Vietnam Journal of Science and Technology ***60*** *(4) (2022) 700-712*

doi:10.15625/2525-2518/17026

NETWORK APPROACHES FOR IDENTIFICATION OF HUMAN GENETIC DISEASE GENES

Minh-Tan Nguyen[1, 3], Tien-Dzung Tran[2, 3, *]

[1]*Center of Information - Library, Hanoi University of Industry, 298 Cau Dien Street, Bac Tu Liem District, Ha Noi, Viet Nam*

[2]*Department of Software Engineering, Faculty of Information Technology, Hanoi University of Industry, 298 Cau Dien Street, Bac Tu Liem District, Ha Noi, Viet Nam*

[3]*Complex Systems and Bioinformatics Lab, Hanoi University of Industry, 298 Cau Dien Street, Bac Tu Liem District, Ha Noi, Viet Nam*

[*]Email: *trantd@haui.edu.vn*

Received: 25 March 2022; Accepted for publication: 28 June 2022

Abstract. The identification of genes causing a genetic disease is still an important issue in the biomedical field because the list of disease genes is still incomplete while it determines the early diagnosis and treatment of fatal genetic diseases such as autism, cancer, drug resistance, and secondary hypertension. Genes associated with a particular disease or similar diseases tend to reside in the same region in a biological network and their location on the network can be predicted. Many network analysis methods have been proposed to solve this problem so far. This review first helps readers access and master the basic concepts of biological networks, disease genes, and their properties. Then, the main content is devoted to the analysis and evaluation of analytical methods recently used to find disease genes on two networks: protein-protein interaction (PPI) and cellular signaling network (CSN). We reported typical problems of identification of primary genes that cause genetic diseases and modern techniques that were widely used for solving those problems. For each technique, we also represented key algorithms so that the audience can exactly implement them for their experiments. In particular, we evaluated the performance of these algorithms in prediction of disease genes and suggested the context for their usage. Finally, the implications of the methods are discussed and some future research directions are proposed. Taken together, disease genes can often be identified from network data by two approaches: network-based methods and machine learning-based methods, and the network-based approach showed better performance in most cases because it works well even if the data size is small.

Keywords: disease gene, network-based approach, biological network, hierarchical closeness.

Classification numbers: 4.8.5

1. INTRODUCTION

Genetic disease is caused by an alteration of one or more genes in combination with environmental factors or by an imbalance of multiple genes [1]. This is the basis of three major

categories of genetic disorders: single gene disorders (mutations in a single gene often cause loss of function), multifactorial conditions (variations of genes that interact with the environment and cause functional alterations), and chromosomal disorders (causing chromosomal imbalances and altered gene dosage). In many cases, genetic diseases are caused by interactions between the environment and genetic factors. Depending on the influence of environmental factors or genetic ones, there will be purely genetic diseases or diseases related to both the environment and heredity. There are some common genetic diseases such as autism, cancer, secondary hypertension, and congenital hemolysis, among which cancer is a very common and painful disease today.

Cancer is the name given to a collection of related genetic diseases [2]. In all types of cancer, some of the body's cells begin to divide and continuously spread to surrounding tissues. Cancer can begin to develop almost anywhere in the human body, where components are made up of trillions of cells. Normally, cells grow and divide to form new cells when the body needs them. When cells become old or damaged, they die and new cells take their place. However, as cancer develops, this process will be disrupted. Cells become more and more abnormal, as old or damaged cells persist even though they must die naturally for new cells when the body doesn't need them, and such damaged cells begin to divide incessantly, which can lead to the formation of tumors. Like other genetic diseases, cancer tends to grow quickly in Viet Nam, which is a main motivation for the identification of disease genes. According to the Global Cancer Organization (GLOBOCAN), in 2012 there were 14.1 million new cancer cases worldwide and 8.2 million deaths [3]. In Viet Nam, on average, there are at least 125,000 new cancer cases each year, and it is forecasted that by 2022 there will be 189,000 cases of this dangerous disease each year [4]. Statistics show that there are 5 common types of cancer: gastrointestinal cancer, breast cancer, gynecological cancer, head and neck cancer, and lung cancer [5]. Cancer is a typical genetic disease to study to identify disease genes, which are genes with disease-causing mutations.

