, which shows that the multimodal DBN successfully learns valuable information in different stages of the training and improves the prediction of disease-gene associations.

5.3.2 Comparison with other algorithms

Fig. 5.4 shows the ROC curves of dgMDL (red), Know-GENE (blue) and PCFM (orange) obtained with 5-fold cross-validation, respectively. dgMDL achieves an AUC of 0.969 which is the best among three competing algorithms. The AUC of Know-GENE is 0.941, which is slightly worse than that of dgMDL. PCFM ranks the 3rd with an AUC of 0.791.

5.3.3 Prediction of new disease-gene associations

To further evaluate dgMDL, we rank the unknown disease-gene pairs according to their probabilities of being associated calculated by the model. Since known disease genes are more likely to be associated with other diseases, we rank the unknown pairs of diseases and existing disease genes in this study. Meanwhile, we also rank the unknown pairs by Know-GENE and PCFM for comparison. Table 5.1 lists the top-10 ranked pairs of dgMDL, Know-GENE, and PCFM, respectively. For dgMDL, 8 out of the 10 pairs have been studied in existing literature. While for Know-GENE and PCFM, only 3 of the 10 pairs have been studied.

In addition to the top-10 prediction, we test the ability of dgMDL in predicting new associated genes for a specific disease. Table 5.2 lists the top 10 unknown genes associated with lung cancer. 9 out of 10 pairs have been studied in existing literature. All these results demonstrate that dgMDL is valuable in predicting new disease-gene associations.

5.4 Conclusion

Integrating multiple types of data with machine learning model is a challenging task, especially for predicting disease genes where the number of known associations is limited. In this study, we have proposed a method to predict disease-gene associations with the cross-modality features obtained by multimodal DBN. The deep learning model learns joint representations from raw features extracted from PPI-based similarity networks and GO-based similarity networks. Results show that the proposed method is overall more accurate than the competing algorithms. Further analysis of the top-10 disease-gene pairs and top-10 lung cancer-related genes also reveal the potential of dgMDL in predicting new disease genes. The current model integrates two types of data. It is possible that a gene is not included in any of these data, and its associations cannot be correctly predicted. In the future, more types of data should be integrated by the multimodal DBN, such as disease-disease associations, protein domain and sequence information, to solve this issue and improve the prediction accuracy.

Acknowledgements

The authors thank Dr. Hongyi Zhou for explaining the functional association strength calculated in Know-GENE. This work is supported by the Natural Science and Engineering Research Council of Canada (NSERC); China Scholarship Council (CSC); the National Natural Science Foundation of China [Grant No. 61772552, 61571052]; and the Science Foundation of Wuhan Institute of Technology [Grant No. K201746].

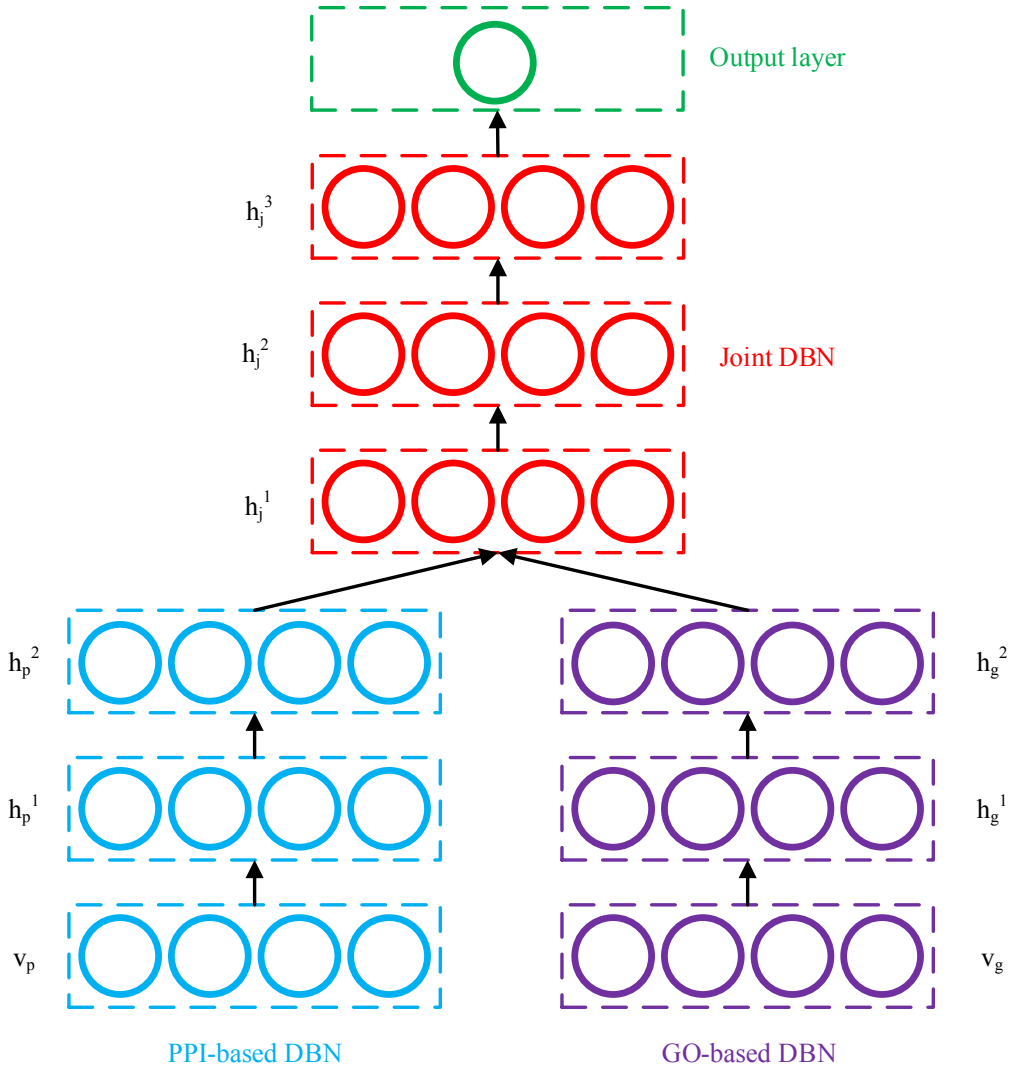

Figure 5.2: Schematic example of a multimodal DBN for disease gene prediction.

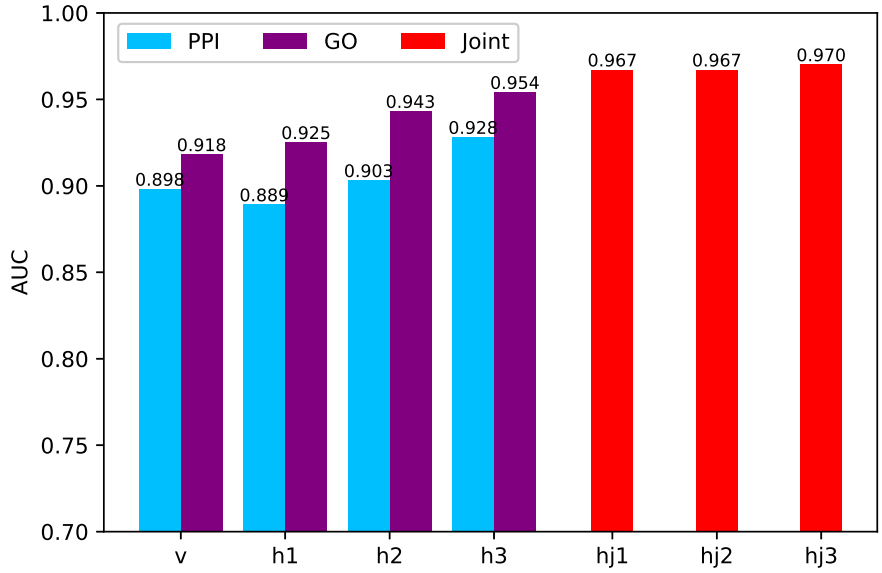

Figure 5.3: AUC of dgMDL in different layers

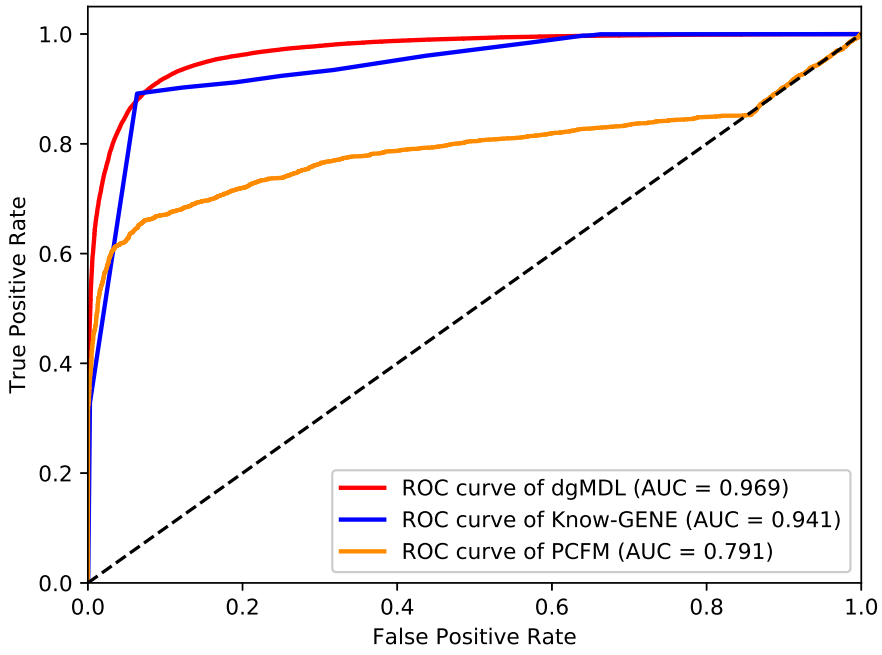

Figure 5.4: ROC curves of the three algorithms

Table 5.1: Top-10 associations predicted by dgMDL, Known-GENE and PCFM

Disease	Gene	Supporting Evidence
dgMDL		
Deafness	PIK3CD	[275]
Deafness	PIK3CA	
Deafness	PIK3R1	[276]
Diabetes	AR	[277]
Deafness	PTPN11	[278]
Diabetes	SMAD4	[279]
Cataract	AR	
Diabetes	GATA3	[280]
Mental retardation	SMAD4	[281]
Deafness	STAT3	[282]
Know-GENE		
Acne inversa familial	NLRP12	
Basal cell nevus syndrome	HGF	
Bladder cancer somatic	PIK3CA	[283]
Bladder cancer somatic	NRAS	
Cardiofaciocutaneous syndrome	EGFR	
Complement factor I deficiency	C3	[284]
LADD syndrome	PIK3CA	
Meckel syndrome	B9D1	[285]
Nevus epidermal somatic	ERBB2	
Nevus epidermal somatic	RET	
PCFM		
Mental retardation	CLCN7	
Mental retardation	PDE3A	
Mental retardation	RBM12	
Mental retardation	BPTF	[286]
Mental retardation	TAP1	
Mental retardation	LAMTOR2	[287]
Mental retardation	DYSF	
Mental retardation	TPRKB	
Mental retardation	HERC1	[288]
Mental retardation	RORC	

Table 5.2: Top-10 susceptible lung cancer-associated genes

Gene	Supporting Evidence
PTPN11	[289]
PIK3R1	[290]
HRAS	[291]
GATA3	[292]
PIK3CD	
JAK2	[293]
STAT3	[294]
C5	[295]
SIK1	[296]
PPM1D	[297]

6

deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks

Prepared as: Ping Luo, Yulian Ding, Xiujuan Lei, and Fang-Xiang Wu. deepdriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. Frontiers in Genetics, 10:13, 2019. PL, YD, XL and FXW discussed about the methods. PL implemented the algorithm, designed and performed the experiments. FXW supervised this study. PL and FXW wrote the manuscript. All authors read, revised and approved the final version of the manuscript.

The previous chapter use multimodal DBN to fuse different types of data, the model would become complex when fusing several types of data. A convolutional layer and pooling layer should help to reduce the dimension of the learned representations. Thus, in this chapter, the CNN model is applied to fuse different types of data and predict disease genes. This chapter fulfills Objective 5 of this thesis.

Abstract

With the advances in high-throughput technologies, millions of somatic mutations have been reported in the past decade. Identifying driver genes with oncogenic mutations from these data is a critical and challenging problem. Many computational methods have been proposed to predict driver genes. Among them, machine learning-based methods usually train a classifier with representations that concatenate various types of features extracted from different kinds of data. Although successful, simply concatenating different types of features may not be the best way to fuse these data. We notice that a few types of data characterize the similarities of genes, to better integrate them with other data and improve the accuracy of driver gene prediction, in this study, a deep learning-based method (deepDriver) is proposed by performing convolution on mutation-based features of genes and their neighbors in the similarity networks. The method allows the convolutional neural network to learn information within mutation data and similarity networks simultaneously, which enhances the prediction of driver genes. deepDriver achieves AUC scores of 0.984 and 0.976 on breast cancer and colorectal cancer, which are superior to the competing algorithms. Further evaluations of the top 10 predictions also demonstrate that deepDriver is valuable for predicting new driver genes.

6.1 Introduction

Cancer is driven by various types of mutations, such as single nucleotide variants (SNVs), insertions or deletions (Indels) and structural variants. Identifying driver genes whose mutations cause cancer could help us decipher the mechanism of cancer, which is beneficial to the development of novel drugs and therapies.

With the advances in next-generation sequencing technologies, massive amounts of cancer genomic data have been published, which elevate the identification of driver genes. Currently, many computational methods have been proposed, and they can be divided into several types. A typical kind of method is those based on the mutation frequency. These methods find "significantly mutated genes" (SMG) whose mutation rates are significantly higher than the background mutation rate and judge them as driver genes. For instance, OncodriveCLUST finds positions with mutation rates higher than the background mutation rate and predicts driver genes from clusters generated based on these seed positions [298]. MutsigCV identifies SMGs by building a patient-specific background mutation model with gene expression data and DNA replication time data [299]. However, due to the heterogeneity of tumors, constructing a reliable background mutation model is difficult [13], which limits the performance of frequency-based methods. Another type of methods predicts driver genes by network analysis. For example, DawnRank predicts driver genes by ranking the genes in a gene interaction network (GIN) with PageRank algorithm [300]. SCS uses network control strategy to find driver mutations that can drive the regulation network from the normal state to disease states [301]. Considering that GINs are downloaded from online databases, such as BioGrid [302] and HPRD [89], which contain many false positives, network-based methods need more accurate GIN to improve their prediction accuracy.

As the increasing number of experimentally validated driver genes, researchers start to use machine learning algorithms to predict new driver genes. These methods usually train a classifier with features characterizing the functional impact of mutations. For instance, CHASM trains a random forest classifier with 86 predictive features [303]. CanDrA trains an SVM with 95 features obtained from ten functional impact-based algorithms, such as SIFT [304] and CHASM. Since the number of driver genes is much smaller than that of passenger genes, selecting gold-standard driver genes (positive data) and a set of high-quality nonfunctional passenger genes (negative data) is difficult for machine learning-based methods. However, with reasonable downsampling, these methods can also achieve better performance than other types of algorithms. Tokheim et al. propose a random forest algorithm (known as 20/20+) and compare it with seven classical driver gene prediction algorithms (ActiveDriver [305], MuSiC [306], MutsigCV [299], OncodriveCLUST [298], OncodriveFM [307], OncodriveFML [308] and TUSON [256]) in [144]. Results show that 20/20+ performs best among the eight algorithms, which demonstrate that machine learning models are able to predict driver genes given the limited known driver-disease associations.

At present, most machine learning-based methods use random forest and SVM as the classifier. To improve the prediction accuracy, various kinds of features extracted from different types of data are used

to train the classifier. Despite the increase of the dimensionality, simply concatenating all these features may not be the best approach to integrate different types of data. Considering that several types of data can be used to characterize the similarities of genes, if we construct similarity networks with these data and combine them with other predictive features, the prediction accuracy of the algorithms should be improved compared to that obtained from a simple feature concatenation. Thus, in this study, a deep learning-based method is proposed to predict driver genes by combining similarity networks with features that characterize the functional impact of mutations (deepDriver). Specifically, candidate driver genes are predicted by a convolutional neural network (CNN) trained with mutation-based feature matrix constructed based on the topological structure of a similarity network. The algorithm leverages the similarity of gene expression patterns and the functional impact of mutations simultaneously, which can better fuse these two types of data and improve the prediction accuracy. To our knowledge, this is the first time that CNN is combined with similarity network to predict driver genes.

In the rest of the paper, Section 6.2 describes the methods and the materials used in the study. Section 6.3 analyzes the results of the evaluation. Section 6.4 draws some conclusions.

6.2 Material and methods

6.2.1 General model

CNN is successful in many areas, such as image classification and speech recognition. The key component of a CNN is the convolutional (CONV) layer, which helps the model to learn local and global structures from the input data. In an image classification problem, these structures include edges, curves, corners, etc. While in a driver gene prediction problem, traditional input data contain distinct features that characterize different properties of genes, which cannot be directly applied to CNN.

We notice that pixels in a small region share the same filters because they have similar grayscale. In a gene similarity network (GSN), genes and their neighbors also have similar properties. If we reconstruct the traditional input data with GSN so that features of similar genes are close to each other, CNN can then be applied to these reconstructed data. Instead of edges and curves learned from the images, topological structures of the similarity networks are learned by CNN with this strategy. In addition, the strategy allows CNN to learn the similarities of genes and the properties of the original input data simultaneously, which can improve the accuracy of driver gene prediction.

Fig. 6.1 depicts a schematic example of a 1-dimensional CNN, which is used in our study. The model consists of five kinds of layers: Input layer, CONV layers, pooling layers, Fully-Connected (FC) layers, and Output layer. Given a feature matrix $\phi_i \in R^{2k \times n_f}$ constructed by the feature vectors of g_i and its k neighbors where n_f is the dimension of the feature vectors of g_i, the output of a CONV layer corresponds to the input

ϕ_i and the filter w_j is calculated as follows

$$A(i,j) = f(w_j \phi_i + b_j) \tag{6.1}$$

where b_j denotes the bias corresponds to w_j, f is an activation function which is ReLU in this study. $w_j \phi_i$ is still the dot product of w_j and ϕ_i except that the calculation is restricted to be local spatially. Each CONV layer is followed by a pooling layer, and the CONV-POOL pattern is repeated for several times. The final structure of the model used for driver gene prediction is determined by grid search, and the results are discussed in Section 6.3.2. The construction of ϕ_i is discussed in the next section.

6.2.2 Network-based convolution

The convolution is performed by combining mutation-based features with gene similarity networks. Many approaches can be used to calculate the similarities of genes. In this study, to characterize the relationships between genes in the disease states, Pearson correlation coefficient (PCC) defined by the following equation is used to calculate the similarities.

$$r(g_i, g_j) = \frac{\sum_{q=1}^{v}(e_{iq} - \bar{e}_i)(e_{jq} - \bar{e}_j)}{\sqrt{\sum_{q=1}^{v}(e_{iq} - \bar{e}_i)^2}\sqrt{\sum_{q=1}^{v}(e_{jq} - \bar{e}_j)^2}} \tag{6.2}$$

where $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{iv})$ denotes the expression values of g_i in v tumor samples, and $\bar{\mathbf{e}}_i$ is the mean of $\mathbf{e}_i$. An undirected network N is constructed by k-nearest neighbors (kNN) algorithm [163] in which every gene is connected to genes that have the k largest PCC scores with itself.

After obtaining N, the construction of ϕ_i used in the convolution is depicted by Fig. 6.2. Assuming we have obtained a feature vector x_i for each gene g_i, and $g_{s1}, g_{s2}, \ldots, g_{sk}$ are the k nearest neighbors of g_i in N, where $pcc(g_i, g_{s1}) > pcc(g_i, g_{s2}) > \cdots > pcc(g_i, g_{sk})$. Feature matrix $\phi_i \in R^{2k \times n_f}$ is built as depicted by the figure. In ϕ_i, features of similar genes are close to each other so that they can share the same filters in the CONV layer.

