frontiers | Frontiers in Aging Neuroscience

TYPE Methods
PUBLISHED 21 February 2023
DOI 10.3389/fnagi.2023.1061892

OPEN ACCESS

EDITED BY
Ailan Wang,
Geneis (Beijing) Co., Ltd., China

REVIEWED BY
Lihong Peng,
Hunan University of Technology, China
Bolin Chen,
Northwestern Polytechnical University, China
Wei Peng,
Kunming University of Science and Technology, China

*CORRESPONDENCE
Ju Xiang
✉ xiang.ju@foxmail.com

SPECIALTY SECTION
This article was submitted to Cellular and Molecular Mechanisms of Brain-aging, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 05 October 2022
ACCEPTED 30 January 2023
PUBLISHED 21 February 2023

CITATION
Ma J, Qin T and Xiang J (2023) Disease-gene prediction based on preserving structure network embedding.
Front. Aging Neurosci. 15:1061892.
doi: 10.3389/fnagi.2023.1061892

Disease-gene prediction based on preserving structure network embedding

Jinlong Ma[1], Tian Qin[1] and Ju Xiang[2,3]*

[1]School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China, [2]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, China, [3]Department of Basic Medical Sciences, Changsha Medical University, Changsha, China

Many diseases, such as Alzheimer's disease (AD) and Parkinson's disease (PD), are caused by abnormalities or mutations of related genes. Many computational methods based on the network relationship between diseases and genes have been proposed to predict potential pathogenic genes. However, how to effectively mine the disease-gene relationship network to predict disease genes better is still an open problem. In this paper, a disease-gene-prediction method based on preserving structure network embedding (PSNE) is introduced. In order to predict pathogenic genes more effectively, a heterogeneous network with multiple types of bio-entities was constructed by integrating disease-gene associations, human protein network, and disease-disease associations. Furthermore, the low-dimension features of nodes extracted from the network were used to reconstruct a new disease-gene heterogeneous network. Compared with other advanced methods, the performance of PSNE has been confirmed more effective in disease-gene prediction. Finally, we applied the PSNE method to predict potential pathogenic genes for age-associated diseases such as AD and PD. We verified the effectiveness of these predicted potential genes by literature verification. Overall, this work provides an effective method for disease-gene prediction, and a series of high-confidence potential pathogenic genes of AD and PD which may be helpful for the experimental discovery of disease genes.

KEYWORDS

disease-gene prediction, heterogeneous network, network embedding, network propagation, human essential genes

1. Introduction

Alzheimer's disease (AD) and Parkinson's disease (PD) are two of the most common forms of neurodegenerative illness that affect people all over the globe. The two diseases are both complicated problems that are associated with aging. AD is the most common cause of dementia as well as other neurodegenerative illnesses, and the symptoms of the condition often include behavioral abnormalities, memory loss, and cognitive impairment (Joe and Ringman, 2019; Li et al., 2021). About 1–2 percent of adults over the age of 60 are diagnosed with PD, making PD the most prevalent motor neurodegenerative illness (Wirdefeldt et al., 2011; Ascherio and Schwarzschild, 2016). Diagnosis of PD can be established when typical features of dyskinesia are combined with other features such as rigidity, tremor, and postural instability (Postuma et al., 2015). For patients, these disease may bring them tremendous emotional pressure and financial pressure. At the moment, patients are mostly treated with pharmaceuticals such as N-methyl-D-aspartic acid receptor antagonists, memantine, and cholinesterase inhibitors. For the time being, there is no all-encompassing therapeutic

solution accessible for this condition (Jevtic et al., 2017). Therefore, it is an urgent need to explore and evaluate novel cell-related biomarkers and therapeutic targets to increase the level of diagnosis and therapy offered to patients with AD or PD. The creation of gene networks may bring about the alterations associated with the etiology and development of AD and PD.