Currently, there are many methods to identify disease genes to support early detection and treatment of cancer genes [6, 7]. Typically, similarity-based methods include POCUS [8], SUSPECTS [9], Endeavor [10], ToppGene [11], and Cardigan [12]. Besides, there have been a number of methods based on machine learning techniques consisting of Decision Tree Learning [13], k-Nearest Neighbor [14], Naive Bayesian [15], Artificial Neural Networks [16], Support Vector Machines [17], Random Forest [18], VarCoPP [19], and Graph Neural Network (GNN) [20]. However, the above methods have some disadvantages, for example, the similarity-based method is limited to cases where known and unknown disease genes are indirectly related to each other or have the same function [21]. Meanwhile, methods based on machine learning techniques will be limited to cases when there are new disease gene samples that are not included in the available training set of known disease genes [21]. In other words, machine learning methods often lack the gold standards of disease genes for the diseases so that they can run well in the learning phase on the training disease gene set. To reduce these limitations, biological network-based methods for disease gene prediction have been proposed and show better efficiency in predicting disease genes [21, 22]. This review paper will present two types of problems in predicting disease genes and introduce network-based methods of disease gene prediction on biological networks such as Hierarchical Closeness [23] and ORIENT [24]. Besides showing a roadmap of network-based methods for disease gene prediction, we also presented an evaluation of the performance of these methods and discussed their application in various contexts.

2. BASIC CONCEPTS AND DISEASE GENE RANKING

2.1. Biological networks

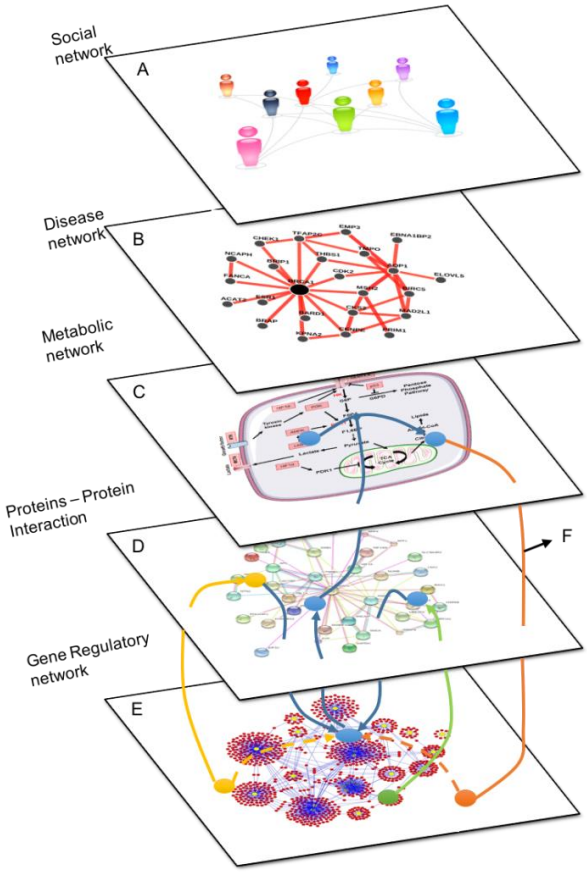

Figure 1. Network layers related to genetic disorder. The biological networks include: Gene Regulatory Network (E), Protein Interaction Network (D), Metabolic Network (C) and Cell Signalling Network (F); Social network (A); Inherited Diseases Network (B).

Biological networks are graphs that are models of biological systems where each node represents a unit such as a gene/protein and each edge indicates an interaction between two units. A network is any system made up of smaller systems that form an integrated whole. Some of the most popular biological networks today are:

- A gene regulatory network (Figure 1E) is a directed network, located in the nucleus of the cell. Gene regulation is a general term for the cellular control of protein synthesis at the transcriptional step. Gene regulation can also be viewed as a cell's response to internal stimuli. Normally, one gene is regulated by another through corresponding proteins (called transcription factors), so that gene regulation is coordinated in a gene regulatory network.
- A protein interaction network (Figure 1D) is a scalar network, each node represents proteins and edges represent interactions, specifically, two proteins are connected if they interact with each other.
- A metabolic network (Figure 1C) is a directed network in which each node represents a

metabolite (molecule) and each edge represents a biochemical reaction. A biochemical reaction is a chemical process that results in the conversion of one biomolecule to another. Metabolism is the biochemical changes of chemical compounds in the cell. The two main processes in metabolism are the synthesis of complex organic molecules and their degradation. Metabolic networks describe cellular metabolic pathways, which include a sequence of reactions that describe how molecules interact with each other and convert to another molecule or compound.