6.2.3 Mutation-based features

For each gene of a specific disease, twelve features are extracted from the mutation datasets. Table 6.1 lists the names and descriptions of these features. Among them, the first eight ones measure the fraction of a specific type of mutation among all the mutations. The tenth and eleventh feature measure the rate of missense mutations and non-silent mutations to silent mutations, respectively. The last two features measure the positional clustering of different types of mutations and are calculated as follows

$$E_i = \frac{-\sum_i p_j \log_2 p_j}{\log_2 m} \tag{6.3}$$

For the normalized missense entropy, m is the total number of missense mutations of g_i, and $p_j = \kappa_j / m$ where κ_j is the number of missense mutations in the j-th codon. For the normalized mutation entropy,

m is the total number of all types of mutations of g_i. Different mutations are binned together based on their types, except for that missense mutations are binned based on their codon positions, different silent mutations are divided into their own bins. Inactivating mutations (nonsense, translation start site, nonstop, splice site) are grouped into a single bin.

These twelve features have been used in many machining learning-based methods [309, 144]. To demonstrate the superiority of our model, we did not use any other features proposed by specific methods. In addition, during the implementation of the competing methods (SVM, 20/20+), only these twelve features are used to train their models.

6.2.4 Data sources

In this study, deepDriver was evaluated on three types of cancer: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD) and lung adenocarcinoma (LUAD). The mutation data and gene expression data of these three diseases were downloaded from the NCI Genomic Data Commons (GDC) [104]. For the mutation data, quality control was applied by filtering out hypermutated samples ($> 1,000$ intragenic somatic variants) [309]. In total, 228,046, 168,746 and 287,667 somatic variants were obtained for BRCA, COAD, and LUAD, respectively. For gene expression, datasets of 1,102 BRCA, 478 COAD and 551 LUAD primary tumor samples measured by RNA-Seq were downloaded. We chose the data normalized by FPKM and converted the values to TPM by the method proposed in [224]. Three steps were then performed to remove the genes that are barely expressed in tumor samples. First, TPM values less than 1 were considered unreliable and replaced by 0. Second, $\log_2(\text{TPM} + 1)$ was applied to all TPM values. Third, genes expressed in less than 10% of all tumor samples were removed.

Gene ids were standardized to the gene names provided by HUGO Gene Nomenclature Committee (downloaded Aug 1, 2018) [310]. Only genes that have both mutation and expression data are kept. Finally, 13,777 genes for BRCA, 11,282 genes for COAD and 13,731 genes for LUAD passed the quality control.

The driver genes were collected from two sources—the Cancer Gene Census category (CGC) [87] and the genes published in [311]. Genes in CGC were divided into two tiers, and we used genes in Tier 1 as driver genes because strong evidence has proved their oncogenic role in cancer genesis. It is of note that both oncogene and tumor suppressor gene (TSG) are regarded as driver gene in this study. In total, 37 driver genes for BRCA, 42 driver genes for COAD and 12 driver genes for LUAD were collected from CGC. The Bailey et al.ś dataset [311] contains 299 driver genes associated with 33 types of cancer. In total, 29 driver genes for BRCA, 20 driver genes for COAD and 20 driver genes for LUAD were collected. These driver genes as well as a few sets of non-disease genes were regarded as "ground truth" in the evaluation. Details of the non-disease genes are discussed in next section.

To validate the performance of the algorithm, the structure of the model was first determined by the grid search using the driver genes of BRCA and COAD collected from CGC. Then, the optimal model was directly applied to LUAD without fine-tuning the hyperparameters. Similarly, when the model was trained

with the driver genes published in [311], the optimal hyperparameters were used without fine-tuning.

6.2.5 Evaluation metrics

The algorithm was evaluated in two steps. In the first step, deepDriver was compared with 20/20+ and SVM in terms of the AUC (area under the receiver operating characteristic (ROC) curve) scores obtained from 10-fold cross-validation. ROC curve plots the false positive rate (FPR) against the true positive rate (TPR) at different thresholds. FPR and TPR are defined as follows

$$\begin{aligned} FPR &= \frac{FP}{FP + TN} \\ TPR &= \frac{TP}{TP + FN} \end{aligned} \tag{6.4}$$

where TP, FP, TN, and FN are the numbers of true positives, false positives, true negatives, and false negatives, respectively. In this study, a true positive is a driver gene predicted as a driver gene, a false positive is a passenger gene predicted as driver gene, a true negative is a passenger gene predicted as a passenger gene, and a false negative is a driver gene predicted as a passenger gene. The larger the AUC is, the better the performance of the algorithm is.

Since the number of passenger genes is much larger than that of the driver genes, a method is needed to solve the imbalanced issue. Currently, two types of methods can be used to solve the imbalanced problem: data level methods and classifier level methods [312]. In this study, a data level method, downsampling, was used to reduce the size of the passenger genes. Specifically, a subset of passenger genes was randomly selected from all the passengers so that the numbers of positive samples (driver genes) and negative samples (passenger genes) are equal. This approach was run for five times which generated five sets of data. During the cross-validation, for each set of data, all the positive and negative samples were randomly split into ten groups, and the CNN model was validated for ten rounds. In each round, one group of samples were used as the testing data while the rest nine groups of samples were used as the training data.

Additionally, since passenger genes are barely reported in existing literature, in this study, genes that have not been reported as cancer driver genes (unknown genes) were regarded as passenger genes. This strategy was used because of the following two reasons. First, the numbers of the selected passenger genes and the undiscovered driver genes are both much less than that of the unknown genes. Potential driver genes only have a small change to be selected as passenger genes [256]. Second, the final results were obtained by taking the average predictions of the five sets of data. This bagging strategy would improve the stability and accuracy of the results and reduce the impact of a potential driver gene selected as a passenger gene. Finally, the 10-fold cross-validation was run for five times for each dataset to reduce the influence of random shuffling, and the average AUC score was used to evaluate the performance of the algorithms.

In the second step, all the unknown genes were ranked by their probabilities of being driver genes, and the top 10 predictions were searched from the existing literature to check whether our predictions are in concert with existing studies. We also ranked the unknown genes by SVM, 20/20+ and OncodriveCLUST

and compared their results with those of deepDriver in terms of the number of genes having been analyzed in existing literature.

6.2.6 Implementation

The algorithm was implemented using Keras [313] with TensorFlow [314] as the backend engine. We have tested the program on both CPU and GPU versions of TensorFlow and the model can be efficiently trained with or without the help of GPU. A reference implementation is available at `cnnhttps://github.com/luoping1004/deepDriver`.

6.3 Results

6.3.1 Hyperparameters

In this study, the architecture of CNN is determined by the following hyperparameters.

1. The number of the CONV layers (*ncl*)
2. The number of the FC layers (*nfl*)
3. The number of the nodes in the CONV layers (*ncn*)
4. The number of the nodes in the FC layers (*nfn*)

These hyperparameters were determined by grid search, with *ncl* searched from {1,2,3,4}, *nfl* searched from {1,2,3}, *ncn* searched from {12,24,48} and *nfn* searched from {24,48,96}. The optimal values of *ncl*, *nfl*, *ncn* and *nfn* are 2, 1, 24 and 48, respectively. In addition, zero padding was used in the CONV layers except the first one. The size of the filters, the window size of the pooling layers and the stride sizes used in the CONV layers and the pooling layers were all empirically set to 2.

The number of neighbors used by kNN algorithms was also determined by grid search. We searched k from {3,5,7,9,11,13,15}, and finally, $k = 9$ and $k = 7$ were chosen for BRCA and COAD, respectively. In fact, the AUC scores were all above 0.950 when $7 \leq k \leq 15$. Based on our previous study, $k = 7$ is enough to generate high-quality similarity networks [78]. Thus, $k = 7$ was used when the dataset of LUAD was analyzed by our deepDriver. Meanwhile, for other types of cancer not discussed in this study, $k = 7$ is also recommended when the similarity network is constructed.

For 20/20+, a random forest of 200 trees was used based on the suggestions of [144]. For SVM, the model was implemented with a linear kernel and RBF kernel. The penalty parameter C was searched from {0.1, 0.01, 0.001, 1, 10, 100, 1000}, and γ was searched from {1/12, 0.001, 0.0001, 0.00001}. Finally, for BRCA and COAD, SVM performed the best with an RBF kernel, when $C = 1$, $\gamma = 0.0001$; for LUAD, SVM performed the best with an RBF kernel, when $C = 1000$, $\gamma = 0.00001$.

6.3.2 Cross-validation

Fig. 6.3, Fig. 6.4 and Fig. 6.5 show the results of the ROC curves and the corresponding AUC scores of deepDriver, 20/20+ and SVM on BRCA, COAD and LUAD, respectively. According to the figures, deepDriver achieved AUC scores of 0.984, 0.976 and 0.998 on BRCA, COAD and LUAD, respectively, which were at least 15.1% higher than those of the two competing algorithms, especially for COAD and LUAD where the AUC scores of the competing algorithms were less than 0.750.

To further demonstrate that the model was not overfitted, the learning curves were plotted using the datasets of the three types of cancer. For each type of cancer, 80% of the total samples were used as training data while the rest 20% samples were left to test the performance of the model. Fig. 6.6, Fig. 6.7 and Fig. 6.8 show the results of the learning curves. The AUC scores obtained from the testing set improved with the increase of the number of the training samples, which demonstrates that the model is not overfitted. In the meantime, the AUC scores obtained with a small amount of samples also demonstrate that the model is able to produce meaningful results even if the number of the known driver genes is less than 10.

In addition to the driver genes collected from CGC, our deepDriver was also validated using the driver genes published in [311]. As discussed in Section 6.2.4, the optimal hyperparameters obtained from the first set of drivers were directly used to evaluate the model. Fig. 6.9 depicts the resulted ROC curves. Our deepDriver obtained AUC scores of 0.985, 0.941 and 0.970 on BRCA, COAD, and LUAD, respectively.

6.3.3 *De novo* study

To further evaluate the performance of deepDriver, the unknown genes were ranked by their probabilities of being driver genes predicted by the model. Similar to the cross-validation, 5 sets of data were used to train the model and the unknown genes were ranked by the average probabilities. Meanwhile, we also ranked the unknown genes using the three competing algorithms and compared their results with those of deepDriver in terms of the number of genes that have been studied as drivers in existing literature.

Table 6.2 shows the top 10 predicted driver genes of deepDriver. 6 out of the 10 genes have been studied in existing literature or databases as potential driver genes of BRCA. The ninth gene 'DST' was found to have the potential to drive ductal carcinoma in situ to breast cancer [315]. 5 out of the 10 genes have been studied as driver genes of COAD in the existing literature. Meanwhile, among the rest 5 genes, 'AMER1' and 'ADAMTSL3' were found to be frequently mutated in COAD [316, 317]. 'LAMA3' were predicted as biomarkers which could be used to diagnose COAD in the early stage [318]. 'KMT2A' belongs to the KMT2 family which is related to COAD [319]. 4 out of 10 genes have been studied as driver genes of LUAD. The tenth gene 'HERC2P3' contains a microsatellite locus that can precisely discriminate LUAD samples and non-tumor samples [320]. As for three competing algorithms, Table 6.3, 6.4 and 6.5 show their prediction results. In summary, deepDriver performed better than the three competing algorithms in predicting new cancer drivers. Its prediction results were in concert with existing studies which further reveal the value of

deepDriver in predicting cancer driver genes.

6.4 Conclusion

In this study, we proposed an algorithm to predict cancer driver genes with CNN. The method combined CNN with similarity networks so that the functional impact of mutations and similarities of gene expression can be learned simultaneously, which improve the accuracy of driver gene prediction. Experiments performed on BRCA, COAD and LUAD then showed that deepDriver was superior to the competing algorithms in terms of both cross-validation and *de novo* prediction.

In the future, similarity networks calculated by different strategies and predictive features extracted by other algorithms can both be used to improve the prediction accuracy. Meanwhile, the algorithm can be applied to the pancancer dataset to predict generic cancer driver genes. Since the total number of cancer driver genes is much higher than that of a specific type of cancer, candidate driver genes can also be further classified into TSG and oncogene on the pancancer dataset.

Acknowledgements

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC)

Table 6.1: 12 features extracted from mutation data.

No.	Name	Description
1	Silent fraction	Fraction of silent mutations
2	Nonsense fraction	Fraction of nonsense mutations
3	Splice site fraction	Fraction of splice site mutations
4	Missense fraction	Fraction of missense mutations
5	Recurrent missense fraction	Fraction of recurrent missense mutations
6	Frameshift indel fraction	Fraction of frameshift indel mutations
7	Inframe indel fraction	Fraction of Inframe indel mutations
8	Lost start and stop fraction	Fraction of Lost start and stop mutations
9	Missense to silent	Ratio of missense to silent mutations
10	Non-silent to silent	Ratio of non-silent to silent mutations
11	Normalized missense position entropy	See Section 6.2.3
12	Normalized mutation entropy	See Section 6.2.3

Table 6.2: Top 10 predictions of deepDriver

Gene Names	Reference
BRCA	
PTEN	[321]
HCFC1	[322, 323]
UTRN	[324]
ZNF517	
STAG2	[322, 323]
ZFP36L1	[325]
ZNF91	
VPS13C	
DST	
FBXW7	[326]
COAD	
AMER1	
SOX9	[327]
NRAS	[328]
MTOR	[329]
ATM	[330]
ADAMTSL3	
ELMO1	[331]
TG	
LAMA3	
KMT2A	
LUAD	
XIST	[332]
MALAT1	[333]
STK11	[334]
USH1C	
HSP90AB2P	
BNIP3P1	
EEF1A1P9	
UBE2MP1	
SMAD4	[335]
HERC2P3	

Table 6.3: Top 10 predictions of 20/20+

Gene Names	Reference
BRCA	
KMT2C	[336]
PTEN	[321]
ANKRD12	
NF1	[337]
ANKHD1-EIF4EBP3	
ARID4B	
MCM7	
MYO6	
MLLT4	[323]
CEP128	
COAD	
ATM	[330]
SOX9	[327]
LAMA3	
ADAMTSL3	
ELMO1	[331]
OLFM1	
BRINP1	
ACVR1B	
CNOT1	
PCDH7	
LUAD	
LRRIQ1	
HECTD4	
EPB41L3	[338]
NF1	[339]
CEP350	
PRKDC	
APC	
MYH9	
POSTN	
FN1	

Table 6.4: Top 10 predictions of SVM

Gene Names	Reference
BRCA	
VPS13C	
UTRN	[324]
HCFC1	[322, 323]
MLLT4	[323]
ZNF91	
STAG2	[322, 323]
FBXW7	[326]
MALAT1	
NRK	
BAZ2B	
COAD	
ATM	[330]
NRAS	[328]
MTOR	[329]
SOX9	[327]
ADAMTSL3	
ELMO1	[331]
AMER1	
KMT2B	
FBN2	
KMT2A	
LUAD	
XIST	[332]
MALAT1	[333]
USH1C	
SNRPN	
STK11	[334]
SMAD4	[335]
POLA1	
MAGEE1	
BRAF	
CTNNB1	

Table 6.5: Top 10 predictions of OncodriveCLUST

Gene Names	Reference
BRCA	
ACTN4	[340]
AFF2	
ATP2B3	
AVPR1B	
CASR	
CMYA5	
DIS3L	
EPB41L2	
FBXW8	
KCND3	
COAD	
AKAP12	[341]
C3orf20	
COL1A2	[342]
DOK1	[343]
FNDC1	
MSRB3	
NCOA2	[344]
NPHS1	
NRAP	
PCDHB13	

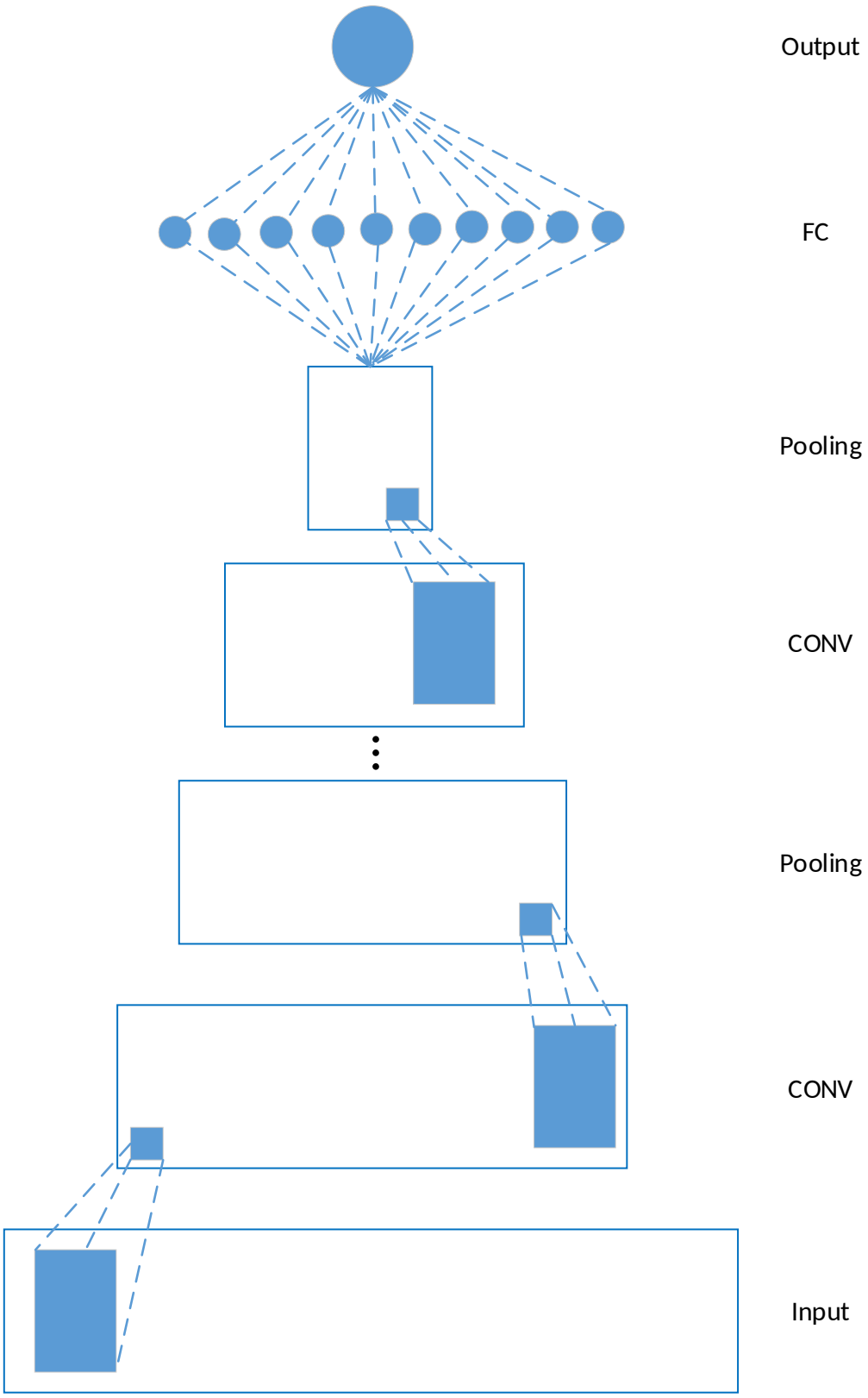

Figure 6.1: Schematic 1-D CNN. In this study, each CONV layer is followed by a pooling layer and the CONV-POOL pattern is repeated for several times. The final structure of the model is determined by grid search.

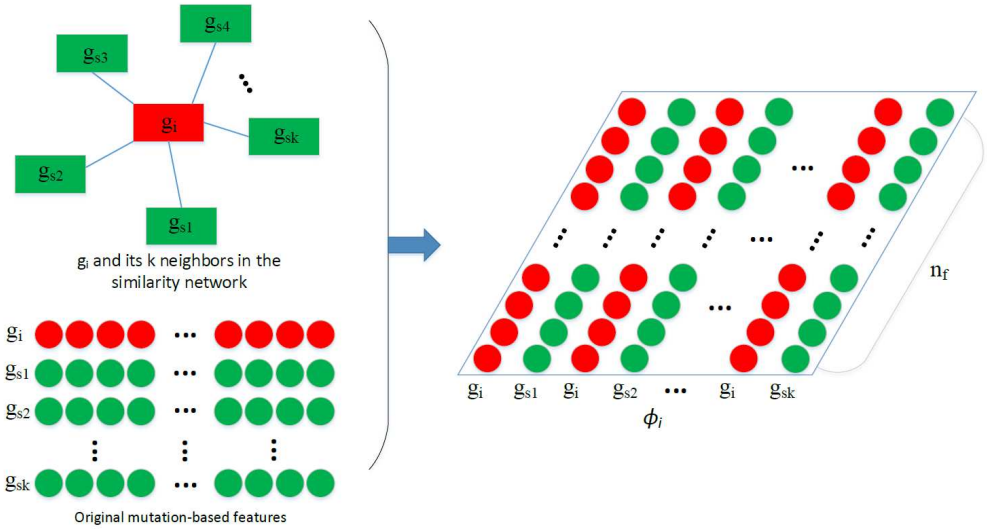

Figure 6.2: Construction of ϕ_i. Given the feature vectors of g_i and its k nearest neighbors $g_{s1}, g_{s2}, \ldots, g_{sk}$, a feature matrix ϕ_i is constructed by arranging the $2k$ vectors into a $2k \times n_f$ matrix, which is then used in the convolution.