The screening and identification of pathogenic genes is one of the research hot spots in the field of modern bio-medicine, which is important to explore the pathogenic mechanism of the disease. The rapid advancement of contemporary biomedical technology has led to the production of a significant amount of data about biological networks, which in turn has facilitated the development of network bio-medicine (Ata et al., 2021; Xiang et al., 2022b). This change has resulted in the development of a novel concept and path for the screening of pathogenic genes. Many network-based algorithms have been proposed to prioritize screening disease-related candidate genes. Hu et al. (2018) proposed a novel disease-gene-prediction method by combining path-based structure with community structure characteristics in human protein–protein networks. Liu et al. (2021) built a biological heterogeneous network on known correlations between the entities from different sets, and proposed a new network embedded representation algorithm to calculate the correlation between disease and genes, using the correlation score to predict pathogenic genes. Xiang et al. (2021a) proposed a framework of network impulsive dynamics on multiplex biological network to predict disease-related genes which could identify disease-related genes by mining the dynamical responses of nodes to impulsive signals being exerted at specific nodes. Xiang et al. (2022a) proposed a hybrid disease-gene prediction method integrating multiscale module structure (HyMM), which could utilize multiscale information from local to global structure to more effectively predict disease-related genes. Ruan and Wang (2021) proposed a Disease-Specific Network Enhancement Prioritization (DiSNEP) framework to improve disease gene prioritization using networks. However, how to effectively mine the disease-gene relationship network to predict disease genes better is still an open problem.

Network embedding, which is an effective way to extract useful information from networks, transforms the nodes of network into low-dimensional spatial vectors while maximizes the information about the network structure and attributes (Mikolov et al., 2013; Perozzi et al., 2014; Tang et al., 2015; Wang et al., 2019). For instance, Li et al. suggested a representation learning method which used joint binary network embedding to conduct an analysis of single-cell RNA-seq data. The proposed heterogeneous network was able to incorporate numerous binary networks, allowing for a low-dimensional representation of a variety of node types to obtain (Li and Patra, 2010). DeepWalk (Perozzi et al., 2014) implements a depth-first search over the network, whereas LINE (Tang et al., 2015) implements a breadth-first searching strategy to generate a context for nodes. Zeng et al. presented the idea of embedding multiview knowledge in order to get an understanding of entity embedding. This was due to the fact that multiview learning might lead to improved generalization performance in order to learn exhaustive entity embedding from various views (Zeng et al., 2016). Xiang et al. (2021b) proposed a method for predicting disease-related genes by using fast network embedding (PrGeFNE), which could integrate multiple types of associations related to diseases and genes.

In this paper, a method called preserved structure network embedding (PSNE) is offered for the prediction of disease genes. Firstly, we collect disease-gene associations, human protein network, and disease-disease associations to construct a heterogeneous network for integrating information. Each node in the heterogeneous network represents a vector that can retain the input network structure. Secondly, the network embedding algorithm is introduced to obtain low-dimensional vector representations of the nodes that make up the network. Thirdly, in order to remove unnecessary information, the low-dimensional vector representations that are retrieved from the nodes are made to be sparse. In order to create the disease-gene network, as well as to rebuild the heterogeneous network made up of diseases and genes, low-dimensional vector representations are used. In the last step, network propagation algorithm is used in order to forecast disease genes utilizing the newly developed two-layer heterogeneous network.

The remaining parts of the article are detailed down below. In Section 2, the disease-related gene data sources used in this investigation are outlined. Then, a method called PSNE is proposed for predicting disease genes. In Section 3, we compare other methods to confirm the excellent performance of PSNE through experiments and analyze the gene prediction ability of the PSNE method in AD and PD. In Section 4, we summarize the work and discussion.