- A cellular signaling network (Figure 1F) is a mixed network. Signaling is a communication process within a cell to coordinate its response to environmental changes. The cell signaling network is a network that directs chemical reactions in a cell from a stimulus to a stimulating effect.
- Disease network (Figure 1B) is a network of human disorders and diseases with reference to their genetic origins or other features. In other words, the network is the map of human disease relationships referring mostly to disease genes. For example, in a human disease network, two diseases are linked if they share at least one associated gene. Note that the disease network represents the disease relationship inside a human body.
- Social network (Figure 1A) is a network that shows the ties between individuals among a community. The married tie between two people may cause new disease such as inbreeding marriage, and therefore social network can impact the disease network.

All these networks are different layers that describe gene interaction from the cell nucleus to human society, in which the evolution of human society reflects that of the gene system.

2.2. Disease gene ranking

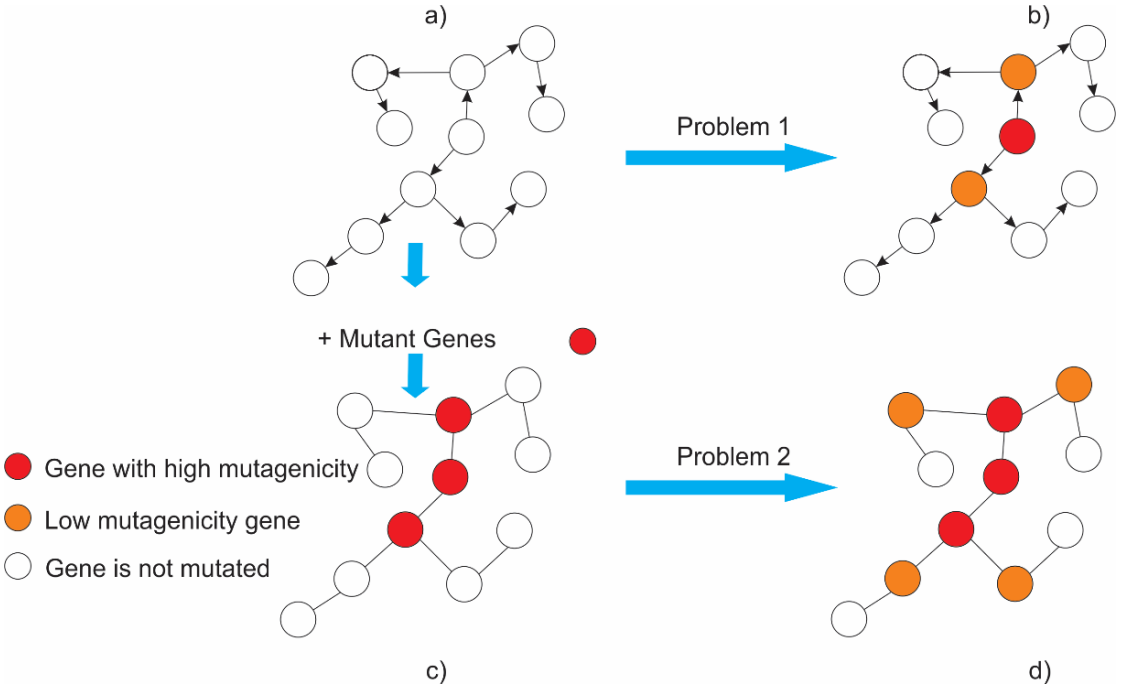

Figure 2. Disease gene ranking problems on biological networks. (a) the original given network; (b) The resulting network of Problem 1; (c) The original network after mapping the mutated genes; d) The network of the results of Problem 2.

Gene ranking is the use of computational methods to sequence human genes so that genes potentially associated with the disease receive a higher rank (Figure 2). After ranking, a small group of genes with high ranking will then be selected for experimental testing. The identification of disease-causing genes is an important problem in biomedicine and molecular biology. Previously, the identification of pathogenic genes was done mainly by biological experiments. This method, performed for hundreds of human genes located on a suspicious chromosome region, requires a lot of time and finance. In fact, there were two types of problems for ranking disease genes, and they will be described in the next sections.