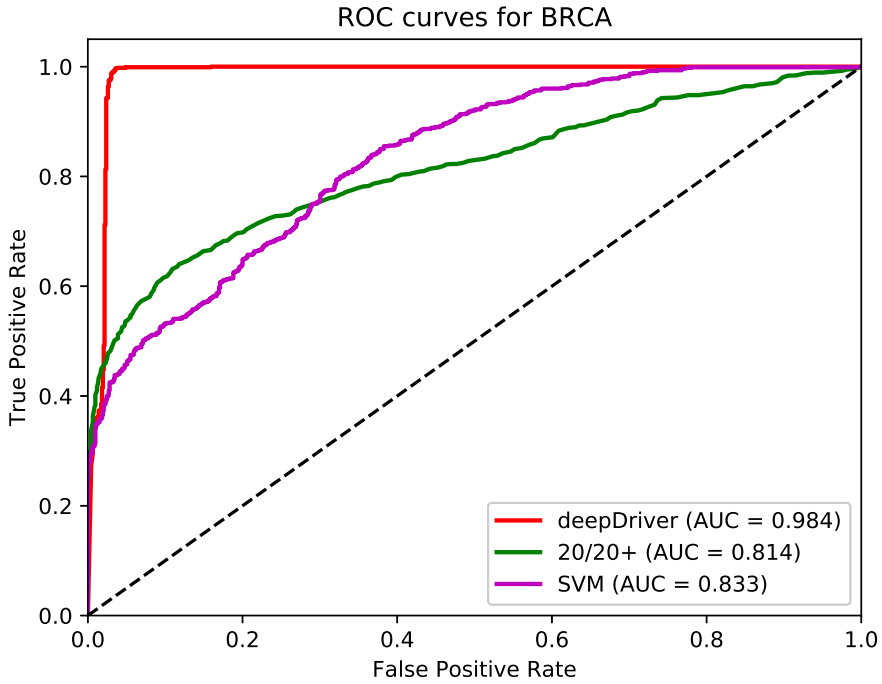

Figure 6.3: ROC curves of the three algorithms obtained on the dataset of BRCA. The red, green and magenta lines depict the ROC curves of deepDriver, 20/20+ and SVM, respectively. The AUC value of deepDriver is 0.984, which is at least 15.1% higher than that of the other two algorithms.

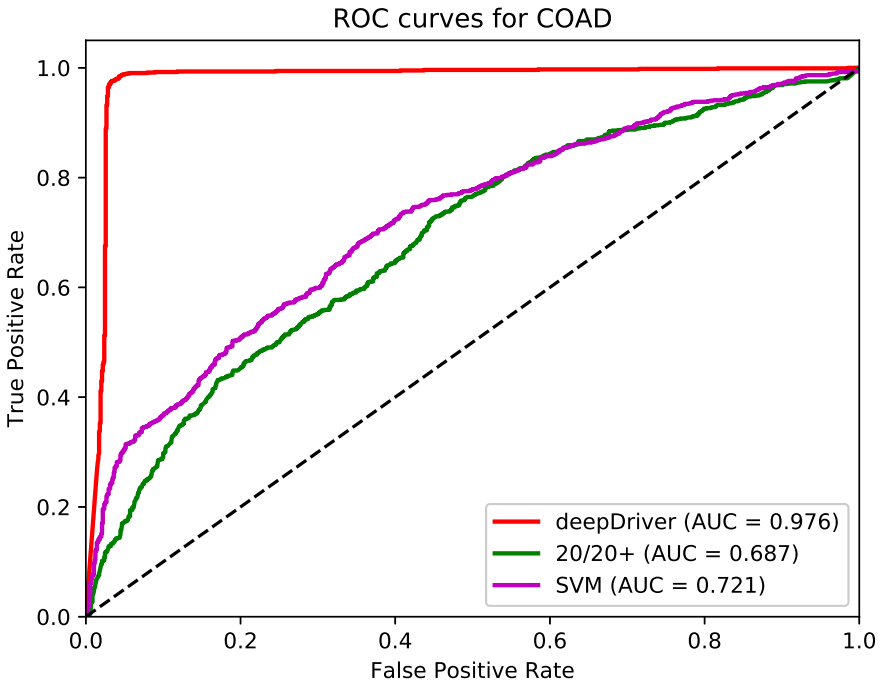

Figure 6.4: ROC curves of the three algorithms obtained on the dataset of COAD. The red, green and magenta lines depict the ROC curves of deepDriver, 20/20+ and SVM, respectively. The AUC value of deepDriver is 0.976, which is at least 25.5% higher than that of the other two algorithms.

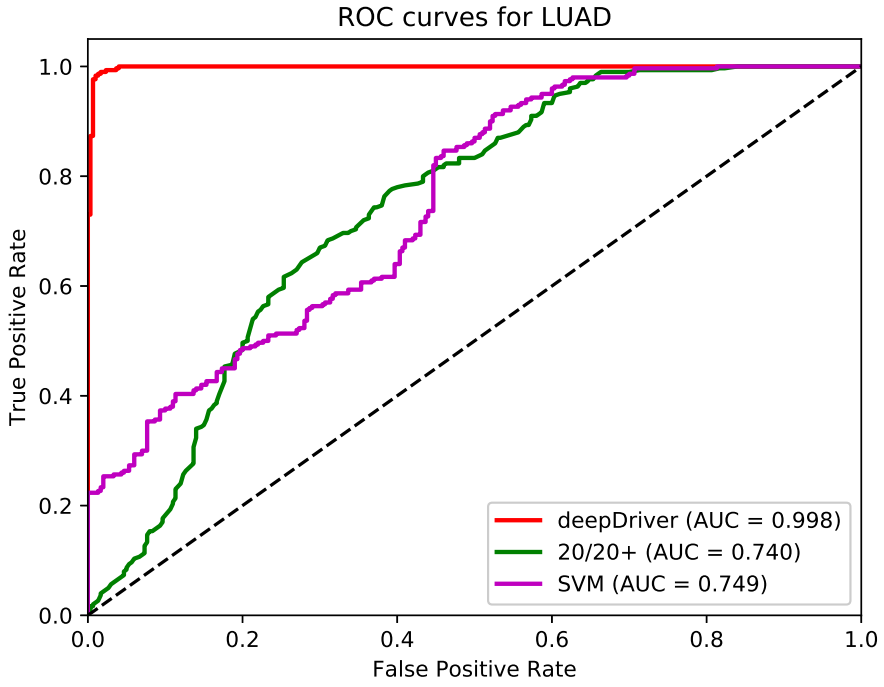

Figure 6.5: ROC curves of the three algorithms obtained on the dataset of LUAD. The red, green and magenta lines depict the ROC curves of deepDriver, 20/20+ and SVM, respectively. The AUC value of deepDriver is 0.998, which is at least 24.9% higher than that of the other two algorithms.

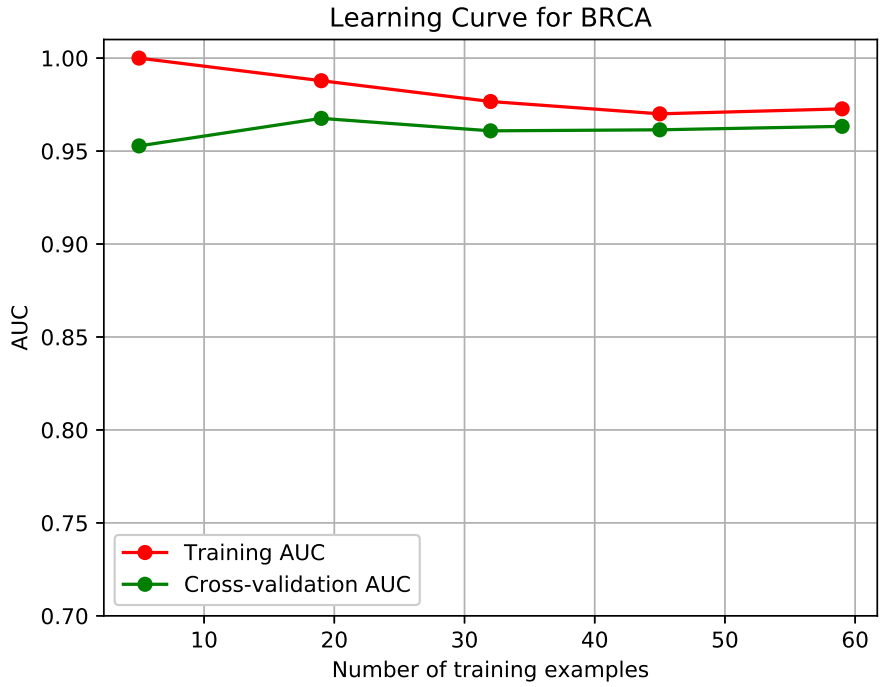

Figure 6.6: Learning curve for BRCA.

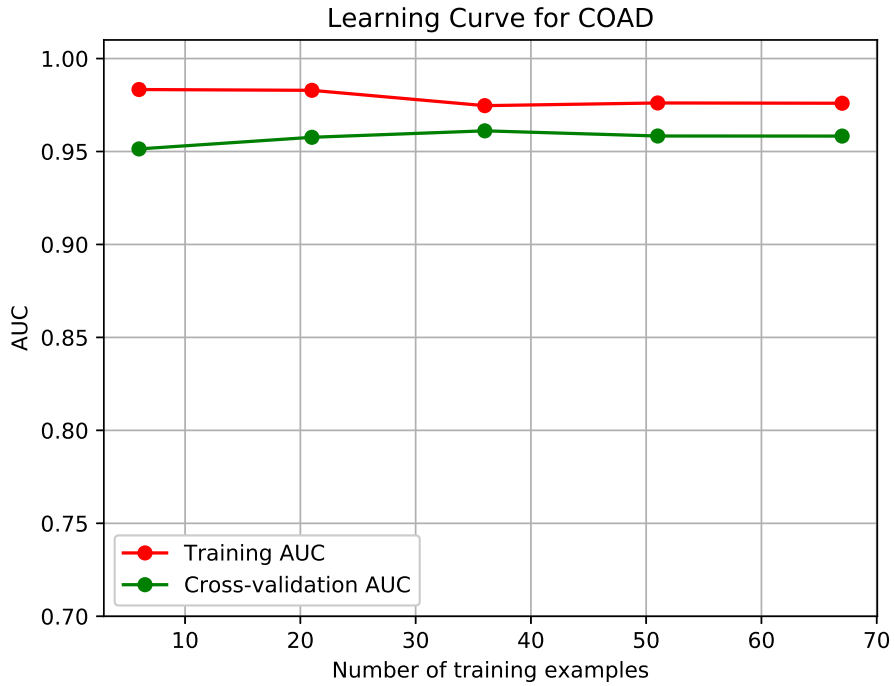

Figure 6.7: Learning curve for COAD.

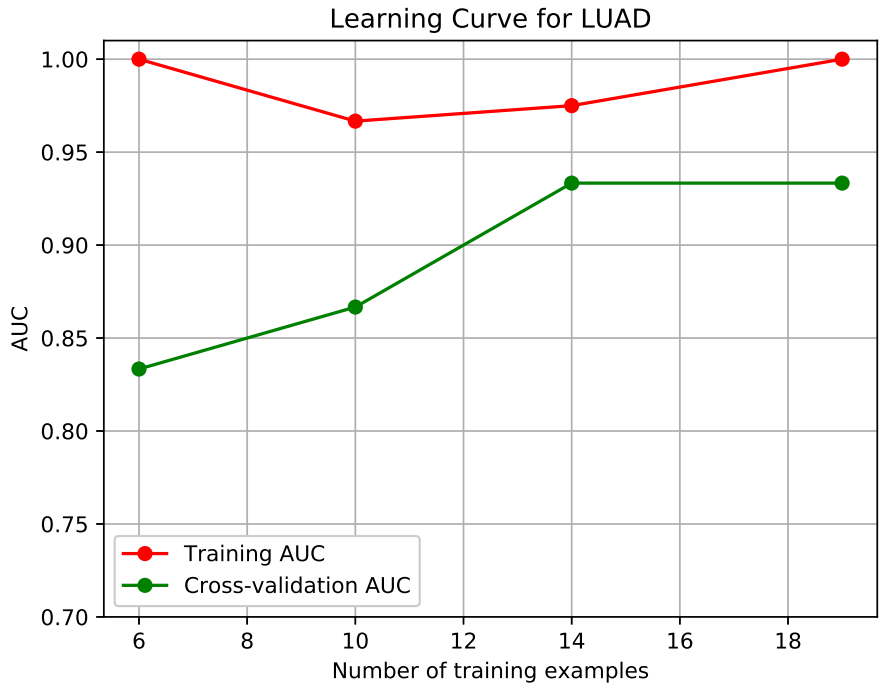

Figure 6.8: Learning curve for LUAD.

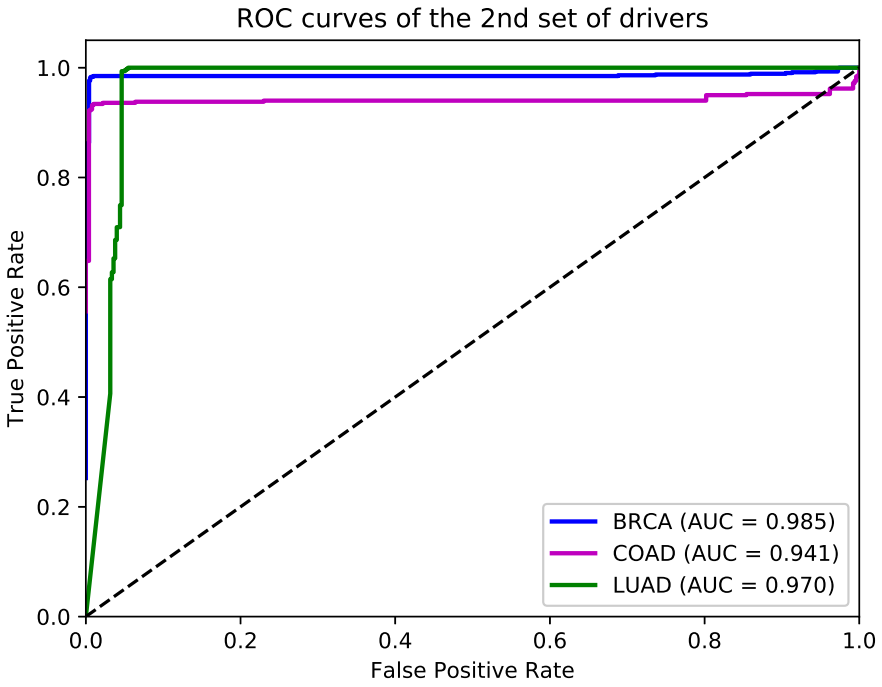

Figure 6.9: ROC curves of deepDriver obtained from the second sets of driver genes.

7

Identifying disease-gene associations with graph-regularized manifold learning

Prepared as: Ping Luo, Qianghua Xiao, Pi-Jing Wei, Bo Liao, and Fang-Xiang Wu. Identifying disease-gene associations with graph-regularized manifold learning. Frontiers in Genetics, 10:270, 2019. PL, QX, PJW, BL and FXW discussed about the methods. PL implemented the algorithm, designed and performed the experiments. FXW supervised this study. PL and FXW wrote the manuscript. All authors read, revised and approved the final version of the manuscript.

In previous chapters, a few supervised models are used to predict disease genes. If the models have to be trained separately for different diseases, the corresponding methods cannot be applied for diseases with only a few or no known associated genes, since the number of the instances is not enough to train the models. To solve this issue, we can extract features for both diseases and genes and predict disease-gene associations instead of associated genes for a specific disease. The number of the positive instances is then equals to the number of all known disease-gene associations, which is large enough to train the model. For instance, algorithms proposed in Chapter 5 has used this strategy. Additionally, we can also use NMF-based methods to solve this issue. These methods define disease-gene prediction as a matrix completion problem. Each entry in the association matrix is regarded as the probability of a disease-gene pair being associated.

In this chapter, we also propose an NMF-based method. However, unlike existing methods which use matrix completion to solve the problem, we map the diseases and genes onto a lower dimensional manifold and use their geodesic distance to determine whether they are associated. Our assumption is that the distance among a disease and its associated genes should be smaller than that among the disease and other genes. This chapter fulfills Objective 6 of this thesis.

Abstract

Complex diseases are known to be associated with disease genes. Uncovering disease-gene associations is critical for diagnosis, treatment, and prevention of diseases. Computational algorithms which effectively predict candidate disease-gene associations prior to experimental proof can greatly reduce the associated cost and time. Most existing methods are disease-specific which can only predict genes associated with a

specific disease at a time. Similarities among diseases are not used during the prediction. Meanwhile, most methods predict new disease genes based on known associations, making them unable to predict disease genes for diseases without known associated genes. In this study, a manifold learning-based method is proposed for predicting disease-gene associations by assuming that the geodesic distance between any disease and its associated genes should be shorter than that of other non-associated disease-gene pairs. The model maps the diseases and genes into a lower dimensional manifold based on the known disease-gene associations, disease similarity and gene similarity to predict new associations in terms of the geodesic distance between disease-gene pairs. In the 3-fold cross-validation experiments, our method achieves scores of 0.882 and 0.854 in terms of the area under of the receiver operating characteristic (ROC) curve (AUC) for diseases with more than one known associated genes and diseases with only one known associated gene, respectively. Further *de novo* studies on Lung Cancer and Bladder Cancer also show that our model is capable of identifying new disease-gene associations.

7.1 Introduction

Complex diseases are caused by a group of genes known as disease genes. Identifying disease-gene associations is of critical importance since it helps us unravel the mechanisms of diseases, which has many applications such as diagnosis, treatment and prevention of disease. With the advances in high-throughput experimental techniques, a large amount of data that indicate associations between diseases and their associated genes have been generated, which could accelerate the identification of disease-associated genes. However, it is expensive and time-consuming to experimentally prove an association between a gene and a disease. Computational methods that translate the experimental data into legible disease-gene associations are necessary for in-depth experimental validation.

Currently, many algorithms have been developed to predict disease-gene associations, and they can be briefly divided into two categories: the machine learning-based methods and the network-based methods. The typical machine learning-based methods extract gene-related features and train models that can discriminate disease genes and passenger genes [3, 78, 41, 220, 345]. Since the features are extracted for genes, these algorithms are usually single-task algorithms which once can only predict disease genes for a specific disease. Thus, for diseases that have a few or no known associated genes, the number of the genes would be too small to train the model. In the meantime, the relationships among diseases are usually not used in the prediction since only one disease is considered at a time. Matrix completion methods, as a type of machine learning methods, can solve the above two issues by jointly predicting disease-gene associations and leveraging the similarities among diseases during the calculation [6, 7]. However, matrix completion methods generally do not have the global optimal solutions and could take a very long time to converge to even a local optimal solution. Network-based methods are based on the assumption that genes close related in the network are associated with the same diseases. Centrality indices, random walk and network energy are used in many

methods to predict disease-gene associations [18, 22, 47, 44]. Although most network-based methods are not affected by the above two issues, their performance is strongly affected by the quality of networks, and they usually perform worse than machine learning-based methods on diseases with many known associated genes [117, 118].

In this study, we propose a manifold learning-based method (dgManifold) to predict disease-gene associations. In our dgManifold, genes and diseases are regarded as points in the same high-dimensional Euclidean space. Our assumption is that diseases and their associated genes should be consistent in some lower dimensional manifold, and the geodesic distance between a disease and its associated genes should be shorter than that of other non-associated disease-gene pairs. Although the Euclidean distance between diseases and genes in the high-dimensional space may not reflect their true geodesic distance, we can map the diseases and genes into a low-dimensional manifold based on the experimentally verified disease-gene associations [346, 347]. Then, the true geodesic distance between all the disease-gene pairs can be calculated. In the meantime, the mapping process is regularized by two affinity graphs, disease similarity network and gene similarity network, so that the learned representations with the similarity information can further increase the prediction accuracy. Additionally, since our dgManifold is a supervised method, and it is difficult (if possible) to learn valuable representations for diseases that only have a few or no known associated genes. A prior information vector calculated with the disease similarities and known disease-gene associations should be combined with the original association data to solve this issue. Similar strategies have been applied to calculate the initial probabilities used in the random walk, which have improved the accuracy of predicting miRNA-disease associations. [348, 349, 350].