2. Materials and methods

2.1. DATASET

2.1.1. Disease-gene association

Data availability, fragmentation, heterogeneity and inconsistency of concept description are problems that must be overcome in disease mechanism research. DisGeNet is a database which collects a large number of mutations and genes related to human diseases (Mendelian genetic diseases, complex diseases, and environmental diseases). DisGeNet is a collection of data obtained by collating and combining the data from public databases, scientific literature, Genome-wide association study catalogs, and animal models. The data collected by the database are annotated by a unified standard. In addition, there are more perfect basic criteria to determine the order of the relationship between genotypes and phenotypes. This information can be accessed through web interfaces, cytoscape applications, R packages, and scripting in several programming languages. DisGeNet is not only a multi-functional platform but also can be used for different research purposes, comprising the molecular essential of specific human diseases and their complications, analyzing pathogenic gene characteristics, constructing drug therapeutic effects and hypotheses of adverse drug reactions, testing candidate disease genes, and evaluating text mining approaches. The latest version of DisGeNet is v7.0, containing 1,134,942 genetic disease associations, between 21,671 genes and 30,170 diseases, symptoms, characteristics, and clinical or abnormal human phenotypes, as

well as 369,554 variant disease associations, between 194,515 variants and 14,155 diseases, characteristics, and phenotypes.

In this paper, the DisGeNet database is combed in search of disease-gene connections and filtered the primary dataset. Firstly, we choose "disease" as "diseaseType" and "Disease or Syndrome" as "diseaseSemanticType." Then, the genes that do not exist in the human protein network are filtered out. If a disease has the same set of disease genes, only one of the diseases is randomly retained. Finally, the number of disease-gene associations is 20,274.

2.1.2. Human protein network

To avoid the incompleteness of human protein networks caused by a single data source, we employ extensive interactions that are generated from the sources listed below (Menche et al., 2015). (1) Regulatory interactions; (2) literature-compiled interactions, which are mostly derived from low-throughput trials; (3) binary interactions derived from numerous yeast two-hybrid high-throughput datasets as well as data gathered from the published literature; (4, 5) interactions of the signaling pathways; (6) pair of kinase and substrate molecules; and (7) protein complexes.

2.1.3. Disease-disease association

By using the same method in MimMiner recently, the disease-disease similarity scores are calculated to construct the disease-disease network. The OMIM IDs are mapped one by one to UMLS IDs in DisGeNet, and then k-nearest neighbor method is used to obtain sparse disease-disease network.

2.2. Methods

Here, the disease-gene prediction method is presented by using preserving structure network embedding (PSNE), which can use multi-source biological information to predict disease-related genes more effectively. The PSNE method consists of four parts: heterogeneous network construction, network embedding algorithm, heterogeneous network reconfiguration and heterogeneous network propagation (see Figure 1). Next, we will describe the details of the four parts.

2.2.1. Heterogeneous network construction

The disease gene prediction method is get start by constructing heterogeneous networks using disease and gene association data from multiple sources. In order to solve the network sparsity problem, the disease-gene network is needed to enrich by using other known human protein relationship networks and disease-disease relationships. Heterogeneous network is constructed by integrating three different types of connected data: disease-gene associations, human protein network, and disease-disease associations.

2.2.2. Network embedding algorithm

Network embedding is an algorithm to get the information from the network, which converts the nodes into a vector of low dimensional space while maximally preserves the network structural information (Dai et al., 2019). There are many network embedding algorithms. The random walk algorithm is used in the PSNE method. The purpose of network embedding is to determine the interconnections that exist between each node and the links that are immediately around the node by using a series of vectors that is created by random walk (Grover and Leskovec, 2016). The functional similarity of two nodes in a network is correlated with the distance between those nodes. Network analysis is utilized to provide a quantitative assessment of the links between genes and diseases. More researches can be done in the disease-gene associations by applying the network embedding algorithm to the structure.

Let $G = (V, E, W)$ denote a heterogeneous network, where vertex $v \in V$ indicates a gene, $e(u, v) \in E$ is the edge of connecting genes v and u, $w(u, v) \in W$ is the weight of edge $e(u, v)$, which is used to characterize the probability of a relationship between u and v. Here, the weight of all edges in the heterogeneous network is set to 1, which means that they are equal to each other. Considering the nature of grouping between heterogeneous networks, the vertex may move toward its adjacent position with different probabilities. In the process of random walk, nodes tend to travel along the edge and have the highest probability of transitioning to their proximity. Given a vertex v_i and an edge $e(v_{i-1}, v_i)$, where vertex v_i visited vertex v_{i-1} in the previous step. By calculating the transition probability $T(v_i, v_{i+1})$ on edge $e(v_i, v_{i+1})$, vertex v is transferred to one of its neighbors v_{i+1}. The transition probability $T(v_i, v_{i+1})$ is defined as follows:

$$T(v_i, v_{i+1}) = \pi(v_{i-1}, v_{i+1}) * W, \tag{1}$$

$$\pi(v_{i-1}, v_{i+1}) \begin{cases} \frac{1}{p}, d_{v_{i-1}v_{i+1}} = 0 \\ 1, \ d_{v_{i-1}v_{i+1}} = 1 \\ \frac{1}{q}, d_{v_{i-1}v_{i+1}} = 2 \end{cases}, \tag{2}$$

where $d_{v_{i-1}v_{i+1}}$ represents the shortest path distance from the previous vertex v_{i-1} to the next vertex v_{i+1}. $d_{v_{i-1}v_{i+1}} = 0$ means that vertex v_{i-1} and v_{i+1} are the same vertex, and vertex v jumps back to its previous vertex v_{i-1}, $d_{v_{i-1}v_{i+1}} = 1$ means that vertex v_{i+1} is the common neighbor of vertex v_{i-1} and v, $d_{v_{i-1}v_{i+1}} = 2$ means that v_{i-1} and v_{i+1} are indirectly connected, and vertex v_{i+1} is not their common neighbor. The parameter p controls the possibility of revisiting the node during random walk. Setting the parameter p to a higher value can ensure that we avoid sampling the nodes that have been visited in the next step. This setting encourages moderate exploration and avoids the redundancy of sampling. If the value of parameter p is very low, it will cause the walk to backtrack one step, which will make the walk close to the starting node. If the parameter q is greater than 1, random walk will tend to be between nodes around the starting point. Such random walk behavior reflects breadth first search. On the contrary, if the parameter q is less than 1, random walk is more likely to visit nodes far away from the node, which reflects depth first search. After calculating the transition probability of each edge in the heterogeneous network, the normalized transition probability matrix T_{norm} is defined to ensure that the sum of the exit probability of each node is 1,

$$T_{norm}(v_i, v_{i+1}) = \frac{T(v_i, v_{i+1})}{\sum_{j \in i} T(v_i, v_j)}. \tag{3}$$