3. PROBLEM 1: GIVING A NETWORK AND PREDICTING DISEASE GENES

3.1. Problem statement

Input: Given a graph $G = (V, E)$ where V is the node set, $V = \{v_1, v_2, \ldots, v_n\}$; E is the edge set, $E = \{(v_i, v_j) | v_i, v_j \in V, i, j = 1, \ldots, n\}$.

Output: a relation $S(V,F)$ where V is the node set; $F \in$ R* indicates the possibility of v being a mutant gene or disease gene.

3.2. Centrality ranking algorithm

The Hierarchical Closeness (HC) algorithm of node v is calculated by the following formula:

$$HC(v) = N_R(v) + CC(v)$$

where, $N_R(v)$ is the number of nodes accessible from the defined node v. $N_R(v) = |\{w \in V | \exists$ a path from v to $w\}|$, is calculated as follows:

Algorithm 1. Computation of Reachability

```
1   function N_R(G, v)
2     // Input: G = (V,E), v∈V
3     // Output: Reachability of v
4   begin
5     count ← 1; Label v as discovered
6     for each vertex w in adjacent List(v)
7       if vertex w is not labeled as discovered then
8           if size of adjacent (w) > 0 then
9             count ← count + N_R(G, w)
10          count ← count + 1
11    return count
12  end
```

$CC(v)$ is the Closeness of node v with its value computed by $CC(v) = \frac{1}{far(v)}$ where the farness of node v denoted by $far(v)$ is calculated by the following formula:

$$far(v) = \sum_{\substack{w \in V \setminus \{v\} \\ d_G(v,w) \neq \infty}} d_G(v,w)$$

where $d_G(v, w)$ is the distance of the shortest path, if any, from v to$w$. If node v is isolated (no edges connected to other nodes), then $CC[v] = 0$. Generally, $CC(v)$ can be computed by the Algorithm 2 as follows:

Algorithm 2. Computation of Closeness

```
1   function CC(v)
2     //Input: G = (V,E), v∈V
3   //Output: CC(v)
4   begin
5     Initialize empty queue Q
6     d[u] ← ∞, ∀u ∈ V\{v}
7     Q.push(v);d[v] ← 0;far[v] ← 0
8     while Q is not empty do
9       u ←Q.pop()
10      for each w in adjacent(u)
11        if d[w] =∞ then
12          Q.push(w)
13        d[w] ←d[u] + 1;far[v] ← far[v] + d[w]
14
15    cc[v] ← 1/far[v]
16    return cc[v]
17  end
```

As reported previously, the Hierarchical Closeness ranking of a graph $G = (V,E)$ represents the ranking of the fragility of nodes [23], and this property of HC was used to rank disease genes for solving this problem where HC is considered attribution F of the output relation $S(V,F)$.

3.3. The performance of the methods solving the first problem

For input data, disease gene data were obtained from the OMIM database (Online Mendelian Inheritance in Man) containing 4350 disease gene data and mapped into a cell signaling network consisting of 1953 nodes and 8579 associations obtained from the KEGG database [25]. After testing on KEGG cellular signaling network, the obtained results showed that Hierarchical Closeness (HC) algorithm had better performance than other ranking algorithms including Degree, Reachability, Closeness, Betweenness, and PageRank (all p-values < 0.05) from 5 % to 25 % on the same cell signaling network [23]. In other words, the benchmark results indicated that HC was the best method to solve problem 1. In addition, the authors also investigated the predictive performance of HC on four specific disease-causing gene groups such as cancer, genetics, immune, and neurodegenerative subtypes. The results also revealed that genes with high HC (K ≤ 60 %) tended to be more pathogenic genes than the rest.

4. PROBLEM 2: GIVING DISEASE GENES AND FINDING THE NEXT DISEASE GENES

4.1. Problem statement

Input: Given a connected and weighted graph $G = (V, E)$ where V is the set of nodes, $V = \{v_1, v_2, ..., v_n\}$; E is the edge set, $E = \{(v_i, v_j)|v_i, v_j \in V, i, j = 1,..., n\}$, a probability of returning c ($0 \leq c \leq 1$), a set of source nodes $S \subseteq V$, p_0 is the initialization vector N x 1 where the value of each element corresponding to a non-source or source node is 0 and 1 /|S| and W is the normal form adjacency matrix of G.