In the rest of the manuscript, Sections 7.2 describes our algorithm as well as the data sources and evaluation metrics used in the study. Section 7.3 discusses the evaluation results. Section 7.4 draws some conclusions.

7.2 Methods and material

7.2.1 General model

Given n diseases and m genes, the associations among them can be represented by a matrix $A \in R^{n \times m}$ in which $a_{ij} = 1$ if disease i is associated with gene j, and otherwise $a_{ij} = 0$. Intuitively, each disease can be represented by a binary m-dimensional row vector while each gene can be represented by a binary n-dimensional column vector. However, in these high-dimensional spaces, it is hard to calculate the actual distance between a disease and a gene.

If we map the diseases and genes into the same manifold with a lower dimensionality and assume that the distance between a disease and its associated genes should be as short as possible on this manifold, predicting disease-gene associations can be solved by computing this mapping based on known disease-gene associations, which can be mathematically formulated as: finding k-dimensional representatives of diseases

$\mathbf{r}_1, \ldots, \mathbf{r}_n$ and k-dimensional representatives of genes $\mathbf{q}_1, \ldots, \mathbf{q}_m$ such that the following objective function is minimized

$$O_k = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\|\mathbf{r}_i - \mathbf{q}_j\|^2. \tag{7.1}$$

However, without any constraints, the objective function (7.1) is not well defined. To illustrate this, if k-dimensional vectors $\mathbf{r}_i^+$ and $\mathbf{q}_j^+$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ minimize the objective function (7.1), then $\epsilon\mathbf{r}_i^+$ and $\epsilon\mathbf{q}_j^+$ can further minimize the objective function when $0 \leq \epsilon < 1$. Especially, when $\epsilon = 0$, any k-dimensional vectors $\mathbf{r}_i^+$ and $\mathbf{q}_j^+$ can minimize the objective function. Therefore, to make the optimization problem well defined, the following constraints are added

$$\sum_{i=1}^{n}\mathbf{r}_i\mathbf{r}_i^T = I_k \quad \text{and} \quad \sum_{j=1}^{m}\mathbf{q}_j\mathbf{q}_j^T = I_k. \tag{7.2}$$

where I_k is the $k \times k$ identity matrix. As a results, the learned representations are unique with these constraints.

To insure that the mapped representations of diseases and genes are in concert with their intrinsic properties, two affinity graphs, disease similarity network and gene similarity network are used to regularize the objective function (7.1), and the new objective function is as follows

$$O_k = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ij}\|\mathbf{r}_i - \mathbf{q}_j\|^2 + \frac{\alpha}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^d\|\mathbf{r}_i - \mathbf{r}_j\|^2 + \frac{\beta}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}^g\|\mathbf{q}_i - \mathbf{q}_j\|^2 \tag{7.3}$$

where S^d and S^g are the adjacency matrices of the disease similarity network and the gene similarity network, respectively. α and β are the regularization coefficients.

Note that

$$\begin{aligned}
O_k =& \sum_{i=1}^{n}(\sum_{j=1}^{m} a_{ij})\mathbf{r}_i^T\mathbf{r}_i + \sum_{j=1}^{m}(\sum_{i=1}^{n} a_{ij})\mathbf{q}_j^T\mathbf{q}_j - 2\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\mathbf{r}_i^T\mathbf{q}_j \\
&+\alpha\sum_{i=1}^{n}(\sum_{j=1}^{n} s_{ij}^d)\mathbf{r}_i^T\mathbf{r}_i - \alpha\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^d\mathbf{r}_i^T\mathbf{r}_j \\
&+\beta\sum_{i=1}^{m}(\sum_{j=1}^{m} s_{ij}^g)\mathbf{q}_i^T\mathbf{q}_i - \beta\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}^g\mathbf{q}_i^T\mathbf{q}_j \\
=& \sum_{i=1}^{n} A_{ri}\mathbf{r}_i^T\mathbf{r}_i + \sum_{j=1}^{m} A_{cj}\mathbf{q}_j^T\mathbf{q}_j - 2\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\mathbf{r}_i^T\mathbf{q}_j \\
&+\alpha\sum_{i=1}^{n} S_i^d\mathbf{r}_i^T\mathbf{r}_i - \alpha\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^d\mathbf{r}_i^T\mathbf{r}_j \\
&+\beta\sum_{j=1}^{m} S_j^g\mathbf{q}_j^T\mathbf{q}_j - \beta\sum_{j=1}^{m}\sum_{i=1}^{m} \mathbf{s}_{ij}^g\mathbf{q}_i^T\mathbf{q}_j \\
=& \sum_{i=1}^{n}(A_{ri} + \alpha S_i^d)\mathbf{r}_i^T\mathbf{r}_i + \sum_{j=1}^{m}(A_{cj} + \beta S_j^d)\mathbf{q}_j^T\mathbf{q}_j \\
&-2\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\mathbf{r}_i^T\mathbf{q}_j - \alpha\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^d\mathbf{r}_i^T\mathbf{r}_j - \beta\sum_{j=1}^{m}\sum_{i=1}^{m} s_{ij}^g\mathbf{q}_i^T\mathbf{q}_j
\end{aligned} \tag{7.4}$$

where $S_i^d = \sum_{j=1}^n s_{ij}^d, S_i^g = \sum_{j=1}^m s_{ij}^g, A_{ri} = \sum_{j=1}^m a_{ij}, A_{cj} = \sum_{i=1}^n a_{ij}$. Let

$$\begin{aligned} L^{11} &= diag[A_{r1} + \alpha S_1^d, A_{r2} + \alpha S_2^d, \ldots, A_{rn} + \alpha S_n^d] - \alpha S^d, \\ L^{22} &= diag[A_{c1} + \beta S_1^g, A_{c2} + \beta S_2^g, \ldots, A_{cm} + \beta S_m^g] - \beta S^g, \end{aligned} \tag{7.5}$$

the objective function (7.3) can be simplified as

$$O_k = \sum_{i=1}^n \sum_{j=1}^n L^{11} \mathbf{r}_i^T \mathbf{r}_j + \sum_{i=1}^m \sum_{j=1}^m L^{22} \mathbf{q}_i^T \mathbf{q}_j - 2 \sum_{i=1}^n \sum_{j=1}^m a_{ij} \mathbf{r}_i^T \mathbf{q}_j \tag{7.6}$$

Furthermore, let

$$r_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix}, q_j = \begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jk} \end{bmatrix}, z_t = \left[\begin{array}{c} x_{1t} \\ \vdots \\ x_{nt} \\ \hdashline y_{1t} \\ \vdots \\ y_{mt} \end{array}\right] = \left[\begin{array}{c} x_t \\ \hdashline y_t \end{array}\right], \tag{7.7}$$

$$\begin{aligned} &A_r = diag[A_{r1}, \ldots, A_{rn}], \quad A_c = diag[A_{c1}, \ldots, A_{cm}], \\ &L^d = diag[S_1^d, \ldots, S_n^d] - S^d, \quad L^g = diag[S_1^g, \ldots, S_m^g] - S^g, \end{aligned} \tag{7.8}$$

$$L = \begin{bmatrix} A_r + \alpha L^d & -A \\ -A^T & A_c + \beta L^g \end{bmatrix}, \tag{7.9}$$

objective function (7.6) can be simplified as

$$\begin{aligned} O_k =& \sum_{t=1}^k \sum_{i=1}^n \sum_{j=1}^n L^{11} x_{it} x_{jt} + \sum_{t=1}^k \sum_{i=1}^m \sum_{j=1}^m L^{22} y_{it} y_{jt} - 2 \sum_{t=1}^k \sum_{i=1}^n \sum_{j=1}^m a_{ij} x_{it} y_{jt} \\ =& \sum_{t=1}^k [\mathbf{x}_t^T L^{11} \mathbf{x}_t + \mathbf{y}_t^T L^{22} \mathbf{y}_t - 2 \mathbf{x}_t^T A \mathbf{y}_t] \\ =& \sum_{t=1}^k [\mathbf{x}_t^T \mathbf{y}_t^T] \begin{bmatrix} L^{11} & -A \\ -A^T & L^{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} \\ =& Tr(Z^T L Z) \end{aligned} \tag{7.10}$$

where $Z = (z_1, \ldots, z_k)$. Therefore, minimizing the objective function (4) with constraints (2) is equivalent to minimize the following function

$$Q_k = Tr(Z^T L Z) \tag{7.11}$$

with constraints

$$Z^T Z = X^T X + Y^T Y = 2 I_k \tag{7.12}$$

According to [351], minimizing objective function (7.11) with constraints (7.12) can be solved by

$$Z^* = (\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}) \tag{7.13}$$

where $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ are k eigenvectors correspond to the k smallest eigenvalues of L. Meanwhile, the smallest eigenvalue is 0, and entries in the corresponding eigenvector $\mathbf{u}_0$ are identical to each other, which does not contribute to the calculation of the geodesic distance. Thus, let $\hat{Z}$ denote the matrix by removing the fist column of Z^*. The first n rows of $\hat{Z}$ are the obtained $(k-1)$-dimensional representations of diseases, and the rest m rows of $\hat{Z}$ are the learned representations of genes. The geodesic distance between a disease i and gene j can be calculated by

$$gdist_{ij} = \|\hat{\mathbf{r}}_i - \hat{\mathbf{q}}_j\|^2. \tag{7.14}$$

7.2.2 Similarity network

Gene similarity

In this study, the learning process is regularized by similarity networks, and the similarities of genes are calculated based on the Gene Ontology (GO). GO database provides a set of vocabularies to describe the function of genes and gene products [125, 126]. The GO terms and their relationships are manifested as a directed acyclic graph (DAG) where nodes represent terms while edges represent semantic relationships. Many algorithms have been proposed to calculate the similarities of genes using ontology data, and the approach proposed by [130] is used in this study.

Let $DAG_h = (T_h, E_h)$ denote GO term h, where T_h contains all the successor GO terms of h in the DAG, and E_h contains the semantic relationships between h and other terms in T_h. Each term t in T_h has a τ-value related to h:

$$\begin{cases} \tau_h(t) = 1, if\ t = h \\ \tau_h(t) = max\{w_e * \tau_h(t^{'}) | t^{'} \in children\ of\ t\}, otherwise \end{cases} \tag{7.15}$$

where w_e is the weight of the edge (semantic relationships) in the DAG. Two types of semantic relationships ("*is_a*" and "*part_of*") are used in the DAG, and the corresponding w_e is set to 0.8 and 0.6, respectively, as recommended in [130].

Given $DAG_h = (T_h, E_h)$ and $DAG_b = (T_b, E_b)$ for GO terms h and b, their similarity can be computed by

$$sgo(h, b) = \frac{\sum_{t \in T_h \cap T_b} (\tau_h(t) + \tau_b(t))}{\sum_{t \in T_h} \tau_h(t) + \sum_{t \in T_b} \tau_b(t)} \tag{7.16}$$

Then, the similarity of one GO term $t^{'}$ and a set of GO terms $GO = \{t_1, t_2, \ldots, t_l\}$ is defined as

$$SGO(t^{'}, GO) = \max_{1 \le i \le l} (SGO(t^{'}, t_i)) \tag{7.17}$$

Finally, the functional similarity of two genes g_1 and g_2 is calculated by

$$s^g_{g_1, g_2} = \frac{\sum_{1 \le i \le n_1} SGO(t_{1i}, GO_2) + \sum_{1 \le j \le n_2} SGO(t_{2j}, GO_1)}{n_1 + n_2} \tag{7.18}$$

where $GO_1 = \{t_{11}, t_{12}, \ldots, t_{1n_1}\}$ and $GO_2 = \{t_{21}, t_{22}, \ldots, t_{2n_2}\}$ are two sets of GO terms that describe g_1 and g_2, respectively.

Disease similarity

The similarities among diseases are also calculated with the ontology data. Instead of GO, the Human Phenotype Ontology (HPO) [127] is used to characterize human diseases. The HPO provides a vocabulary of phenotypic terms related to human diseases. Each term represents a clinical abnormality, and all the terms are structured as a DAG, in which every term is related to their parent terms by "is_a" relationships. Although diseases are not directly described by the HPO, the annotation file provided by HPO contains terms associated with every disease, and thus Eq. 7.17, 7.18 can be used to compute the similarities of diseases. When we calculate the similarities of phenotypic terms based on the DAG, w_e in Eq. 7.15 is set to 0.7 as recommended in [352].

7.2.3 Prior information

For diseases with only a few associated genes, the limited information would affect the performance of any computational algorithms. This problem is especially serious for diseases with no known associated genes. To solve this problem, we add some prior information for diseases with no known associations.

Specifically, given a disease $i^{'}$, $\mathbf{p}_{i^{'}}$ is added to the $i^{'}$-th row of the matrix A as prior information so that the shortage of known information can be alleviated. The j-th entry of $\mathbf{p}_{i^{'}}$ is calculated by

$$p_{i^{'}j} = \Big(\sum_{i=1, i \neq i^{'}}^{n} s^{d}_{ii^{'}} a_{ij} \Big) / \Big(\sum_{i=1, i \neq i^{'}}^{n} a_{ij} \Big) \tag{7.19}$$

In our experiments, when cross-validation is used to evaluate the algorithm, the prior information is added to the i-th row of matrix A as long as one of the associated genes of disease i is left to test the model. Meanwhile, in the *de novo* study, prior information is also added to the diseases used for evaluation.

7.2.4 Data sources

The disease-gene association data are downloaded from the Online Mendelian Inheritance in Man (OMIM) database [274] in August 2018. The Morbid Map at OMIM contains nearly seventy-five hundred entries sorted alphabetically by disorder names. Each entry represents an association between a gene and a disease. Different entries are labeled with different tags ('(3)', '[]' and '?') which indicate their reliabilities. To obtain a reliable association dataset, based on [154], three steps were performed to preprocess the originally downloaded data. First, entries with the tag '(3)' are selected while others are abandoned. We adopt this strategy because diseases with tag '(3)' indicate that the molecular basis of these diseases is known and the associations are reliable, while entries with '[]' represent abnormal laboratory test values, and entries with '?' represent provisional disease-gene associations. Second, disease entries are classified into distinct diseases by merging disease subtypes based on their given disorder names. For instance, 17 entries of "Leigh syndrome" are merged into disease "Leigh syndrome", and the 19 complementary terms of "Lung cancer somatic" are merged into "Lung Cancer". Third, 74 diseases are removed because they are not annotated by any HPO

terms. During the classification, string match was used to classify adjacent entries, followed by a manual verification. Finally, we obtain a dataset consisting of 4,770 associations between 1,537 diseases and 3320 genes. Among the 1,537 diseases, 917 have only one associated gene (single-gene disease), while the rest diseases have at least two associated genes (multiple-gene disease).

The ontology data of genes and phenotypes are downloaded from the GO database [125, 126], and the HPO database [127], respectively. The PPI network used in the competing algorithms is downloaded from the InWeb_InBioMap database (version 2016_09_12) [170].

7.2.5 Evaluation metrics

In this study, the algorithm is evaluated in two steps. In the first step, our dgManifold is compared with two competing algorithms: PCFM [7] and Katz [220]. PCFM is a matrix completion method which integrates disease similarities and gene similarities to predict disease-gene associations. Katz is a classic network-based method which uses Katz centrality to rank the disease-gene associations. We choose these two algorithms because they are all multi-task algorithms which can predict all disease-gene associations as our dgManifold does. The AUC (area under of the receiver operating characteristic (ROC) curve) scores calculated from 3-fold cross-validation are used to compare these three algorithms.

ROC curve plots the true positive rate [TP/(TP+FN)] verses the false positive rate [FP/(FP+TN)] at different thresholds, and a larger AUC represents better overall performance. In this study, a true positive (TP) is a known disease-gene association (positive sample) predicted as a disease-gene association, while a false positive (FP) is a non-disease-gene association (negative sample) predicted as a disease-gene association. A false negative (FN) is a positive sample predicted as negative while a true negative (TN) is a negative sample predicted as negative. Since negative samples are not included in existing databases, we randomly select a set of unknown disease-gene pairs as negative samples. The number of negative samples is equal to that of positive samples. Considering that the selected negative samples may have small possibilities to be a real disease-gene association, the random selection was run for five times to generate 5 sets of negative samples. The final AUC score is the average score obtained from the 5 sets of samples.

During the cross-validation, the known disease-gene associations are split into 3 groups, and the algorithm is run for 3 rounds. In each round, one group of associations is regarded as unknown ($a_{ij} = 0$), while the rest two groups of associations are used to train the model. The prior information is recomputed during every round of the cross-validation. Considering that single-gene diseases would have no known associated genes if they are left for testing the model during the cross-validation, predicting disease genes for these diseases is similar to predict disease genes for a completely new disease. Thus, the three algorithms are compared on multiple-gene diseases and single-gene diseases separately. Additionally, to show the effect of the prior information, the AUC scores of our method without prior information are also calculated.

In the second step, the model is trained with all the known associations, and the geodesic distance between every unknown disease-gene pairs is calculated. To find out whether our new predictions are in

concert with existing experimental studies, the top-10 predictions of two diseases, Lung Cancer and Bladder Cancer, are searched from the existing literature. In our dataset, Lung Cancer has 16 associated genes, and Bladder Cancer has 4 associated genes. We choose these two types of cancer because they are experimentally well studied which could better prove our results.

7.3 Results

7.3.1 Model parameters

In our study, several parameters affect the performance of the model. To obtain the optimal parameters, the grid search is conducted by searching k from $\{20, 30, 50, 100, 500, 800, 1000, 1200, 1500\}$ and α from $\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. β is set to be equal to α. The AUC score is used to determine whether the selected parameters are optimal. Finally, for multiple-gene diseases, the model performs best when $k = 30, \alpha = \beta = 0.2$, and for single-gene diseases, the optimal parameters are $k = 30, \alpha = \beta = 0.1$.

7.3.2 Cross-validation

Fig. 7.1 and Fig. 7.2 show the resulted ROC curves and AUC scores of the three competing algorithms on multiple-gene diseases and single-gene diseases, respectively. For multiple-gene diseases, our dgManifold achieves AUC score of 0.882 with prior information and 0.873 without prior information, while the AUC scores of Katz and PCFM are 0.742 and 0.636, respectively. For single-gene diseases, the AUC score of our dgManifold is 0.854 when prior information is used and 0.485 with no prior information, while the AUC scores of Katz and PCFM are 0.455 and 0.322, respectively. These results show that our method is superior to the competing methods in terms of the AUC scores.

It is worth noting that the AUC scores of all three algorithms are less than 0.5 when they are applied to single-gene diseases. This is mainly because that single-gene diseases have no known associated genes during the cross-validation, and algorithms can only use disease similarities and association data of other diseases to perform the prediction. These data are not enough to generate accurate results, especially for supervised algorithms. Thus, prior information is necessary for the algorithm. In fact, the results of our experiments have shown that the prior information is beneficial to the prediction of disease-gene associations, especially when the diseases have no known associated genes.

7.3.3 *De novo* study

In addition to AUC scores, we evaluate the performance of our dgManifold in predicting new disease-gene associations. Specifically, Lung Cancer and Bladder Cancer are selected, and prior information corresponded to these two diseases is added to matrix A. Then, all known disease-gene associations are used to train the model ($k = 30, \alpha = \beta = 0.2$), and the geodesic distance between all the unknown disease-gene pairs is

calculated. For each of the two selected diseases, the unknown disease-gene pairs are ranked based on the geodesic distance in ascending order, and the top-10 predictions are searched from existing literature.

Table 7.1 shows the results of *de novo* studies. 5 out of 10 predicted genes have been experimentally confirmed as associated with Lung Cancer. Among these genes, KCNK9 is a potential therapeutic target [353]. HTRA1 contributes to the tumor formation by inhibiting the TGF-beta pathway [354]. ATP6AP1 and MYL2 are two potential biomarkers [355, 356]. Mutation of C282Y allele in HFE is associated with Lung Cancer [357]. Although SEMA4A is not proved to be associated with Lung Cancer yet, it is related to Lung Inflammation and Colorectal Cancer, and its role in Lung Cancer genesis might be discovered in the future [358]. For Bladder Cancer, 3 out of 10 genes have been experimentally verified. Among them, SMAD3 mediates epithelial-mesenchymal transition which affects the invasion and migration of Bladder Cancer [359]. DMP1 is a tumor suppressor gene of Bladder Cancer [360]. CALR is potential biomarker [361]. These results show that our predictions are in concert with existing reports, and thus our dgManifold is valuable for predicting new disease-gene associations.