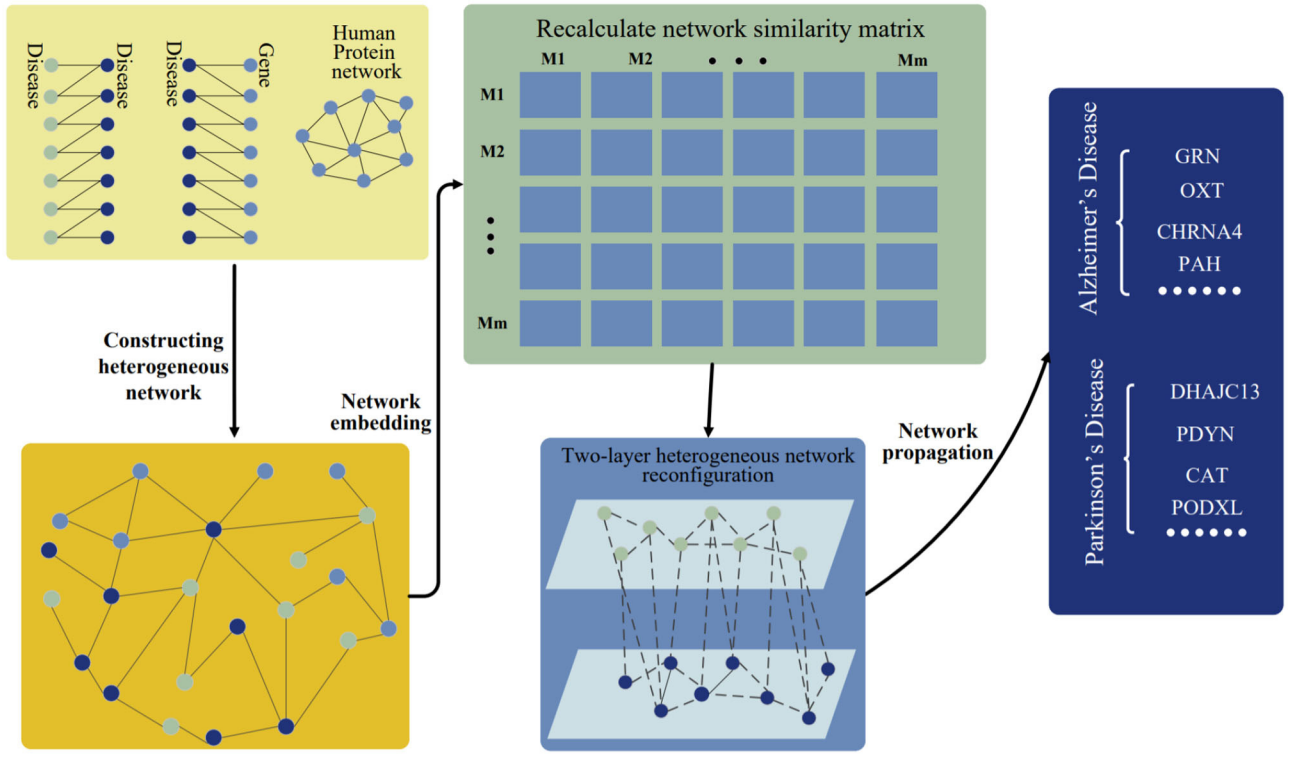

FIGURE 1
Workflow of PSNE. The disease-disease associations, disease-gene associations, and human protein network are connected together and constructed as a heterogeneous network. The network embedding algorithm is used to realize the low-dimensional vector representation of the network and then a new two-layer heterogeneous network of diseases and genes is constructed. Finally, the network propagation algorithm is used to predict the pathogenic genes.

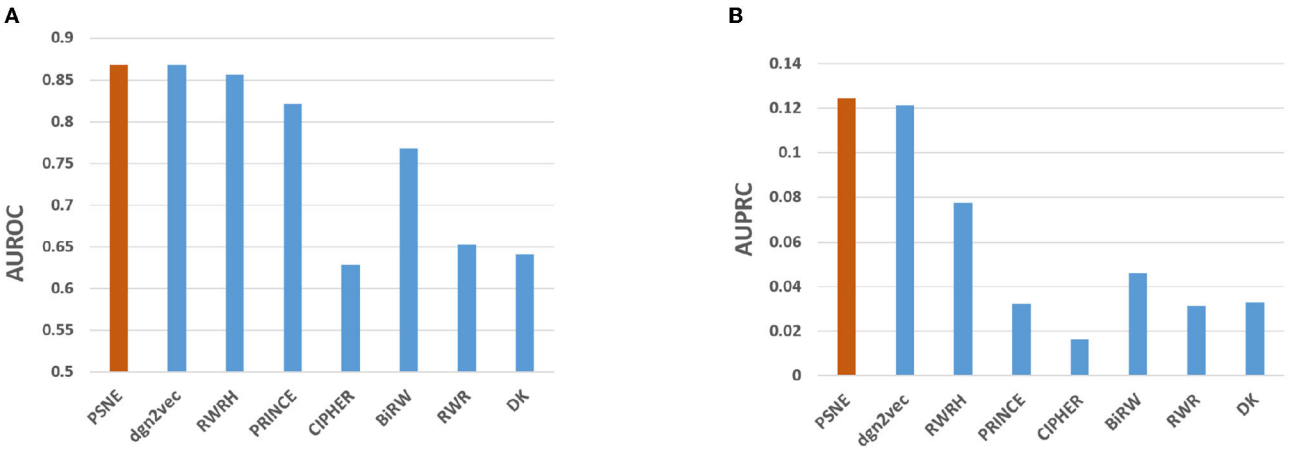

FIGURE 2
Performance evaluation of PSNE in terms of **(A)** AUROC and **(B)** AUPRC, with comparison to the state-of-the-art methods.