Output: p is a probability vector N x 1 of $|V|$ node.

4.2. Random walk method

The random walk algorithm on a network or graph is defined as a process of moving from a current node to any random neighbour node, starting from a source node [26]. The random walk with restart (RWR) algorithm is a variant of the random walk algorithm and at any time during the migration it allows to return (restart) the source nodes (also called starting nodes) with a probability (c), also known as the back-probability. Given a connected and weighted graph $G(V, E)$ with a set of nodes $V = \{v_1, v_2, ..., v_n\}$ and a set of links $E = \{(v_i, v_j)|v_i, v_j \in V\}$, a set of source nodes $S \subseteq V$ and an adjacency matrix W of size NxN, RWR can be described as follows:

$$P_{t+1} = (1 - c).W.P_t + c.P_0$$

where, P_t is a probability vector N x 1 of $|V|$ node at time step t (the i-th element represents the step at node $v_i \in V$) and P_0 is an initialization vector N x 1 whose value for each element corresponding to a non-source node or a source node is 0 and $1/|S|$. In the case of an unweighted graph, it can be easily converted to a weighted graph by assigning an arbitrary weight to all interactions. The adjacency matrix W is represented by a column matrix of normal form $(W)_{ij}$ where the element (i, j) of W represents the probability that a step at v_i moves to v_j in the interval $V \setminus \{v_i\}$.

The Random Walk with Restart is an extended version of the random walk algorithm. Recently it has been used by many researchers for gene prioritization [27-29]. The algorithm is presented as follows:

Algorithm 3. Random Walk with Restart

```
1   function RwR(G, S, c)
2     //Input:G = (V,E)
3           seeding nodes S
4                 restart probability c ∈ [0, 1]
5   //Output: Stationary vector from Random Walk with Restart at S
6   begin
7     threshold← 1e-10; e← 1; t← 1;
8     W← create adjacent matrix from G(V, E);
9     P_t←W[0];
10    for each i∈S
11      P_t[i]← S[i]
```

12 $P_t \leftarrow P_t / \sum_i P_t[i]; P_0 \leftarrow P_t$
13 **while** $e >= threshold$ **do**
14 $P_{t+1} \leftarrow (1 - c).W.P_t + c.P_0$
15 $e \leftarrow \sqrt{\sum_i (P_{t+1}[\mathrm{i}] - P_t[\mathrm{i}])^2}$
16 $t \leftarrow t+1$
17 **return** P_t
18 **end**

4.3. ORIENT method

The ORIENT method (neighbor-favoring weight reinforcement) is an approach to improve the RWR method performance by enhancing the weights of neighboring interactions with known disease genes [24]. In the ORIENT method, there are two factors that affect the performance of the method, which are the back-probability and the weight-reinforcement rate. The probability of returning to the source node (c) is the probability that a node on the graph will return to the source node where it came from, or when c has a relatively large value, the nodes on the graph tend to frequently return to the source node, and the neighbours around that source node are ranked higher [24]. The ORIENT method has the best performance when the probability of returning to the source node c has the smallest value. In other words, when the value of c is high, it will limit the ability to reach the neighbours far away from the source node, but only concentrate on the nodes located near the source node [24]. Another factor that affects the performance of the method is the weight-reinforcement rate. ORIENT will have the best performance when the weight increment ratio takes on a sufficiently large value.

4.4. HumanNet model

HumanNet is a set of gene networks inferenced for predicting disease genes. From each of these networks, it is possible to predict different disease genes by network-based gene prioritization algorithms from a number of known disease genes [30]. Next, HumanNet v2 as a new feature of the updated HumanNet is a four-level inclusive hierarchy of the human gene networks [31]. Finally, HumanNet v3 has significant improvements such as: the number of networks is reduced to a three-tier model with a larger network size, and it outperforms HumanNet v2 in disease gene prediction [32]. HumanNet provides a practical resource for the study of a wide variety of human diseases.

4.5. The performance of the methods solving the second problem

For input data, two types of gene/protein networks, functional linking networks (FLNs) [33] and protein interaction networks were obtained from the STRING database [34]. The FLN is a weighted graph with 21,657 genes and 22,388,609 links. The experimental results show that the ORIENT method has better performance than the traditional random walk with restart method from 5 % to 15 % and has the best performance when the weights of the interaction edges are related only to adjacent genes closest to the enhanced pathogenic gene. In addition, the performance of the ORIENT method is inversely proportional to the probability of returning to the source node (c), while the performance of other traditional methods without increasing the weights is proportional in the gene/protein networks [24]. However, the performance of the ORIENT method on popular networks is not better than those on HumanNet networks [32].