7.4 Conclusion

In this study, we have proposed dgManifold to predict disease-gene associations with manifold learning. Our dgManifold assumes that the distance between diseases and their associated genes should be shorter than that of other non-associated disease-gene pairs and maps the diseases and genes into a lower dimensional manifold based on known disease-gene associations, disease similarity and gene similarity. The prediction of new associations can be achieved by sorting the geodesic distance between unknown disease-gene pairs. The cross-validation results show that our model outperforms the competing algorithms in terms of AUC scores for both multiple-gene diseases and single-gene diseases. The further *de novo* studies also demonstrate that our dgManifold is valuable in predicting new disease-gene associations.

Note that dgManifold is only regularized by disease similarities and gene similarities at the current version, and the prior information is also obtained from the disease similarities. In the future, we can improve our method by regularizing the objective function with more types of data and computing the prior information with clinical evidences.

7.5 Acknowledgements

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC).

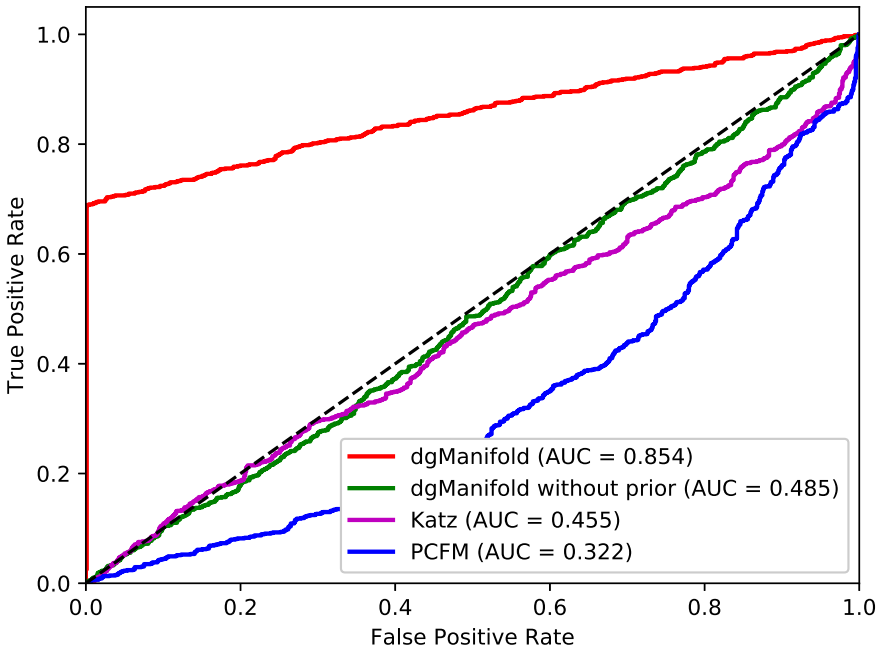

Figure 7.1: ROC curves of the three competing algorithms on multiple-gene diseases.

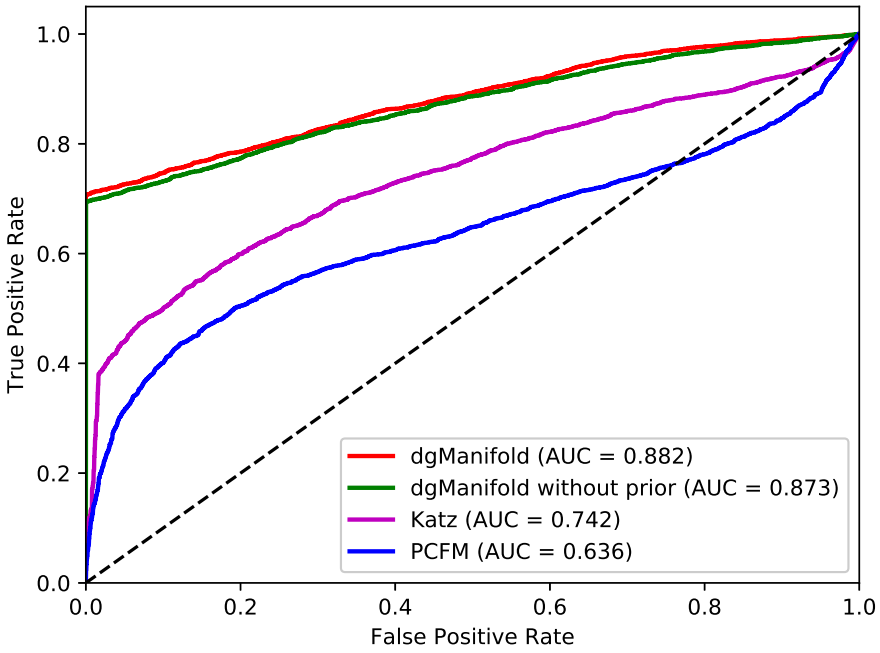

Figure 7.2: ROC curves of the three competing algorithms on single-gene diseases.

Table 7.1: Top 10 predictions for lung cancer and bladder cancer

Gene symbol	Reference
Lung Cancer	
SEMA4A	
KCNK9	[353]
MYL2	[356]
DENND5A	
HTRA1	[354]
GABRA1	
ATP6AP1	[355]
KCTD17	
HFE	[357]
BCS1L	
Bladder Cancer	
PDYN	
DKC1	
SMAD3	[359]
MCC	
DMP1	[360]
MGP	
CALR	[361]
CASQ2	
SOX18	
GATM	

8

Summary and future work

8.1 Summary

Disease-gene prediction is a critical yet challenging issue. The appropriate integration of multi-level biological data is the key to improving prediction accuracy. This thesis aims at fusing multiples types of data with multimodal deep learning to advance the performance of existing algorithms. In the meantime, several issues that limit the accuracy of prediction are also addressed. In total, six objectives are proposed in Chapter 1, and Chapters 2 to 7 have achieved these objectives.

Chapter 2 comprehensively reviews the computational algorithms for disease-gene prediction and achieves Objective 1. Classic and state-of-the-art algorithms, databases and evaluation methods are summarized in this chapter, and several future perspectives are discussed for designing new algorithms.

Chapter 3 designs a strategy to select negative data and applies the network energy-based model on both a PPI network and a differential co-expression network to predict disease genes.

Chapter 4 first proposes an approach to construct sample-specific networks using static PPI network and clinical gene expression data. Then, an ensemble strategy is used to predict disease genes from all the single sample-based networks with centrality-based features.

Chapter 5 presents a method that uses node2vec to extract raw network embeddings from different modalities (PPI networks and GO data in this study) and fuse them with multimodal DBN. The latent representations learned by the model then significantly improve the prediction accuracy.

Chapter 6 presents a strategy to fuse raw features (mutation-based features) and similarity information by a CNN model and use it to predict cancer driver genes.

Chapter 7 proposes an NMF-based method by using manifold learning to map diseases and genes onto a lower-dimensional manifold. The mapping process is based on the known disease-gene associations and regularized by disease similarities and gene similarities. After the mapping, the geodesic distance between each disease-gene pair is used to prioritize disease genes.

With our proposed algorithms, the accuracy of computational disease gene prediction has been greatly improved, and biochemists can combine the results of our prediction with their experiments to accelerate the identification of disease genes. Meanwhile, the proposed models could be applied to other areas to enhance the biological analysis. For instance, algorithms proposed in Chapters 5, 6 and 7 can be used to address

linkage prediction problems, such as the prediction of protein interactions, drug-target associations, and mRNA-disease associations, etc.

Note that our methods are not optimal. For instance, the algorithm proposed in Chapter 3 should combine different expression with differential co-expression instead of only using the latter information. Hierarchical clustering used in Chapter 4 should be compared with other algorithms to improve the clustering accuracy. Dropout should be added to the models proposed in Chapters 5 and 6 to improve their stability. Therefore, more efforts should be done to improve the performance of our algorithms in the future.

8.2 Future work

Based on the studies proposed in this thesis, several future directions for disease-gene prediction are proposed as follows:

1. Using multi-omics data to predict disease genes.

   Multi-omics data characterize different stages of cellular activities, and analyzing omic data is believed to improve the accuracy of computational prediction. However, for disease gene prediction, most algorithms still focus on genomic data, and only a few algorithms have used multi-omics data in their studies [30]. Therefore, new methods should apply other omic data (transcriptomic, proteomic, etc.) in their studies to discover the appropriate approaches to apply these data for disease-gene prediction.

2. Developing algorithms for personalized prediction.

   Complex diseases might be associated with many disease genes; however, only a subset of malfunctioning genes would lead to a disease, and the same disease on different patients might be caused by different sets of genes. Therefore, predicting patient-specific disease genes should be useful for personalized treatment. A previous study had used deep Boltzmann machine to predict personalized mutations [362]. The results are promising, and the accuracy could be further improved with more available samples.

3. Comparing the state-of-the-art methods, and developing a software package to implement these methods.

   In Chapter 2, we address that many algorithms have not been applied in real disease-gene prediction studies since they are not easy to use. Moreover, metrics alone cannot reveal the true prediction power of an algorithm, and a comprehensive study should be conducted to address the performance of the existing state-of-the-art methods. Therefore, a study should be proposed to compare the state-of-the-art methods and implement the best ones by a user-friendly package or web tool.

References

[1] Martin Oti and Han G Brunner. The modular nature of genetic diseases. *Clinical Genetics*, 71(1):1–11, 2007.

[2] Yana Bromberg. Disease gene prioritization. *PLoS Computational Biology*, 9(4):e1002902, 2013.

[3] 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.

[4] Manoj Pratim Samanta and Shoudan Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences*, 100(22):12579–12583, 2003.

[5] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International Conference on Machine Learning Workshop*, volume 79, 2012.

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

[7] 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, 2017.

[8] M Dawn Teare and Jennifer H Barrett. Genetic linkage studies. *The Lancet*, 366(9490):1036–1044, 2005.

[9] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356, 2008.

[10] Michael Boutros and Julie Ahringer. The art and design of genetic screens: RNA interference. *Nature Reviews Genetics*, 9(7):554, 2008.

[11] Rosario M Piro and Ferdinando Di Cunto. Computational approaches to disease-gene prediction: rationale, classification and successes. *The FEBS Journal*, 279(5):678–696, 2012.

[12] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association

loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[13] Feixiong Cheng, Junfei Zhao, and Zhongming Zhao. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in Bioinformatics*, 17(4):642–656, 2015.

[14] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6(1):227, 2005.

[15] Richard A George, Jason Y Liu, Lina L Feng, Robert J Bryson-Richardson, Diane Fatkin, and Merridee A Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Research*, 34(19):e130–e130, 2006.

[16] Martin Oti, Berend Snel, Martijn A Huynen, and Han G Brunner. Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, 43(8):691–698, 2006.

[17] Lude Franke, Harm Van Bakel, Like Fokkens, Edwin D De Jong, Michael Egmont-Petersen, and Cisca Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*, 78(6):1011–1025, 2006.

[18] 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.

[19] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular Systems Biology*, 4(1):189, 2008.

[20] William S Noble, Rui Kuang, Christina Leslie, and Jason Weston. Identifying remote protein homologs by network propagation. *The FEBS Journal*, 272(20):5119–5128, 2005.

[21] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gabor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics*, 138(3):333–408, 2013.

[22] 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):e1000641, 2010.

[23] Sinan Erten, Gurkan Bebek, and Mehmet Koyutürk. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *Journal of Computational Biology*, 18(11):1561–1574, 2011.

[24] 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.

[25] Oded Magger, Yedael Y Waldman, Eytan Ruppin, and Roded Sharan. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Computational Biology*, 8(9):e1002690, 2012.

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

[27] Jiawei Luo and Shiyu Liang. Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. *Journal of Biomedical Informatics*, 53:229–236, 2015.

[28] Rui Jiang. Walking on multiple disease-gene networks to prioritize candidate genes. *Journal of Molecular Cell Biology*, 7(3):214–230, 2015.

[29] 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, 2018.

[30] Xiujuan Lei and Yuchen Zhang. Predicting disease-genes based on network information loss and protein complexes in heterogeneous network. *Information Sciences*, 479:386–400, 2019.

[31] Linton C Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[32] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[33] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[34] Jianxin Wang, Min Li, Huan Wang, and Yi Pan. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):1070–1080, 2012.

[35] Jianxin Wang, Wei Peng, and Fang-Xiang Wu. Computational approaches to predicting essential proteins: a survey. *PROTEOMICS–Clinical Applications*, 7(1-2):181–192, 2013.

[36] Muhammed A Yıldırım, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási, and Marc Vidal. Drug—target network. *Nature Biotechnology*, 25(10):1119, 2007.

[37] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[38] U Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O Woods, Inderjit S Dhillon, and Edward M Marcotte. Correction: Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PloS One*, 8(9), 2013.

[39] Gamage Upeksha Ganegoda, JianXin Wang, Fang-Xiang Wu, and Min Li. Prediction of disease genes using tissue-specified gene-gene network. *BMC Systems Biology*, 8(3):S3, 2014.

[40] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805, 2006.

[41] 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, 2012.

[42] Emad Ramadan, Sadiq Alinsaif, and Md Rafiul Hassan. Network topology measures for identifying disease-gene association in breast cancer. *BMC Bioinformatics*, 17(7):274, 2016.

[43] Ping Luo, Li-Ping Tian, Bolin Chen, Qianghua Xiao, and Fang-Xiang Wu. Predicting disease genes from clinical single sample-based PPI networks. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 247–258. Springer, 2018.

[44] 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.

[45] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–225, 1974.

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

[47] BoLin Chen, Min Li, JianXin Wang, and Fang-Xiang Wu. Disease gene identification by using graph kernels and markov random fields. *Science China Life Sciences*, 57(11):1054–1063, 2014.

[48] Bolin Chen, Min Li, Jianxin Wang, Xuequn Shang, and Fang-Xiang Wu. A fast and high performance multiple data integration algorithm for identifying human disease genes. *BMC Medical Genomics*, 8(Suppl 3):S2, 2015.

[49] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[50] Yang Yang, Leng Han, Yuan Yuan, Jun Li, Nainan Hei, and Han Liang. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, 5:3231, 2014.

[51] Lisette JA Kogelman, Susanna Cirera, Daria V Zhernakova, Merete Fredholm, Lude Franke, and Haja N Kadarmideen. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA sequencing in a porcine model. *BMC Medical Genomics*, 7(1):57, 2014.

[52] Guanming Wu and Lincoln Stein. A network module-based method for identifying cancer prognostic signatures. *Genome Biology*, 13(12):R112, 2012.

[53] Jesse Gillis and Paul Pavlidis. "guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444, 2012.

[54] Lin Hou, Min Chen, Clarence K Zhang, Judy Cho, and Hongyu Zhao. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Human Molecular Genetics*, 23(10):2780–2790, 2013.

[55] Nicholas J Hudson, Antonio Reverter, and Brian P Dalrymple. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Computational Biology*, 5(5):e1000382, 2009.

[56] David Amar, Hershel Safer, and Ron Shamir. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Computational Biology*, 9(3):e1002955, 2013.

[57] Bruno M Tesson, Rainer Breitling, and Ritsert C Jansen. Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11(1):497, 2010.

[58] Quan Wang, Hui Yu, Zhongming Zhao, and Peilin Jia. Ew_dmgwas: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics*, 31(15):2591–2594, 2015.

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

[60] Yuan Fang, Wenqing Lin, Vincent W Zheng, Min Wu, Kevin Chen-Chuan Chang, and Xiao-Li Li. Semantic proximity search on graphs with metagraph-based learning. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 277–288. IEEE, 2016.

[61] 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.

[62] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[63] Hongyi Zhou and Jeffrey Skolnick. A knowledge-based approach for predicting gene–disease associations. *Bioinformatics*, 32(18):2831–2838, 2016.

[64] Ping Luo, Yuanyuan Li, Li-Ping Tian, and Fang-Xiang Wu. Enhancing the prediction of disease-gene associations with multimodal deep learning. *Bioinformatics*, 2019. doi:10.1093/bioinformatics/btz155.

[65] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International Conference on Machine Learning*, pages 2014–2023, 2016.

[66] 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*, page 532226, 2019.

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

[68] Ping Luo, Qianghua Xiao, Pi-Jing Wei, Bo Liao, and FangXiang Wu. Identifying disease-gene associations with graph-regularized manifold learning. *Frontiers in Genetics*, 10:270, 2019.

[69] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537, 2006.

[70] Léon-Charles Tranchevent, Amin Ardeshirdavani, Sarah ElShal, Daniel Alcaide, Jan Aerts, Didier Auboeuf, and Yves Moreau. Candidate gene prioritization with endeavour. *Nucleic Acids Research*, 44(W1):W117–W121, 2016.

[71] Sinan Erten, Gurkan Bebek, Rob M Ewing, and Mehmet Koyutürk. Dada: degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, 4(1):19, 2011.

[72] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, and Lars Juhl Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015.

[73] Jeongkyun Kim, Seongeun So, Hee-Jin Lee, Jong C Park, Jung-jae Kim, and Hyunju Lee. Digsee: disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Research*, 41(W1):W510–W517, 2013.

[74] Jeongkyun Kim, Jung-jae Kim, and Hyunju Lee. An analysis of disease-gene relationship from medline abstracts by digsee. *Scientific Reports*, 7:40154, 2017.

[75] Sarah ElShal, Léon-Charles Tranchevent, Alejandro Sifrim, Amin Ardeshirdavani, Jesse Davis, and Yves Moreau. Beegle: from literature mining to disease-gene discovery. *Nucleic acids research*, 44(2):e18–e18, 2015.

[76] 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 Research*, 45(D1):D833–D839, 2016.

[77] Anthony Kulesa, Martin Krzywinski, Paul Blainey, and Naomi Altman. Sampling distributions and the bootstrap. *Nature Methods*, 12:477, 2015.

[78] Ping Luo, Li-Ping Tian, Jishou Ruan, and Fang-Xiang Wu. Disease gene prediction by integrating PPI networks, clinical RNA-seq data and OMIM data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1):222–232, 2019.

[79] Xiwei Tang, Xiaohua Hu, Xuejun Yang, and Yuan Sun. A algorithm for identifying disease genes by incorporating the subcellular localization information into the protein-protein interaction networks. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 308–311. IEEE, 2016.

[80] Ada Hamosh, Alan F Scott, Joanna Amberger, Carol Bocchini, David Valle, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, 2002.

[81] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, 2015.

[82] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. The genetic association database. *Nature Genetics*, 36(5):431, 2004.

[83] 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, 2018.

[84] Alba Gutiérrez-Sacristán, Solène Grosdidier, Olga Valverde, Marta Torrens, Àlex Bravo, Janet Piñero, Ferran Sanz, and Laura I Furlong. Psygenet: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics*, 31(18):3075–3077, 2015.

[85] Alba Gutiérrez-Sacristán, Àlex Bravo, Marta Portero-Tresserra, Olga Valverde, Antonio Armario, MC Blanco-Gandía, Adriana Farré, Lierni Fernández-Ibarrondo, Francina Fonseca, Jesús Giraldo, et al. Text mining and expert curation to develop a database on psychiatric diseases and their genes. *Database*, 2017, 2017.

[86] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177, 2004.

[87] Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. Cosmic: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2016.

[88] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

[89] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database—2009 update. *Nucleic Acids Research*, 37(suppl_1):D767–D772, 2008.

[90] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. Mint, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861, 2011.

[91] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(suppl_1):D449–D451, 2004.

[92] Michael J Meyer, Jishnu Das, Xiujuan Wang, and Haiyuan Yu. Instruct: a database of high-quality 3d structurally resolved protein interactome networks. *Bioinformatics*, 29(12):1577–1579, 2013.

[93] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, 2016.

[94] Taibo Li, Rasmus Wernersson, Rasmus B Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T Workman, Olga Rigina, Kristoffer Rapacki, Hans H Stærfeldt, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nature Methods*, 14(1):61, 2017.

[95] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, et al. The intact molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846, 2011.

[96] Mark J Cowley, Mark Pinese, Karin S Kassahn, Nic Waddell, John V Pearson, Sean M Grimmond, Andrew V Biankin, Sampsa Hautaniemi, and Jianmin Wu. Pina v2. 0: mining interactome modules. *Nucleic Acids Research*, 40(D1):D862–D865, 2011.

[97] Martin H Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E Wanker, and Miguel A Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PloS One*, 7(2):e31826, 2012.

[98] Jishnu Das and Haiyuan Yu. Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92, 2012.

[99] Sabry Razick, George Magklaras, and Ian M Donaldson. irefindex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405, 2008.

[100] Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. Mentha: a resource for browsing integrated protein-interaction networks. *Nature Methods*, 10(8):690, 2013.