Let $G = (V, E, W)$ denote a heterogeneous network. Let f denote the mapping function from node to feature representation to learn the downstream prediction task. Here, d is a parameter that specifies the dimension of feature representation, which is set to 128. Equivalently, f is a parameter matrix of size $|V| \times d$. For each source node $u \in V$, $N_S(u) \subset V$ is defined as the network neighborhood generated by node u through domain sampling strategy S. Feature learning in networks is described as a model that maximizes the log-probability($\log Pr$) of neighbors $N_S(u)$,

$$\max_f \sum_{u \in V} \log Pr\left(N_S(u) \mid f(u)\right). \tag{4}$$

TABLE 1 Performance evaluation of PSNE in terms of top-k Recall (k = 1, 5, 10, 50, 100, and 200), with comparison to the state-of-the-art methods.

Methods	Recall_1	Recall_5	Recall_10	Recall_20	Recall_50	Recall_100	Recall_200
PSNE	**0.078**	**0.176**	**0.227**	**0.287**	**0.359**	**0.418**	**0.499**
dgn2vec	0.075	0.174	0.223	0.279	0.349	0.411	0.497
RWRH	0.046	0.121	0.172	0.237	0.322	0.393	0.475
PRINCE	0.015	0.057	0.093	0.143	0.211	0.257	0.307
CIPHER	0.007	0.029	0.049	0.081	0.137	0.184	0.232
BiRW	0.027	0.080	0.117	0.164	0.234	0.286	0.432
RWR	0.019	0.056	0.077	0.102	0.137	0.166	0.195
DK	0.021	0.058	0.077	0.101	0.131	0.155	0.182

Bold values represent the maximum values in each column of data.

TABLE 2 For Alzheimer's disease, the performance evaluation of PSNE in terms of top-k Recall (k = 1, 5, 10, 50, 100, and 200) with comparison to the state-of-the-art methods.

Methods	Recall_1	Recall_5	Recall_10	Recall_20	Recall_50	Recall_100	Recall_200
PSNE	**0.022**	**0.042**	**0.062**	**0.132**	**0.222**	**0.325**	**0.467**
dgn2vec	0.008	0.037	0.053	0.073	0.160	0.245	0.352
RWRH	0.017	0.026	0.042	0.087	0.149	0.269	0.412
PRINCE	0	0.004	0.009	0.009	0.016	0.085	0.164
CIPHER	0	0	0	0	0.006	0.018	0.052
BiRW	0	0	0.007	0.035	0.110	0.197	0.294
RWR	0.008	0.013	0.013	0.029	0.072	0.146	0.265
DK	0.015	0.031	0.057	0.077	0.146	0.215	0.290

Bold values represent the maximum values in each column of data.

TABLE 3 For Alzheimer's disease, the performance evaluation of PSNE in terms of top-k Precision (k = 1, 5, 10, 50, 100, and 200), with comparison to the state-of-the-art methods.

Methods	Prec_1	Prec_5	Prec_10	Prec_20	Prec_50	Prec_100	Prec_200
PSNE	**0.320**	**0.136**	**0.100**	**0.106**	**0.069**	**0.050**	**0.036**
dgn2vec	0.120	0.104	0.080	0.058	0.050	0.038	0.027
RWRH	0.240	0.080	0.064	0.070	0.047	0.041	0.032
PRINCE	0	0.016	0.016	0.008	0.006	0.013	0.013
CIPHER	0	0	0	0	0.002	0.003	0.004
BiRW	0	0	0.012	0.028	0.035	0.031	0.022
RWR	0.120	0.040	0.020	0.024	0.023	0.023	0.020
DK	0.240	0.104	0.092	0.060	0.046	0.034	0.022

Bold values represent the maximum values in each column of data.