5. CONCLUSIONS AND DISCUSSIONS

After reviewing and testing disease gene ranking methods, we made some important conclusions that each method has effective application in different usage contexts. To solve the first problem where no known disease genes are given, the Hierarchical closeness method has better performance than other ranking algorithms in predicting disease genes on directed networks, especially for cancer, genetic, immunological, and neurodegenerative disease genes, in other words, the HC algorithm cannot predict all disease genes [23]. Therefore, from Hierarchical closeness, there have been some advanced researches to improve its performance in some specific diseases such as cancer [7]. To solve the second problem where some disease genes are seeded, there are often two primary approaches: machine learning methods and network-based methods. Because machine learning methods have a limitation that requires a lot of known disease genes for the learning phase, while there is no gold standard for disease genes of various diseases, network-based approaches with natural connections (e.g. in protein-interaction networks or gene networks) between known and unknown disease genes in the form of 'disease module' are usually used for disease gene prediction [35, 36]. With network-based approach, the ORIENT method has better performance than the traditional random walk with restart method and has the best performance when the weights of the interactions involving only nearest neighboring disease genes have been enhanced. In other words, the performance of the ORIENT method will depend on the quality of the source gene dataset and is inversely proportional to the probability of returning to the source node (c) [24]. However, the performance of the ORIENT method on popular networks is not better than those on HumanNet networks.

The identification of disease genes is still a very attractive issue in the field of biomedicine. In this day and age, the development of science and technology has given birth to many methods of finding disease genes. The paper has introduced the basic concepts of biological networks, disease genes and their characteristics. After that, we analyzed and evaluated two methods of detecting disease genes: Hierarchical closeness and ORIENT on cell signaling networks and protein interaction networks. For further studies, it may continue to analyze and evaluate more deeply the methods of ranking pathogenic genes as well as propose to test the two methods Hierarchical closeness and ORIENT on other networks such as metabolism networks, gene regulatory networks, accordingly investigate how to overcome the limitations of existing disease gene ranking methods. Besides, we can improve the network quality for disease gene prediction by network inference methods such as Graph neural networks [37].

Acknowledgments. This research was funded by Hanoi University of Industry under a funding grant in 2022 for the research team led by the corresponding author.

CRediT authorship contribution statement. Minh-Tan Nguyen: made the survey, drafted the manuscript. Tien-Dzung Tran: conceived the study, edited, reviewed, and gave guidance on the theoretical and mathematical issues and contexts. Both authors provided critical feedback and corrections.

Declaration of competing interest. We declare that we have no competing interests.

REFERENCES

1. Simon C. and Farndon P. - What Causes Genetic Disorders? InnovAiT **1** (8) (2008) 544-553.
2. Schram F. R. and P. K. L. Ng - What is Cancer?, Journal of Crustacean Biology **32** (4) (2012) 665-672.