[101] Kevin R Brown and Igor Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*, 8(5):R95, 2007.

[102] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[103] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2012.

[104] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[105] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[106] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 2018.

[107] Vivian G Cheung, Renuka R Nayak, Isabel Xiaorong Wang, Susannah Elwyn, Sarah M Cousins, Michael Morley, and Richard S Spielman. Polymorphic cis-and trans-regulation of human gene expression. *PLoS Biology*, 8(9):e1000480, 2010.

[108] Alkes L Price, Nick Patterson, Dustin C Hancks, Simon Myers, David Reich, Vivian G Cheung, and Richard S Spielman. Effects of cis and trans genetic ancestry on gene expression in african americans. *PLoS Genetics*, 4(12):e1000294, 2008.

[109] Xin He, Chris K Fuller, Yi Song, Qingying Meng, Bin Zhang, Xia Yang, and Hao Li. Sherlock: detecting gene-disease associations by matching patterns of expression qtl and gwas. *The American Journal of Human Genetics*, 92(5):667–680, 2013.

[110] Jun Wang, Jiashun Zheng, Zengmiao Wang, Hao Li, and Minghua Deng. Inferring gene-disease association by an integrative analysis of eqtl genome-wide association study and protein-protein interaction data. *Human Heredity*, 83(3):117–129, 2018.

[111] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2016.

[112] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[113] Yi-An Ko, Huiguang Yi, Chengxiang Qiu, Shizheng Huang, Jihwan Park, Nora Ledo, Anna Köttgen, Hongzhe Li, Daniel J Rader, Michael A Pack, et al. Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease. *The American Journal of Human Genetics*, 100(6):940–953, 2017.

[114] Ke Hao, Yohan Bossé, David C Nickle, Peter D Paré, Dirkje S Postma, Michel Laviolette, Andrew Sandford, Tillie L Hackett, Denise Daley, James C Hogg, et al. Lung eqtls to help reveal the molecular underpinnings of asthma. *PLoS Genetics*, 8(11):e1003029, 2012.

[115] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309, 2007.

[116] Laura D Wood, D Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J Leary, Dong Shen, Simina M Boca, Thomas Barber, Janine Ptak, et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113, 2007.

[117] Bolin Chen, Xuequn Shang, Min Li, Jianxin Wang, and Fang-Xiang Wu. A two-step logistic regression algorithm for identifying individual-cancer-related genes. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 195–200. IEEE, 2015.

[118] Bolin Chen, Xuequn Shang, Min Li, Jianxin Wang, and Fang-Xiang Wu. Identifying individual-cancer-related genes by rebalancing the training samples. *IEEE Transactions on Nanobioscience*, 15(4):309–315, 2016.

[119] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in kegg. *Nucleic Acids Research*, 47(D1):D590–D595, 2019.

[120] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[121] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[122] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 2017.

[123] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, et al. Wikipathways: a multi-faceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, 2018.

[124] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl_1):D685–D690, 2010.

[125] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25, 2000.

[126] Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2016.

[127] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027, 2018.

[128] Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7, 2005.

[129] Alex J Cornish, Alessia David, and Michael JE Sternberg. Phenorank: reducing study bias in gene prioritization through simulation. *Bioinformatics*, 34(12):2087–2095, 2018.

[130] 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.

[131] Maxat Kulmanov and Robert Hoehndorf. Evaluating the effect of annotation size on measures of semantic similarity. *Journal of Biomedical Semantics*, 8(1):7, 2017.

[132] 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):e97079, 2014.

[133] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[134] Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, Christian Stolte, Seán I O'Donoghue, Reinhard Schneider, and Lars Juhl Jensen. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

[135] Josefine Sprenger, J Lynn Fink, Seetha Karunaratne, Kelly Hanson, Nicholas A Hamilton, and Rohan D Teasdale. Locate: a mammalian protein subcellular localization database. *Nucleic Acids Research*, 36(suppl_1):D230–D233, 2007.

[136] Gaoshi Li, Min Li, Jianxin Wang, Jingli Wu, Fang-Xiang Wu, and Yi Pan. Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinformatics*, 17(8):279, 2016.

[137] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6(1):55, 2005.

[138] Alice Bossi and Ben Lehner. Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5(1):260, 2009.

[139] 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, 2016.

[140] Guanming Wu, Xin Feng, and Lincoln Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, 11(5):R53, 2010.

[141] Jui-Hung Hung. Gene set/pathway enrichment analysis. In *Data Mining for Systems Biology*, pages 201–213. Springer, 2013.

[142] Xiangxiang Zeng, Yinglai Lin, Yuying He, Linyuan Lv, Xiaoping Min, and Alfonso Rodríguez-Paton. Deep collaborative filtering for prediction of disease genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[143] Arnon Mazza, Konrad Klockmeier, Erich Wanker, and Roded Sharan. An integer programming framework for inferring disease complexes from network data. *Bioinformatics*, 32(12):i271–i277, 2016.

[144] Collin J Tokheim, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, 2016.

[145] Yong Mao, Han Chen, Han Liang, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen. Candra: cancer-specific driver missense mutation annotation with optimized features. *PloS One*, 8(10):e77945, 2013.

[146] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234. ACM, 2016.

[147] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2017.

[148] Chaolin Zhang and Yufeng Shen. A cell type-specific expression signature predicts haploinsufficient autism-susceptibility genes. *Human Mutation*, 38(2):204–215, 2017.

[149] Nancy J Butcher, Daniele Merico, Mehdi Zarrei, Lucas Ogura, Christian R Marshall, Eva WC Chow, Anthony E Lang, Stephen W Scherer, and Anne S Bassett. Whole-genome sequencing suggests mechanisms for 22q11. 2 deletion-associated parkinson's disease. *PloS One*, 12(4):e0173944, 2017.

[150] Peilin Jia, Siyuan Zheng, Jirong Long, Wei Zheng, and Zhongming Zhao. dmgwas: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, 27(1):95–102, 2011.

[151] Tune H Pers, Piotr Dworzyński, Cecilia Engel Thomas, Kasper Lage, and Søren Brunak. Metaranker 2.0: a web server for prioritization of genetic variation data. *Nucleic Acids Research*, 41(W1):W104–W108, 2013.

[152] Warren A Cheung, BF Francis Ouellette, and Wyeth W Wasserman. Inferring novel gene-disease associations using medical subject heading over-representation profiles. *Genome Medicine*, 4(9):75, 2012.

[153] Jean-Fred Fontaine, Florian Priller, Adriano Barbosa-Silva, and Miguel A Andrade-Navarro. Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Research*, 39(suppl 2):W455–W461, 2011.

[154] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[155] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517, 2005.

[156] Wei Zhang, Jia Xu, Yuanyuan Li, and Xiufen Zou. Detecting essential proteins based on network topology, gene expression data, and gene ontology information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 15(1):109–116, 2018.

[157] Jun Meng, Xin Zhang, and Yushi Luan. Global propagation method for predicting protein function by integrating multiple data sources. *Current Bioinformatics*, 11(2):186–194, 2016.

[158] Ping Luo, Li-Ping Tian, Jishou Ruan, and Fang-Xiang Wu. Identifying disease genes from PPI networks weighted by gene expression under different conditions. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 1259–1264. IEEE, 2016.

[159] Rui Hu, Xing Qiu, Galina Glazko, Lev Klebanov, and Andrei Yakovlev. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. *BMC Bioinformatics*, 10(1):1, 2009.

[160] Lin Hou, Min Chen, Clarence K Zhang, Judy Cho, and Hongyu Zhao. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Human Molecular Genetics*, 23(10):2780–2790, 2014.

[161] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[162] Eric Jones, Travis Oliphant, Pearu Peterson, et al. Scipy: Open source scientific tools for python, 2001. *URL http://www.scipy.org*, 73:86, 2015.

[163] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[164] Daniela Börnigen, Léon-Charles Tranchevent, Francisco Bonachela-Capdevila, Koenraad Devriendt, Bart De Moor, Patrick De Causmaecker, and Yves Moreau. An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088, 2012.

[165] Bing Zhang, Stefan Kirov, and Jay Snoddy. Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33(suppl 2):W741–W748, 2005.

[166] Jing Wang, Dexter Duncan, Zhiao Shi, and Bing Zhang. Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic Acids Research*, 41(W1):W77–W83, 2013.

[167] Jing Wang, Suhas Vasaikar, Zhiao Shi, Michael Greer, and Bing Zhang. Webgestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, 45(W1):W130–W137, 2017.

[168] Claudia Scheckel, Elodie Drapeau, Maria A Frias, Christopher Y Park, John Fak, Ilana Zucker-Scharff, Yan Kou, Vahram Haroutunian, Avi Ma'ayan, Joseph D Buxbaum, et al. Regulatory consequences of neuronal elav-like protein binding to coding and non-coding rnas in human brain. *eLife*, 5:e10421, 2016.

[169] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[170] Taibo Li, Rasmus Wernersson, Rasmus B Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T Workman, Olga Rigina, Kristoffer Rapacki, Hans H Stærfeldt, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, 14(1):61–64, 2016.

[171] Gregory A Brent. Mechanisms of thyroid hormone action. *The Journal of Clinical Investigation*, 122(9):3035–3043, 2012.

[172] Won Gu Kim and Sheue-yann Cheng. Thyroid hormone receptors and cancer. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1830(7):3928–3936, 2013.

[173] Joanna Magdalena Zarzynska. Two faces of tgf-beta1 in breast cancer. *Mediators of Inflammation*, 2014, 2014.

[174] Zeinab A Yahia, Ameera AM Adam, Magdeldin Elgizouli, Ayman Hussein, Mai A Masri, Mayada Kamal, Hiba S Mohamed, Kamal Alzaki, Ahmed M Elhassan, Kamal Hamad, et al. Epstein barr virus: a prime candidate of breast cancer aetiology in sudanese patients. *Infectious Agents and Cancer*, 9(1):9, 2014.

[175] Hai Hu, Man-Li Luo, Christine Desmedt, Sheida Nabavi, Sina Yadegarynia, Alex Hong, Panagiotis A Konstantinopoulos, Edward Gabrielson, Rebecca Hines-Boykin, German Pihan, et al. Epstein–barr virus infection of mammary epithelial cells promotes malignant transformation. *EBioMedicine*, 9:148–160, 2016.

[176] Vishnu Prasad Adhikari, Lin-Jie Lu, and Ling-Quan Kong. Does hepatitis b virus infection cause breast cancer? *Chinese Clinical Oncology*, 5(6):81, 2016.

[177] Maria Francesca Santolla, S Avino, M Pellegrino, EM De Francesco, P De Marco, R Lappano, A Vivacqua, F Cirillo, DC Rigiracciolo, A Scarpelli, et al. Sirt1 is involved in oncogenic signaling mediated by gper in breast cancer. *Cell Death & Disease*, 6(7):e1834, 2015.

[178] Berit Maria Müller, Lisa Jana, Atsuko Kasajima, Annika Lehmann, Judith Prinzler, Jan Budczies, Klaus-Jürgen Winzer, Manfred Dietel, Wilko Weichert, and Carsten Denkert. Differential expression of histone deacetylases hdac1, 2 and 3 in human breast cancer-overexpression of hdac2 and hdac3 is associated with clinicopathological indicators of disease progression. *BMC Cancer*, 13(1):215, 2013.

[179] Colleen S Sinclair, Matthew Rowley, Ali Naderi, and Fergus J Couch. The 17q23 amplicon and breast cancer. *Breast Cancer Research and Treatment*, 78(3):313–322, 2003.

[180] Dingxie Liu, Peng Hou, Zhi Liu, Guojun Wu, and Mingzhao Xing. Genetic alterations in the phosphoinositide 3-kinase/akt signaling pathway confer sensitivity of thyroid cancer cells to therapeutic targeting of akt and mammalian target of rapamycin. *Cancer Research*, 69(18):7311–7319, 2009.

[181] Laura A Marlow, Christina A von Roemeling, Simon J Cooper, Yilin Zhang, Stephen D Rohl, Shilpi Arora, Irma M Gonzales, David O Azorsa, Honey V Reddi, Han W Tun, et al. Foxo3a drives proliferation in anaplastic thyroid carcinoma through transcriptional regulation of cyclin a1: a paradigm shift that impacts current therapeutic strategies. *J Cell Sci*, 125(18):4253–4263, 2012.

[182] Enrique A Mesri, Mark A Feitelson, and Karl Munger. Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host & Microbe*, 15(3):266–282, 2014.

[183] Dimitris P Stamatiou, Stavros P Derdas, Odysseas L Zoras, and Demetrios A Spandidos. Herpes and polyoma family viruses in thyroid cancer (review). *Oncology Letters*, 11(3):1635–1644, 2016.

[184] Chi Yeon Kim, Seung Hun Lee, and Chee Won Oh. Cutaneous malignant melanoma associated with papillary thyroid cancer. *Annals of Dermatology*, 22(3):370–372, 2010.

[185] Duojia Pan. The hippo signaling pathway in development and cancer. *Developmental Cell*, 19(4):491–505, 2010.

[186] Young Keun Kim, Jeon-Soo Shin, and Moon H Nahm. Nod-like receptors in infection, immunity, and diseases. *Yonsei MJdical journal*, 57(1):5–14, 2016.

[187] Jenny Wong. Neurotrophin signaling and alzheimer's disease neurodegeneration- focus on bdnf/trkb signaling. In *Trends in Cell Signaling Pathways in Neuronal Fate Decision*. InTech, 2013.

[188] Marina Montagnani Marelli, Roberta M Moretti, Stefania Mai, Oliver Müller, Johan C Van Groeninghen, and Patrizia Limonta. Novel insights into gnrh receptor activity: role in the control of human glioblastoma cell proliferation. *Oncology Reports*, 21(5):1277–1282, 2009.

[189] Ruth F Itzhaki. Herpes simplex virus type 1 and alzheimer's disease: increasing evidence for a major role of the virus. *Frontiers in Aging Neuroscience*, 6, 2014.

[190] Yaomin Chen, Xiumei Huang, Yun-wu Zhang, Edward Rockenstein, Guojun Bu, Todd E Golde, Eliezer Masliah, and Huaxi Xu. Alzheimer's β-secretase (bace1) regulates the camp/pka/creb pathway independently of β-amyloid. *Journal of Neuroscience*, 32(33):11390–11395, 2012.

[191] Andrew F Teich, Russell E Nicholls, Daniela Puzzo, Jole Fiorito, Rosa Purgatorio, Ottavio Arancio, et al. Synaptic therapy in alzheimer's disease: a creb-centric approach. *Neurotherapeutics*, 12(1):29–41, 2015.

[192] Shinichiro Yamamoto, Teruaki Wajima, Yuji Hara, Motohiro Nishida, and Yasuo Mori. Transient receptor potential channels in alzheimer's disease. *Biochimica Et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1772(8):958–967, 2007.

[193] Rudy J Castellani, George Perry, and Mark A Smith. Prion disease and alzheimer's disease: pathogenic overlap. *Acta Neurobiologiae Experimentalis*, 64(1):11–18, 2004.

[194] Chang Liu, Guohong Cui, Meiping Zhu, Xiangping Kang, and Haidong Guo. Neuroinflammation in alzheimer's disease: chemokines produced by astrocytes and chemokine receptors. *International Journal of Clinical and Experimental Pathology*, 7(12):8342, 2014.

[195] Tiago Gil Oliveira and Gilbert Di Paolo. Phospholipase d in brain function and alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1801(8):799–805, 2010.

[196] Matthias Dumpich and Carsten Theiss. Vegf in the nervous system: an important target for research in neurodevelopmental and regenerative medicine. *Neural Regeneration Research*, 10(11):1725, 2015.

[197] Laura M Heiser, Nicholas J Wang, Carolyn L Talcott, Keith R Laderoute, Merrill Knapp, Yinghui Guan, Zhi Hu, Safiyyah Ziyad, Barbara L Weber, Sylvie Laquerre, et al. Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome Biology*, 10(3):R31, 2009.

[198] Eun-Yeong Oh, Stephen M Christensen, Sindhu Ghanta, Jong Cheol Jeong, Octavian Bucur, Benjamin Glass, Laleh Montaser-Kouhsari, Nicholas W Knoblauch, Nicholas Bertos, Sadiq MI Saleh, et al. Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biology*, 16(1):128, 2015.

[199] Choongseob Oh, Soonyong Park, Eun Kyung Lee, and Yung Joon Yoo. Downregulation of ubiquitin level via knockdown of polyubiquitin gene ubb as potential cancer therapeutic intervention. *Scientific Reports*, 3:2623, 2013.

[200] Sabina Signoretti, Lucia Di Marcotullio, Andrea Richardson, Sridhar Ramaswamy, Beth Isaac, Montserrat Rue, Franco Monti, Massimo Loda, and Michele Pagano. Oncogenic role of the ubiquitin ligase subunit skp2 in human breast cancer. *The Journal of Clinical Investigation*, 126(11):4387, 2016.

[201] Henrik J Johansson, Betzabe C Sanchez, Filip Mundt, Jenny Forshed, Aniko Kovacs, Elena Panizza, Lina Hultin-Rosenberg, Bo Lundgren, Ulf Martens, Gyöngyvér Máthé, et al. Retinoic acid receptor alpha is associated with tamoxifen resistance in breast cancer. *Nature Communications*, 4, 2013.

[202] Eugenia V Broude, Balazs Gyorffy, Alexander A Chumanevich, Mengqian Chen, Martina SJ McDermott, Michael Shtutman, James F Catroppo, and Igor B Roninson. Expression of cdk8 and cdk8-interacting genes as potential biomarkers in breast cancer. *Current Cancer Drug Targets*, 15(8):739–749, 2015.

[203] Jiajun Cui, Katherine Germer, Tianying Wu, Jiang Wang, Jia Luo, Shao-chun Wang, Qianben Wang, and Xiaoting Zhang. Cross-talk between her2 and med1 regulates tamoxifen resistance of human breast cancer cells. *Cancer Research*, 72(21):5625–5634, 2012.

[204] Susan C Pitt, Roland A Hernandez, Matthew A Nehs, Atul A Gawande, Francis D Moore, Daniel T Ruan, and Nancy L Cho. Identification of novel oncogenic mutations in thyroid cancer. *Journal of the American College of Surgeons*, 222(6):1036–1043, 2016.

[205] Subhasree Nag, Jiangjiang Qin, Kalkunte S Srivenugopal, Minghai Wang, and Ruiwen Zhang. The mdm2-p53 pathway revisited. *J Biomed Res*, 27(4):254–271, 2013.

[206] Rongrong Yang, Lei Yang, Fuman Qiu, Lisha Zhang, Hui Wang, Xiaorong Yang, Jieqiong Deng, Wenxiang Fang, Yifeng Zhou, and Jiachun Lu. Functional genetic polymorphisms in pp2a subunit genes confer increased risks of lung cancer in southern and eastern chinese. *PloS One*, 8(10):e77285, 2013.

[207] Kevin E Fisher, Jigna C Jani, Sarah B Fisher, Cora Foulks, Charles E Hill, Collin J Weber, Cynthia Cohen, and Jyotirmay Sharma. Epidermal growth factor receptor overexpression is a marker for adverse pathologic features in papillary thyroid carcinoma. *Journal of Surgical Research*, 185(1):217–224, 2013.

[208] Nancy L Cho, Chi-Iou Lin, Jinyan Du, Edward E Whang, Hiromichi Ito, Francis D Moore, and Daniel T Ruan. Global tyrosine kinome profiling of human thyroid tumors identifies src as a promising target for invasive cancers. *Biochemical and Biophysical Research Communications*, 421(3):508–513, 2012.

[209] Aniello Cerrato, Valentina De Falco, and Massimo Santoro. Molecular genetics of medullary thyroid carcinoma: the quest for novel therapeutic targets. *Journal of Molecular Endocrinology*, 43(4):143–155, 2009.

[210] Monika Ray and Weixiong Zhang. Analysis of alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. *BMC Systems Biology*, 4(1):136, 2010.

[211] Laurence Booth, Jane L Roberts, and Paul Dent. Hspa5/dna k may be a useful target for human disease therapies. *DNA and Cell Biology*, 34(3):153–158, 2015.

[212] Steven P Braithwaite, Jeffry B Stock, Paul J Lombroso, and Angus C Nairn. Protein phosphatases and alzheimer's disease. *Progress in Molecular Biology and Translational Science*, 106:343, 2012.

[213] Xue-Qiu Jian, Ke-Sheng Wang, Tie-Jian Wu, Joel J Hillhouse, and Jerald E Mullersman. Association of adam10 and camk2a polymorphisms with conduct disorder: evidence from family-based studies. *Journal of Abnormal Child Psychology*, 39(6):773, 2011.