Two standard assumptions are made in order to help the optimization problem easy to handle. One is conditional independence. Given the characteristic representation of nodes, the possibility is decomposed by assuming the possibility of observing neighborhood nodes is independent of observing any other neighborhood nodes:

$$Pr\left(N_S(u) \mid f(u)\right) = \prod_{n_i \in N_S(u)} Pr\left(n_i \mid f(u)\right), \tag{5}$$

And the other is symmetry of feature space. Source nodes and neighborhood nodes have symmetrical influence on each other in the feature space. Therefore, the condition of each node pair is modeled as a unit, and parameterize the node pair through the point product of their characteristics:

$$Pr\left(n_i \mid f(u)\right) = \frac{\exp\left(f\left(n_i\right) \cdot f(u)\right)}{\sum_{v \in V} \exp(f(v) \cdot f(u))}. \tag{6}$$

With the above assumptions, the objective in Equation 4 is simplified as:

TABLE 4 Top 20 related genes of AD predicted by PSNE.

Top 1–10 gene	Evidence	Top 11–20 gene	Evidence
GRN	PMID: 35039149	POMC	PMID: 32982666
OXT	PMID: 31775875	TOMM40	PMID: 29777097
CHRNA4	PMID: 23701948	PON1	PMID: 30714958
PAH	PMID: 27294413	ATP13A2	PMID: 25056458
TYROBP	PMID: 25052481	SOD1	PMID: 33402227
IL6	PMID: 30104698	END1	PMID: 33723589
TFB1M	PMID: 32497722	CAT	PMID: 27809706
NOS2	PMID: 32087283	NGF	PMID: 30804738
IFNG	PMID: 20213229	IL10	PMID: 31879236
CBS	PMID: 32754109	PTGS2	PMID: 29784049

TABLE 5 For Parkinson's disease, the performance evaluation of PSNE in terms of top-k Recall (k = 1, 5, 10, 50, 100, and 200), with comparison to the state-of-the-art methods.

Methods	Recall_1	Recall_5	Recall_10	Recall_20	Recall_50	Recall_100	Recall_200
PSNE	0.008	**0.067**	**0.119**	**0.188**	**0.324**	**0.430**	**0.510**
dgn2vec	**0.014**	0.064	0.089	0.148	0.298	0.403	0.489
RWRH	0.012	0.020	0.043	0.069	0.140	0.235	0.381
PRINCE	0	0	0	0.027	0.038	0.056	0.078
CIPHER	0	0	0	0	0	0.006	0.043
BiRW	0	0	0	0.027	0.038	0.056	0.079
RWR	0	0.002	0.010	0.020	0.041	0.051	0.093
DK	0	0.002	0.006	0.013	0.028	0.044	0.056

Bold values represent the maximum values in each column of data.

TABLE 6 For Parkinson's disease, the performance evaluation of PSNE in terms of top-k Precision (k = 1, 5, 10, 50, 100, and 200), with comparison to the state-of-the-art methods.

Methods	Prec_1	Prec_5	Prec_10	Prec_20	Prec_50	Prec_100	Prec_200
PSNE	0.160	**0.264**	**0.240**	**0.192**	**0.134**	**0.089**	**0.052**
dgn2vec	**0.280**	0.264	0.180	0.150	0.123	0.083	0.050
RWRH	0.240	0.088	0.088	0.069	0.058	0.049	0.039
PRINCE	0	0	0	0.028	0.016	0.012	0.008
CIPHER	0	0	0	0	0	0.001	0.005
BiRW	0	0	0	0.028	0.016	0.012	0.008
RWR	0	0.008	0.020	0.020	0.018	0.011	0.010
DK	0	0.008	0.012	0.014	0.012	0.010	0.006

Bold values represent the maximum values in each column of data.

$$\max_{f} \sum_{u \in V} \left[-\log Z_u + \sum_{n_i \in N_S(u)} f(n_i) \cdot f(u) \right]. \quad (7)$$

For large scale networks, the calculation cost of partition function $Z_u = \sum_{v \in V} \exp(f(u) \cdot f(v))$ of each node is very high, the negative sampling is used to approximate the partition function Z_u. The stochastic gradient ascent algorithm is used to optimize the Equation 7 on the model parameters that define the feature f.