3. Globocan W. - Estimated cancer incidence, mortality and prevalence worldwide in 2012. Int Agency Res. Cancer (2012) 43-50.
4. Duc-Tinh Pham, M. T. N., Ha-Nam Nguyen, Tien-Dzung Tran - Analyzing cancer data in North Vietnam by complex network technique, Journal of Science and Technology: Issue on Information and Communications Technology **19** (12.2) (2021).
5. Braithwaite D., Demb J., and Henderson L. - American Cancer Society: cancer facts and figures 2016, Atlanta, GA: American Cancer Society, 2016, p. 53
6. Tran T. D. and Pham D. T. - Identification of anticancer drug target genes using an outside competitive dynamics model on cancer signaling networks, Scientific Reports **11** (1) (2021) 14-21
7. Tran T. D. and Kwon Y. K. - Hierarchical closeness-based properties reveal cancer survivability and biomarker genes in molecular signaling networks, PLOS ONE **13** (6) (2018) e0199109.
8. Turner F. S., Clutterbuck D. R., and Semple C. A. M. - POCUS: mining genomic sequence annotation to predict disease genes, Genome Biology **4** (11) (2003) R75.
9. Adie E. A., et al. - SUSPECTS: enabling fast and effective prioritization of positional candidates, Bioinformatics **22** (6) (2006) 773-774.
10. Aerts S., et al. - Gene prioritization through genomic data fusion, Nature Biotechnology **24** (5) (2006) 537-544.
11. Chen J., et al. - Improved human disease candidate gene prioritization using mouse phenotype, BMC Bioinformatics **8** (1) (2007) 392.
12. Cáceres J. J. and Paccanaro A. - Disease gene prediction for molecularly uncharacterized diseases, PLOS Computational Biology **15** (7) (2019) e1007078.
13. Adie E. A., et al. - Speeding disease gene discovery by sequence based candidate prioritization, BMC Bioinformatics **6** (1) (2005) 55.
14. Kuncheva L. I. - Editing for the k-nearest neighbors rule by a genetic algorithm, Pattern Recognition Letters **16** (8) (1995) 809-814.
15. Moore J. H., et al. - A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, Journal of Theoretical Biology **241** (2) (2006) 252-261.
16. Khan J., et al. - Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine **7** (6) (2001) 673-679.
17. Guyon I., et al. - Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning **46** (1) (2002) 389-422.
18. Jiang R., et al. - A random forest approach to the detection of epistatic interactions in case-control studies, BMC Bioinformatics **10** (1) (2009) S65.
19. Papadimitriou S., et al. - Predicting disease-causing variant combinations, Proceedings of the National Academy of Sciences **116** (24) (2019) 11878-11887.
20. Shu J., et al. - Disease gene prediction with privileged information and heteroscedastic dropout, Bioinformatics **37** (Supplement_1) (2021) i410-i417.
21. Le D. H., Xuan Hoai N., and Kwon Y. K. - A Comparative Study of Classification-Based Machine Learning Methods for Novel Disease Gene Prediction, In: Knowledge and Systems Engineering, Cham: Springer International Publishing, 2015.
22. Tran T. D. and Kwon Y. K. - The relationship between modularity and robustness in signalling networks, Journal of The Royal Society Interface **10** (88) (2013) 20130771.

23. Tran T. D. and Kwon Y. K. - Hierarchical closeness efficiently predicts disease genes in a directed signaling network, Computational Biology and Chemistry **53** (2014) 191-197.
24. Le D. H. and Kwon Y. K. - Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization, Computational Biology and Chemistry **44** (2013) 1-8.
25. Kim J. R., et al. - Reduction of Complex Signaling Networks to a Representative Kernel, Science Signaling **4** (175) (2011) ra35-ra35.
26. Köhler S., *et al.* - Walking the Interactome for Prioritization of Candidate Disease Genes, The American Journal of Human Genetics **82** (4) (2008) 949-958.
27. Lei X. and Bian C. - Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association, Scientific Reports **10** (1) (2020) 1943.
28. Li A., *et al*. - A novel miRNA-disease association prediction model using dual random walk with restart and space projection federated method, PLOS ONE **16** (6) (2021) e0252971.
29. Joodaki M., *et al.* - A scalable random walk with restart on heterogeneous networks with Apache Spark for ranking disease-related genes through type-II fuzzy data fusion, Journal of Biomedical Informatics **115** (2021) 103688.
30. Lee I., *et al*. - Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome research **21** (7) (2011) 1109-1121.
31. Hwang S., *et al*. - HumanNet v2: human gene networks for disease research, Nucleic Acids Research **47** (D1) (2018) D573-D580.
32. Kim C. Y., *et al*. - HumanNet v3: an improved database of human gene networks for disease research, Nucleic Acids Research **50** (D1) (2021) D632-D639.
33. Linghu B., *et al.* - Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biology **10** (9) (2009) R91.
34. Szklarczyk D., *et al*. - The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, Nucleic Acids Research **39** (suppl_1) (2010) D561-D568.
35. Le D. H. - Machine learning-based approaches for disease gene prediction, Briefings in Functional Genomics **19** (5-6) (2020) 350-363.
36. Ata S. K., *et al*. - Recent advances in network-based methods for disease gene prediction, Briefings in Bioinformatics **22** (4) (2020).
37. Zhang X. M., *et al.* - Graph Neural Networks and Their Current Applications in Bioinformatics, Frontiers in Genetics **12** (2021).