[214] Hung-Jin Huang, Cheng-Chun Lee, and Calvin Yu-Chian Chen. Lead discovery for alzheimer's disease related target protein rbap48 from traditional chinese medicine. *BioMed Research International*, 2014, 2014.

[215] Susan E Moody, Jesse S Boehm, David A Barbie, and William C Hahn. Functional genomics and cancer drug target discovery. *Current Opinion in Molecular Therapeutics*, 12(3):284–293, 2010.

[216] Jesse Gillis and Paul Pavlidis. The impact of multifunctional genes on" guilt by association" analysis. *PloS One*, 6(2):e17258, 2011.

[217] Bolin Chen, Weiwei Fan, Juan Liu, and Fang-Xiang Wu. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics*, 15(2):177–194, 2013.

[218] Peilin Jia, Siyuan Zheng, Jirong Long, Wei Zheng, and Zhongming Zhao. dmgwas: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, 27(1):95–102, 2010.

[219] Christof Winter, Glen Kristiansen, Stephan Kersting, Janine Roy, Daniela Aust, Thomas Knösel, Petra Rümmele, Beatrix Jahnke, Vera Hentrich, Felix Rückert, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Computational Biology*, 8(5):e1002511, 2012.

[220] 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):e58977, 2013.

[221] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333, 2014.

[222] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[223] Zhengdong Lu, Berkant Savas, Wei Tang, and Inderjit S Dhillon. Supervised link prediction using multiple sources. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, pages 923–928. IEEE, 2010.

[224] Lior Pachter. Models for transcript quantification from RNA-seq. *arXiv preprint*, 1104(3889), 2011.

[225] Yixuan Chen, Wenhui Wang, Yingyao Zhou, Robert Shields, Sumit K Chanda, Robert C Elston, and Jing Li. In silico gene prioritization by integrating multiple data sources. *PloS One*, 6(6):e21137, 2011.

[226] Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(suppl_2):W305–W311, 2009.

[227] Morgan O'hayre, José Vázquez-Prado, Irina Kufareva, Eric W Stawiski, Tracy M Handel, Somasekar Seshagiri, and J Silvio Gutkind. The emerging mutational landscape of g proteins and g-protein-coupled receptors in cancer. *Nature Reviews Cancer*, 13(6):412, 2013.

[228] Anne Bruun Krøigård, Martin Jakob Larsen, Charlotte Brasch-Andersen, Anne-Vibeke Lænkholm, Ann S Knoop, Jeanette Dupont Jensen, Martin Bak, Jan Mollenhauer, Mads Thomassen, and Torben A Kruse. Genomic analyses of breast cancer progression reveal distinct routes of metastasis emergence. *Scientific Reports*, 7:43813, 2017.

[229] Hakan Uzunoglu, Tugcan Korak, Emel Ergul, Nihal Uren, Ali Sazci, N Zafer Utkan, Ertuğrul Kargi, Çağrı Triyaki, and Oktay Yirmibesoglu. Association of the nibrin gene (nbn) variants with breast cancer. *Biomedical Reports*, 4(3):369–373, 2016.

[230] Valeria Ossovskaya, Ingrid Chou Koo, Eric P Kaldjian, Christopher Alvares, and Barry M Sherman. Upregulation of poly (adp-ribose) polymerase-1 (parp1) in triple-negative breast cancer and other primary human tumor types. *Genes & Cancer*, 1(8):812–821, 2010.

[231] Annalisa Mazzotta, Giulia Partipilo, Simona De Summa, Francesco Giotta, Giovanni Simone, and Anita Mangia. Nuclear parp1 expression and its prognostic significance in breast cancer patients. *Tumor Biology*, 37(5):6143–6153, 2016.

[232] Luduo Zhang, Chun Gong, Samantha LY Lau, Nan Yang, Oscar GW Wong, Annie NY Cheung, Janice WH Tsang, Kelvin YK Chan, and Ui-Soon Khoo. Splicearray profiling of breast cancer reveals a novel variant of ncor2/smrt that is associated with tamoxifen resistance and control of erα transcriptional activity. *Cancer Research*, 73(1), 2012.

[233] David L Crowe and Roshantha AS Chandraratna. A retinoid x receptor (rxr)-selective retinoid reveals that rxr-α is potentially a therapeutic target in breast cancer cell lines, and that it potentiates antiproliferative and apoptotic responses to peroxisome proliferator-activated receptor ligands. *Breast Cancer Research*, 6(5):R546, 2004.

[234] Raghavendra A Shamanna, Huiming Lu, Deborah L Croteau, Arvind Arora, Devika Agarwal, Graham Ball, Mohammed A Aleskandarany, Ian O Ellis, Yves Pommier, Srinivasan Madhusudan, et al. Camptothecin targets wrn protein: mechanism and relevance in clinical breast cancer. *Oncotarget*, 7(12):13269, 2016.

[235] Meng Zhang, Duran Zhao, Cunye Yan, Li Zhang, and Chaozhao Liang. Associations between nine polymorphisms in exo1 and cancer susceptibility: a systematic review and meta-analysis of 39 case-control studies. *Scientific Reports*, 6:29270, 2016.

[236] Ananya Gupta, Muhammad Mosaraf Hossain, Nicola Miller, Michael Kerin, Grace Callagy, and Sanjeev Gupta. Ncoa3 coactivator is a transcriptional target of xbp1 and regulates perk–eif2α–atf4 signalling in breast cancer. *Oncogene*, 35(45):5860, 2016.

[237] Yorito Yamamoto, Aki Tsuchida, Takashi Ushiwaka, Ryuhei Nagai, Mitsuhiro Matsumoto, Junko Komatsu, Hiromi Kinoshita, Susumu Minami, and Kazutoshi Hayashi. Comparison of 4 risk-of-malignancy indexes in the preoperative evaluation of patients with pelvic masses: a prospective study. *Clinical Ovarian and Other Gynecologic Cancer*, 7(1-2):8–12, 2014.

[238] Pinki Chowdhury, Gregory E Lin, Kang Liu, Yongcheng Song, Fang-Tsyr Lin, and Weei-Chin Lin. Targeting topbp1 at a convergent point of multiple oncogenic pathways for cancer therapy. *Nature Communications*, 5:5476, 2014.

[239] Mingzhao Xing. Clinical utility of ras mutations in thyroid cancer: a blurred picture now emerging clearer. *BMC Medicine*, 14(1):12, 2016.

[240] Iacopo Petrini, Paul S Meltzer, In-Kyu Kim, Marco Lucchi, Kang-Seo Park, Gabriella Fontanini, James Gao, Paolo A Zucali, Fiorella Calabrese, Adolfo Favaretto, et al. A specific missense mutation in gtf2i occurs at high frequency in thymic epithelial tumors. *Nature Genetics*, 46(8):844, 2014.

[241] Francesca Galdiero, Anna Maria Bello, Anna Spina, Anna Capiluongo, Sophie Liuu, Margot De Marco, Alessandra Rosati, Mario Capunzo, Maria Napolitano, Emilia Vuttariello, et al. Identification of bag3 target proteins in anaplastic thyroid cancer cells by proteomic analysis. *Oncotarget*, 9(8):8016, 2018.

[242] Massimo Santoro and Francesca Carlomagno. Central role of ret in thyroid cancer. *Cold Spring Harbor Perspectives in Biology*, 5(12):a009233, 2013.

[243] Daniela Bossi, Francesca Carlomagno, Isabella Pallavicini, Giancarlo Pruneri, Maurizio Trubia, Paola Rafaniello Raviele, Alessandra Marinelli, Suresh Anaganti, Maria Christina Cox, Giuseppe Viale, et al. Functional characterization of a novel fgfr1op-ret rearrangement in hematopoietic malignancies. *Molecular Oncology*, 8(2):221–231, 2014.

[244] M Papadakis, A Meyer, F Schuster, N Weyerbrock, C Corinth, and C Dotzenrath. Follicular variant of papillary thyroid cancer in alström syndrome. *Familial Cancer*, 14(4):599–602, 2015.

[245] Ce Xie and Tomohiro Miyasaka. The role of the carboxyl-terminal sequence of tau and map2 in the pathogenesis of dementia. *Frontiers in Molecular Neuroscience*, 9:158, 2016.

[246] AJ Russo. Decreased mitogen inducible gene 6 (mig-6) associated with symptom severity in children with autism. *Biomarker Insights*, 9:BMI–S15218, 2014.

[247] Lei Song, Yue Gu, Jing Jie, Xiaoxue Bai, Ying Yang, Chaoying Liu, and Qun Liu. Dab2 attenuates brain injury in app/ps1 mice via targeting transforming growth factor-beta/smad signaling. *Neural Regeneration Research*, 9(1):41, 2014.

[248] Héctor J De Jesús-Cortés, Carlos J Nogueras-Ortiz, Marla Gearing, Steven E Arnold, and Irving E Vega. Amphiphysin-1 protein level changes associated with tau-mediated neurodegeneration. *Neuroreport*, 23(16):942, 2012.

[249] Sarah M Neuner, Lynda A Wilmott, Brian R Hoffmann, Khyobeni Mozhui, and Catherine C Kaczorowski. Hippocampal proteomics defines pathways associated with memory decline and resilience in normal aging and alzheimer's disease mouse models. *Behavioural Brain Research*, 322:288–298, 2017.

[250] Vivek Gautam, Carla D'Avanzo, Oksana Berezovska, Rudolph E Tanzi, and Dora M Kovacs. Synaptotagmins interact with app and promote aβ generation. *Molecular Neurodegeneration*, 10(1):31, 2015.

[251] Weiwei Zhang, Bin Jiao, Tingting Xiao, Chuzheng Pan, Xixi Liu, Lin Zhou, Beisha Tang, and Lu Shen. Mutational analysis of prnp in alzheimer's disease and frontotemporal dementia in china. *Scientific Reports*, 6:38435, 2016.

[252] Xue Fu, Meiling Ke, Weihua Yu, Xia Wang, Qian Xiao, Min Gu, and Yang Lü. Periodic variation of aak1 in an aβ 1–42-induced mouse model of alzheimer's disease. *Journal of Molecular Neuroscience*, 65:179–189, 2018.

[253] Mengmeng Wu, Wanwen Zeng, Wenqiang Liu, Yijia Zhang, Ting Chen, and Rui Jiang. Integrating embeddings of multiple gene networks to prioritize complex disease-associated genes. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 208–215. IEEE, 2017.

[254] Insuk Lee, U Martin Blom, Peggy I Wang, Jung Eun Shim, and Edward M Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121, 2011.

[255] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519, 2012.

[256] Teresa Davoli, Andrew Wei Xu, Kristen E Mengwasser, Laura M Sack, John C Yoon, Peter J Park, and Stephen J Elledge. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962, 2013.

[257] Xiangxiang Zeng, Ningxiang Ding, Alfonso Rodríguez-Patón, and Quan Zou. Probability-based collaborative filtering model for predicting gene–disease associations. *BMC Medical Genomics*, 10(S5):76, 2017.

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

[259] 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, 2016.

[260] Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[261] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of Proteome Research*, 16(4):1401–1409, 2017.

[262] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 12(4):928–937, 2015.

[263] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Tech. Rep. Computer Science Department, University of Toronto*, 1(4), 2009.

[264] KyungHyun Cho, Alexander Ilin, and Tapani Raiko. Improved learning of gaussian-bernoulli restricted boltzmann machines. In *International Conference on Artificial Neural Networks*, pages 10–17. Springer, 2011.

[265] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009.

[266] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[267] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine learning*, pages 1064–1071. ACM, 2008.

[268] KyungHyun Cho, Tapani Raiko, and Alexander Ilin. Parallel tempering is efficient for learning restricted boltzmann machines. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.

[269] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[270] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, pages 153–160, 2007.

[271] 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):1257601, 2015.

[272] Peng Ni, Jianxin Wang, Ping Zhong, Yaohang Li, Fangxiang Wu, and Yi Pan. Constructing disease similarity networks based on disease module theory. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018. doi:10.1109/TCBB.2018.2817624.

[273] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade*, pages 599–619. Springer, 2012.

[274] Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. OMIM.org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, 2014.

[275] Jing Zou, Xiangqiang Duan, Guiliang Zheng, Zhen Zhao, Shiyue Chen, Pu Dai, and Hongliang Zheng. A novel pik3cd c896t mutation detected in bilateral sudden sensorineural hearing loss using next generation sequencing: An indication of primary immunodeficiency. *Journal of Otology*, 11(2):78–83, 2016.

[276] Magali Avila, David A Dyment, Jørn V Sagen, Judith St-Onge, Ute Moog, Brian HY Chung, S Mo, S Mansour, A Albanese, S Garcia, et al. Clinical reappraisal of short syndrome with pik3r1 mutations: toward recommendation for molecular testing and management. *Clinical Genetics*, 89(4):501–506, 2016.

[277] I-Chen Yu, Hung-Yun Lin, Janet D Sparks, Shuyuan Yeh, and Chawnshang Chang. Androgen receptor roles in insulin resistance and obesity in males: the linkage of androgen-deprivation therapy to metabolic syndrome. *Diabetes*, 63(10):3180–3188, 2014.

[278] G Bademci, FB Cengiz, J Foster II, D Duman, L Sennaroglu, O Diaz-Horta, T Atik, T Kirazli, L Olgun, H Alper, et al. Variations in multiple syndromic deafness genes mimic non-syndromic hearing loss. *Scientific Reports*, 6:31622, 2016.

[279] Donghee Kim, Song Mi Lee, and Hee-Sook Jun. Impact of t-cell-specific smad4 deficiency on the development of autoimmune diabetes in nod mice. *Immunology & Cell Biology*, 95(3):287–296, 2017.

[280] Koji Muroya, Takahiro Mochizuki, Maki Fukami, Manami Iso, Keinosuke Fujita, Mitsuo Itakura, and Tsutomu Ogata. Diabetes mellitus in a japanese girl with hdr syndrome and gata3 mutation. *Endocrine Journal*, 57(2):171–174, 2010.

[281] Viviana Caputo, Luciano Cianetti, Marcello Niceta, Claudio Carta, Andrea Ciolfi, Gianfranco Bocchinfuso, Eugenio Carrani, Maria Lisa Dentici, Elisa Biamino, Elga Belligni, et al. A restricted spectrum of mutations in the smad4 tumor-suppressor gene underlies myhre syndrome. *The American Journal of Human Genetics*, 90(1):161–169, 2012.

[282] Teresa Wilson, Irina Omelchenko, Sarah Foster, Yuan Zhang, Xiaorui Shi, and Alfred L Nuttall. Jak2/stat3 inhibition attenuates noise-induced hearing loss. *PLoS One*, 9(10):e108276, 2014.

[283] Lucie C Kompier, Irene Lurkin, Madelon NM van der Aa, Bas WG van Rhijn, Theo H van der Kwast, and Ellen C Zwarthoff. Fgfr3, hras, kras, nras and pik3ca mutations in bladder cancer and their potential as biomarkers for surveillance and therapy. *PloS One*, 5(11):e13821, 2010.

[284] María Alba-Domínguez, Alberto López-Lera, Sofía Garrido, Pilar Nozal, Ignacio González-Granado, Josefa Melero, Pere Soler-Palacín, Carmen Cámara, and Margarita López-Trascasa. Complement factor i deficiency: a not so rare immune defect. characterization of new mutations and the first large gene deletion. *Orphanet Journal of Rare Diseases*, 7(1):42, 2012.

[285] Katharina Hopp, Christina M Heyer, Cynthia J Hommerding, Susan A Henke, Jamie L Sundsbak, Shail Patel, Priyanka Patel, Mark B Consugar, Peter G Czarnecki, Troy J Gliem, et al. B9d1 is revealed as a novel meckel syndrome (mks) gene by targeted exon-enriched next-generation sequencing and deletion analysis. *Human Molecular Genetics*, 20(13):2524–2534, 2011.

[286] Paweł Stankiewicz, Tahir N Khan, Przemyslaw Szafranski, Leah Slattery, Haley Streff, Francesco Vetrini, Jonathan A Bernstein, Chester W Brown, Jill A Rosenfeld, Surya Rednam, et al. Haploinsufficiency of the chromatin remodeler bptf causes syndromic developmental and speech delay, postnatal microcephaly, and dysmorphic features. *The American Journal of Human Genetics*, 101(4):503–515, 2017.

[287] Fatma Mujgan Sonmez, Eyyup Uctepe, Dilek Aktas, and Mehmet Alikasifoglu. Microdeletion of chromosome 1q21. 3 in fraternal twins is associated with mental retardation, microcephaly, and epilepsy. *Intractable & Rare Diseases Research*, 6(1):61–64, 2017.

[288] Lam Son Nguyen, Taiane Schneider, Marlène Rio, Sébastien Moutton, Karine Siquier-Pernet, Florine Verny, Nathalie Boddaert, Isabelle Desguerre, Arnold Munich, José Luis Rosa, et al. A nonsense variant in herc1 is associated with intellectual disability, megalencephaly, thick corpus callosum and cerebellar atrophy. *European Journal of Human Genetics*, 24(3):455, 2016.

[289] Anirudh Prahallad, Guus JJE Heynen, Giovanni Germano, Stefan M Willems, Bastiaan Evers, Loredana Vecchione, Valentina Gambino, Cor Lieftink, Roderick L Beijersbergen, Federica Di Nicolantonio, et al. Ptpn11 is a central node in intrinsic and acquired resistance to targeted cancer drugs. *Cell Reports*, 12(12):1978–1985, 2015.

[290] Lydia WT Cheung and Gordon B Mills. Targeting therapeutic liabilities engendered by pik3r1 mutations for cancer treatment. *Pharmacogenomics*, 17(3):297–307, 2016.

[291] Michael K Kiessling, Alessandra Curioni-Fontecedro, Panagiotis Samaras, Kirstin Atrott, Jesus Cosin-Roger, Silvia Lang, Michael Scharl, and Gerhard Rogler. Mutant hras as novel target for mek and mtor inhibitors. *Oncotarget*, 6(39):42183, 2015.

[292] Markku Miettinen, Peter A Mc Cue, Maarit Sarlomo-Rikala, Janusz Rys, Piotr Czapiewski, Krzysztof Wazny, Renata Langfort, Piotr Waloszczyk, Wojciech Biernat, Jerzy Lasota, et al. Gata 3–a multispecific but potentially useful marker in surgical pathology–a systematic analysis of 2500 epithelial and non-epithelial tumors. *The American Journal of Surgical Pathology*, 38(1):13, 2014.

[293] Yanjun Xu, Juan Jin, Jiawei Xu, Yang W Shao, and Yun Fan. Jak2 variations and functions in lung adenocarcinoma. *Tumor Biology*, 39(6):1010428317711140, 2017.

[294] Beatrice Grabner, Daniel Schramek, Kristina M Mueller, Herwig P Moll, Jasmin Svinka, Thomas Hoffmann, Eva Bauer, Leander Blaas, Natascha Hruschka, Katalin Zboray, et al. Disruption of stat3 signalling promotes kras-induced lung tumorigenesis. *Nature Communications*, 6:6285, 2015.

[295] Ruben Pio, Leticia Corrales, and John D Lambris. The role of complement in tumor growth. In *Tumor Microenvironment and Cellular Stress*, pages 229–262. Springer, 2014.

[296] Yuan-Hu Yao, Yan Cui, Xiang-Nan Qiu, Long-Zhen Zhang, Wei Zhang, Hao Li, and Jin-Ming Yu. Attenuated lkb1-sik1 signaling promotes epithelial-mesenchymal transition and radioresistance of non–small cell lung cancer cells. *Chinese Journal of Cancer*, 35(1):50, 2016.

[297] Artur Zajkowicz, Dorota Butkiewicz, A Drosik, Monika Giglok, Rafał Suwiński, and Marek Rusin. Truncating mutations of ppm1d are found in blood dna samples of lung cancer patients. *British Journal of Cancer*, 112(6):1114, 2015.

[298] David Tamborero, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244, 2013.

[299] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495, 2014.

[300] Jack P Hou and Jian Ma. Dawnrank: discovering personalized driver genes in cancer. *Genome medicine*, 6(7):56, 2014.

[301] Wei-Feng Guo, Shao-Wu Zhang, Li-Li Liu, Fei Liu, Qian-Qian Shi, Lei Zhang, Ying Tang, Tao Zeng, and Luonan Chen. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics*, 34(11):1893–1903, 2018.

[302] Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, et al. The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379, 2017.

[303] Wing Chung Wong, Dewey Kim, Hannah Carter, Mark Diekhans, Michael C Ryan, and Rachel Karchin. Chasm and snvbox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, 27(15):2147–2148, 2011.

[304] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073, 2009.

[305] Jüri Reimand and Gary D Bader. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology*, 9(1):637, 2013.

[306] Nathan D Dees, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, Daniel C Koboldt, Thomas B Mooney, Matthew B Callaway, David Dooling, Elaine R Mardis, et al. Music: identifying mutational significance in cancer genomes. *Genome research*, 2012.

[307] Abel Gonzalez-Perez and Nuria Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic acids research*, 40(21):e169–e169, 2012.

[308] Loris Mularoni, Radhakrishnan Sabarinathan, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas. Oncodrivefml: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*, 17(1):128, 2016.

[309] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[310] Bethan Yates, Bryony Braschi, Kristian A Gray, Ruth L Seal, Susan Tweedie, and Elspeth A Bruford. Genenames. org: the hgnc and vgnc resources in 2017. *Nucleic acids research*, page gkw1033, 2016.

[311] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[312] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[313] François Chollet et al. Keras. `https://keras.io`, 2015.

[314] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[315] Sangjun Lee, Sheila Stewart, Iris Nagtegaal, Jingqin Luo, Yun Wu, Graham Colditz, Dan Medina, and D Craig Allred. Differentially expressed genes regulating the progression of ductal carcinoma in situ to invasive breast cancer. *Cancer research*, 72(17):4574–4586, 2012.

[316] Rebeca Sanz-Pamplona, Adriana Lopez-Doriga, Laia Paré-Brunet, Kira Lázaro, Fernando Bellido, M Henar Alonso, Susanna Aussó, Elisabet Guinó, Sergi Beltrán, Francesc Castro-Giner, et al. Exome sequencing reveals amer1 as a frequently mutated gene in colorectal cancer. *Clinical Cancer Research*, 21(20):4709–4718, 2015.

[317] Bon-Hun Koo, Tiina Hurskainen, Katrina Mielke, Phyu Phyu Aung, Graham Casey, Helena Autio-Harmainen, and Suneel S Apte. Adamtsl3/punctin-2, a gene frequently mutated in colorectal tumors, is widely expressed in normal and malignant epithelial cells, vascular endothelial cells and other cell types, and its mrna is reduced in colon cancer. *International journal of cancer*, 121(8):1710–1716, 2007.

[318] Mi Ryoung Choi, Chang Hyeok An, Nam Jin Yoo, and Sug Hyung Lee. Laminin gene lamb 4 is somatically mutated and expressionally altered in gastric and colorectal cancers. *Apmis*, 123(1):65–71, 2015.

[319] Rajesh C Rao and Yali Dou. Hijacked in cancer: the kmt2 (mll) family of methyltransferases. *Nature Reviews Cancer*, 15(6):334, 2015.

[320] KR Velmurugan, RT Varghese, NC Fonville, and HR Garner. High-depth, high-accuracy microsatellite genotyping enables precision lung cancer risk classification. *Oncogene*, 36(46):6383, 2017.

[321] Petros Kechagioglou, Rigini M Papi, Xeni Provatopoulou, Eleni Kalogera, Elli Papadimitriou, Petros Grigoropoulos, Aphroditi Nonni, George Zografos, Dimitrios A Kyriakidis, and Antonia Gounaris. Tumor suppressor pten in breast cancer: heterozygosity, mutations and protein expression. *Anticancer research*, 34(3):1387–1400, 2014.

[322] Carlota Rubio-Perez, David Tamborero, Michael P Schroeder, Albert A Antolín, Jordi Deu-Pons, Christian Perez-Llamas, Jordi Mestres, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer cell*, 27(3):382–396, 2015.

[323] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. Intogen-mutations identifies cancer drivers across tumor types. *Nature methods*, 10(11):1081, 2013.

[324] Stéphanie Cornen, Arnaud Guille, José Adélaïde, Lynda Addou-Klouche, Pascal Finetti, Marie-Rose Saade, Marwa Manai, Nadine Carbuccia, Ismahane Bekhouche, Anne Letessier, et al. Candidate luminal b breast cancer genes identified by genome, gene expression and dna methylation profiling. *PLoS One*, 9(1):e81843, 2014.

[325] Xin Yi Loh, Ling Wen Ding, and H Phillip Koeffler. Tumor suppressive role of zfp36l1 by suppressing hif1α and cyclin d1 in bladder and breast cancer. In *AACR Annual Meeting 2017*, volume 77, page Abstract nr 4494. AACR, 2017.

[326] Jun Cao, Ming-Hua Ge, and Zhi-Qiang Ling. Fbxw7 tumor suppressor: A vital regulator contributes to human tumorigenesis. *Medicine*, 95(7), 2016.

[327] Corinne Prévostel and Philippe Blache. The dose-dependent effect of sox9 and its incidence in colorectal cancer. *European Journal of Cancer*, 86:150–157, 2017.

[328] Fausto Meriggi, William Vermi, Paola Bertocchi, and Alberto Zaniboni. The emerging role of nras mutations in colorectal cancer patients selected for anti-egfr therapies. *Reviews on recent clinical trials*, 9(1):8–12, 2014.

[329] Xiao-Wen Wang and Yan-Jie Zhang. Targeting mtor network in colorectal cancer therapy. *World Journal of Gastroenterology: WJG*, 20(15):4178, 2014.

[330] Saud H AlDubayan, Marios Giannakis, Nathanael D Moore, G Celine Han, Brendan Reardon, Tsuyoshi Hamada, Xinmeng Jasmine Mu, Reiko Nishihara, Zhirong Qian, Li Liu, et al. Inherited dna-repair defects in colorectal cancer. *The American Journal of Human Genetics*, 102(3):401–414, 2018.

[331] Xiao-bin Zheng, Chi Zhou, Hai-chun Cheng, Tuo Hu, Hua-shan Liu, Xuan-hui Liu, Xian-rui Wu, Feng-wei Wang, Yu-feng Chen, Jian-ping Wang, et al. Elmo1 promotes metastasis in colorectal cancer

cells via activation of mapk/erk signaling pathway. In *AACR Annual Meeting 2017*, volume 77, page Abstract nr 4849. AACR, 2017.

[332] Haoyou Wang, Qiming Shen, Xin Zhang, Chunlu Yang, Su Cui, Yanbin Sun, Liming Wang, Xiaoxi Fan, and Shun Xu. The long non-coding RNA xist controls non-small cell lung cancer proliferation and invasion by modulating mir-186-5p. *Cellular Physiology and Biochemistry*, 41(6):2221–2229, 2017.

[333] Sen Li, Zhoufang Mei, Hai-Bo Hu, and Xin Zhang. The lncrna malat1 contributes to non-small cell lung cancer development via modulating mir-124/stat3 axis. *Journal of cellular physiology*, 233(9):6679–6688, 2018.

[334] Nicolas Pécuchet, Pierre Laurent-Puig, Audrey Mansuet-Lupo, Antoine Legras, Marco Alifano, Karine Pallier, Audrey Didelot, Laure Gibault, Claire Danel, Pierre-Alexandre Just, et al. Different prognostic impact of stk11 mutations in non-squamous non-small-cell lung cancer. *Oncotarget*, 8(14):23831, 2017.

[335] Sarah M Haeger, Joshua J Thompson, Sean Kalra, Timothy G Cleaver, Daniel Merrick, Xiao-Jing Wang, and Stephen P Malkoski. Smad4 loss promotes lung cancer formation but increases sensitivity to dna topoisomerase inhibitors. *Oncogene*, 35(5):577, 2016.

[336] Kinisha Gala, Qing Li, Amit Sinha, Pedram Razavi, Madeline Dorso, Francisco Sanchez-Vega, Young Rock Chung, Ronald Hendrickson, James J Hsieh, Michael Berger, et al. Kmt2c mediates the estrogen dependence of breast cancer through regulation of erα enhancer function. *Oncogene*, page 1, 2018.

[337] Elina Uusitalo, Roope A Kallionpää, Samu Kurki, Matti Rantanen, Janne Pitkäniemi, Pauliina Kronqvist, Pirkko Härkönen, Riikka Huovinen, Olli Carpen, Minna Pöyhönen, et al. Breast cancer in neurofibromatosis type 1: overrepresentation of unfavourable prognostic factors. *British journal of cancer*, 116(2):211, 2017.

[338] Shinji Kikuchi, Daisuke Yamada, Takeshi Fukami, Mari Masuda, Mika Sakurai-Yageta, Yuko N Williams, Tomoko Maruyama, Hisao Asamura, Yoshihiro Matsuno, Masataka Onizuka, et al. Promoter methylation of dal-1/4.1 b predicts poor prognosis in non–small cell lung cancer. *Clinical Cancer Research*, 11(8):2954–2961, 2005.

[339] Amanda J Redig, Marzia Capelletti, Suzanne E Dahlberg, Lynette M Sholl, Stacy L Mach, Caitlin Fontes, Yunling Shi, Poornima Chalasani, and Pasi A Janne. Clinical and molecular characteristics of nf1 mutant lung cancer. *Clinical Cancer Research*, pages clincanres–2377, 2016.

[340] Kazufumi Honda. The biological role of actinin-4 (actn4) in malignant phenotypes of cancer. *Cell & bioscience*, 5(1):41, 2015.

[341] Ping He, Ke Li, Shi-Bao Li, Ting-Ting Hu, Ming Guan, Fen-Yong Sun, and Wei-Wei Liu. Upregulation of akap12 with hdac3 depletion suppresses the progression and migration of colorectal cancer. *International journal of oncology*, 52(4):1305–1316, 2018.

[342] Yifan Yu, Dongliang Liu, Zhenghao Liu, Shuqiang Li, Yang Ge, Wei Sun, and Baolin Liu. The inhibitory effects of col1a2 on colorectal cancer cell proliferation, migration, and invasion. *Journal of Cancer*, 9(16):2953, 2018.

[343] Teresa Friedrich, Michaela Söhn, Tobias Gutting, Klaus-Peter Janssen, Hans-Michael Behrens, Christoph Röcken, Matthias PA Ebert, and Elke Burgermeister. Subcellular compartmentalization of docking protein-1 contributes to progression in colorectal cancer. *EBioMedicine*, 8:159–172, 2016.

[344] J Yu, WKK Wu, Q Liang, N Zhang, J He, X Li, X Zhang, L Xu, MTV Chan, SSM Ng, et al. Disruption of ncoa2 by recurrent fusion with lactb2 in colorectal cancer. *Oncogene*, 35(2):187, 2016.

[345] Ping Luo, Yulian Ding, Xiujuan Lei, and Fang-Xiang Wu. deepdriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Frontiers in Genetics*, 10:13, 2019.

[346] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[347] Jihun Ham, Daniel D Lee, and Lawrence K Saul. Semisupervised alignment of manifolds. In *AISTATS*, pages 120–127, 2005.

[348] Min Chen, Xingguo Lu, Bo Liao, Zejun Li, Lijun Cai, and Changlong Gu. Uncover mirna-disease association by exploiting global network similarity. *PloS One*, 11(12):e0166509, 2016.

[349] Min Chen, Bo Liao, and Zejun Li. Global similarity method based on a two-tier random walk for the prediction of microrna–disease association. *Scientific Reports*, 8, 2018.

[350] Min Chen, Yan Peng, Ang Li, Zejun Li, Yingwei Deng, Wenhua Liu, Bo Liao, and Chengqiu Dai. A novel information diffusion method based on network consistency for identifying disease related micrornas. *RSC Advances*, 8(64):36675–36690, 2018.

[351] Marianna Bolla. *Spectral clustering and biclustering: Learning large graphs and contingency tables*. John Wiley & Sons, 2013.

[352] Jiang Li, Binsheng Gong, Xi Chen, Tao Liu, Chao Wu, Fan Zhang, Chunquan Li, Xiang Li, Shaoqi Rao, and Xia Li. Dosim: an r package for similarity between diseases based on disease ontology. *BMC Bioinformatics*, 12(1):266, 2011.

[353] Han Sun, Liqun Luo, Bachchu Lal, Xinrong Ma, Lieping Chen, Christine L Hann, Amy M Fulton, Daniel J Leahy, John Laterra, and Min Li. A monoclonal antibody against kcnk9 k+ channel extracellular domain inhibits tumour growth and metastasis. *Nature Communications*, 7:10339, 2016.

[354] Vincenzo Esposito, Mara Campioni, Antonio De Luca, Enrico P Spugnini, Feliciano Baldi, Roberto Cassandro, Alessandro Mancini, Bruno Vincenzi, Angela Groeger, Mario Caputi, et al. Analysis of htra1 serine protease expression in human lung cancer. *Anticancer Research*, 26(5A):3455–3459, 2006.

[355] Siamack Sabrkhany, Marijke JE Kuijpers, Jaco C Knol, Steven WM Olde Damink, Anne-Marie C Dingemans, Henk M Verheul, Sander R Piersma, Thang V Pham, Arjan W Griffioen, Mirjam GA oude Egbrink, et al. Exploration of the platelet proteome in patients with early-stage cancer. *Journal of Proteomics*, 177:65–74, 2018.

[356] Chun-Li Che, Yi-Mei Zhang, Hai-Hong Zhang, Yu-Lan Sang, Ben Lu, Fu-Shi Dong, Li-Juan Zhang, and Fu-Zhen Lv. Dna microarray reveals different pathways responding to paclitaxel and docetaxel in non-small cell lung cancer cell line. *International Journal of Clinical and Experimental Pathology*, 6(8):1538, 2013.

[357] Jerry McLarty, Yixuang Ma, Mylinh Smith, and Jonathan Glass. Iron metabolism and the risk of lung and head and neck cancers. In *AACR Annual Meeting*, volume 68, pages 3923–3923. AACR, 2008.

[358] Apoorva Iyer and Svetlana Chapoval. Neuroimmune semaphorin 4a in cancer angiogenesis and inflammation: A promoter or a suppressor? *International Journal of Molecular Sciences*, 20(1):124, 2019.

[359] Hang Tong, Hubin Yin, Mohammad Arman Hossain, Yiyang Wang, Feixiang Wu, Xiaoyong Dong, Shun Gao, Kai Zhan, and Weiyang He. Starvation-induced autophagy promotes the invasion and migration of human bladder cancer cells via tgf-β1/smad3-mediated epithelial-mesenchymal transition activation. *Journal of Cellular Biochemistry*, 120(4):5118–5127, 2018.

[360] Yang Peng, Wen Dong, Tian-xin Lin, Guang-zheng Zhong, Bei Liao, Bo Wang, Peng Gu, Li Huang, Yun Xie, Fu-ding Lu, et al. Microrna-155 promotes bladder cancer growth by repressing the tumor suppressor dmtf1. *Oncotarget*, 6(18):16043, 2015.

[361] Susumu Kageyama, Takahiro Isono, Hideaki Iwaki, Yoshihiko Wakabayashi, Yusaku Okada, Keiichi Kontani, Koji Yoshimura, Akito Terai, Yoichi Arai, and Tatsuhiro Yoshiki. Identification by proteomic analysis of calreticulin as a marker for bladder cancer and evaluation of the diagnostic accuracy of its detection in urine. *Clinical Chemistry*, 50(5):857–866, 2004.

[362] Yifeng Li, François Fauteux, Jinfeng Zou, André Nantel, and Youlian Pan. Personalized prediction of genes with tumor-causing somatic mutations based on multi-modal deep boltzmann machine. *Neurocomputing*, 324:51–62, 2019.

Appendix A

List of Publications

Refereed journal publications:

1. **Ping Luo**, Yuanyuan Li, Li-Ping Tian, and Fang-Xiang Wu. Enhancing the prediction of disease-gene associations with multimodal deep learning. Bioinformatics, 2019. doi:10.1093/bioinformatics/btz155.

2. **Ping Luo**, Qianghua Xiao, Pi-Jing Wei, Bo Liao, and Fang-Xiang Wu. Identifying disease-gene associations with graph-regularized manifold learning. Frontiers in Genetics, 10:270, 2019.

3. **Ping Luo**, Yulian Ding, Xiujuan Lei, and Fang-Xiang Wu. deepdriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. Frontiers in Genetics, 10:13, 2019.

4. **Ping Luo**, Li-Ping Tian, Jishou Ruan, and Fang-Xiang Wu. Disease gene prediction by integrating PPI networks, clinical RNA-seq data and OMIM data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(1):222-232, 2019.

5. Li-Ping Tian #, **Ping Luo** #, Haiying Wang, Huiru Zheng, and Fang-Xiang Wu. CASNMF: A converged algorithm for symmetrical nonnegative matrix factorization. Neurocomputing, 275:2031-2040, 2018. co-first author.

6. **Ping Luo**, Li-Ping Tian, Bolin Chen, Qianghua Xiao, and Fang-Xiang Wu. Ensemble disease gene prediction by clinical sample-based networks. BMC Bioinformatics, accepted, 2019.

7. Qianghua Xiao, **Ping Luo**, Min Li, Jianxin Wang, Fang-Xiang Wu. A novel core-attachment based method to identify dynamic protein complexes based on gene expression profiles and PPI Network. Proteomics, 19(5):1800129, 2019.

Refereed conference publications:

1. **Ping Luo**, Li-Ping Tian, Bolin Chen, Qianghua Xiao, and Fang-Xiang Wu. Predicting gene-disease associations with manifold learning. In International Symposium on Bioinformatics Research and Applications, pages 265-271. Springer, 2018.

2. **Ping Luo**, Li-Ping Tian, Bolin Chen, Qianghua Xiao, and Fang-Xiang Wu. Predicting disease genes from clinical single sample-based PPI networks. In International Conference on Bioinformatics and Biomedical Engineering, pages 247-258. Springer, 2018.

3. **Ping Luo**, Li-Ping Tian, Jishou Ruan, and Fang-Xiang Wu. Identifying disease genes from PPI networks weighted by gene expression under different conditions. In Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, pages 1259-1264. IEEE, 2016.

4. Yuanyuan Li, **Ping Luo**, Yi Lu, and Fang-Xiang Wu. Improved spectral clustering method for identifying cell types from single-cell data. In 15th International Conference on Intelligent Computing (ICIC), accepted, Nanchang, 2019.

Under revision:

1. **Ping Luo**, Bo Liao, and Fang-Xiang Wu. Predicting disease-associated genes: computational methods, databases, and evaluations. WIREs: Data Mining and Knowledge Discovery, under revision, 2019.

Appendix B

Copyright Permissions

Copyright forms of thesis-related publications are attached in the following pages.

Copyright Clearance Center RightsLink®

Home | Account Info | Help

IEEE Requesting permission to reuse content from an IEEE publication

Title: Disease Gene Prediction by Integrating PPI Networks, Clinical RNA-Seq Data and OMIM Data

Author: Ping Luo

Publication: Computational Biology and Bioinformatics, IEEE/ACM Transactions on

Publisher: IEEE

Date: 1 Jan.-Feb. 2019

Copyright © 2019, IEEE

Logged in as:
Ping Luo
University of Saskatchewan
Account #: 3001480237

LOGOUT

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK | CLOSE WINDOW

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Copyright Clearance Center RightsLink®

Home | Account Info | Help

OXFORD UNIVERSITY PRESS

Title: Enhancing the prediction of disease–gene associations with multimodal deep learning
Author: Luo, Ping; Li, Yuanyuan
Publication: Bioinformatics
Publisher: Oxford University Press
Date: 2019-03-02
Copyright © 2019, Oxford University Press

Logged in as:
Ping Luo
University of Saskatchewan
Account #: 3001480237
LOGOUT

Order Completed

Thank you for your order.

This Agreement between University of Saskatchewan -- Ping Luo ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

printable details

License Number	4678300315915
License date	Sep 29, 2019
Licensed Content Publisher	Oxford University Press
Licensed Content Publication	Bioinformatics
Licensed Content Title	Enhancing the prediction of disease–gene associations with multimodal deep learning
Licensed Content Author	Luo, Ping; Li, Yuanyuan
Licensed Content Date	Mar 2, 2019
Licensed Content Volume	35
Licensed Content Issue	19
Type of Use	Thesis/Dissertation
Requestor type	Author of this OUP content
Format	Electronic
Portion	Text Extract
Number of pages requested	8
Will you be translating?	No
Title	Identifying disease-associated genes by multimodal deep learning
Institution name	University of Saskatchewan
Expected presentation date	Oct 2019

Portions	full paper
Requestor Location	University of Saskatchewan 57 Campus Dr. Saskatoon, SK S7N 5A9 Canada Attn: University of Saskatchewan
Publisher Tax ID	GB125506730
Total	0.00 CAD

ORDER MORE CLOSE WINDOW

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com