2.2.3. Heterogeneous network reconfiguration

Network structures are concise and efficient data structures, which are used to describe related problems in bio-informatics (Hohmann, 2010). The connections between nodes represent their interactions, such as diseases caused by gene expression, reactions between proteins and other interactions. If two nodes in the network are connected, the information shown by the two nodes is considered related. The heterogeneous network representation learning algorithm embeds the rich

TABLE 7 Top 20 related genes of PD predicted by PSNE.

Top 1–10 gene	Evidence	Top 11–20 gene	Evidence
DNAJC13	PMID: 24218364	NOS3	PMID: 18663495
PDYN	PMID: 17203488	NOS2	PMID: 32003282
CAT	–	KMT2B	–
PODXL	PMID: 26864383	SERPINA1	PMID: 27184740
GSR	PMID: 30156440	C2orf50	–
DRD4	PMID: 23232665	ADH1C	PMID: 15642852
IL1B	PMID: 32003282	GLUD2	PMID: 33093440
NTRK2	PMID: 31991178	XK	–
ALDH5A1	–	POLG	PMID: 32364361
TH	–	PON1	PMID: 32002976

structural and semantic information in the heterogeneous network into the low-dimensional node representation, which is convenient for downstream applications. Low-dimensional node vectors represent useful high-level correlation information in heterogeneous networks. The data are used to reconstruct a two-layer heterogeneous network. In specifically, $C_{i,j} = \frac{\vec{s}_i^T \cdot \vec{s}_j}{|\vec{s}_i| \cdot |\vec{s}_j|}$ is used to calculate the cosine similarity between different diseases, where $\vec{s}$ represents the embedding vector of node and $\vec{s}^T$ is the transposition of $\vec{s}$, and the network embedding algorithm is used to the enhanced disease network. Similarly, an enhanced gene network is made. After that, there is the revised gene association network as well as the disease association network, and disease-gene network is integrated to form a new disease-gene two-layer heterogeneous network D, which can be represented by,

$$D = \begin{pmatrix} H_I & H_G \\ H_G^T & H_N \end{pmatrix}, \tag{8}$$

where H_I represents the improved disease network, H_N represents an improved gene network, H_G indicates a disease-gene association and H_G^T is the transposition of H_G. In contrast to the heterogeneous networks described earlier, the two-layer heterogeneous networks include more concentrated and more pertinent information.

2.2.4. Heterogeneous network propagation

The network propagation is simulated in the new two-layer heterogeneous network of diseases and genes to determine the likelihood of disease-associated genes. A disease network H_I, a gene network H_N, a disease-gene network H_G and H_E which is the transposition of H_G are given, each of them is denoted by a symbol. Then, we come up with the following diagonal matrix D_I, D_N, D_G and D_E, where the elements of the diagonal are specified by $(D_I)_{i,j} = \sum_j (H_I)_{j,i}$, $(D_N)_{i,j} = \sum_j (H_N)_{j,i}$, $(D_G)_{i,j} = \sum_j (H_G)_{j,i}$ and $(D_E)_{i,j} = \sum_j (H_E)_{j,i}$. Then, the normalized matrices of H_I, H_N, H_G and H_E can be written as,

$$\left.\begin{aligned} \hat{H}_I &= H_I D_I^{-1} \\ \hat{H}_N &= H_N D_N^{-1} \\ \hat{H}_G &= H_G D_G^{-1} \\ \hat{H}_E &= H_E D_E^{-1} \end{aligned}\right\} \tag{9}$$

These normalized matrices allow us to construct a new matrix.

$$\hat{D} = \begin{pmatrix} (1-\beta)\hat{H}_I & \beta\hat{H}_G \\ \beta\hat{H}_E & (1-\beta)\hat{H}_N \end{pmatrix}, \tag{10}$$

where β is the jump probability between layers. The random walk can jump to the gene network with probability β when applying to the disease network, or remain in the original network with the probability $1-\beta$.

The inter-layer jump can take place only when the node reached by the random walk algorithm is linked to the node of another layer's. In such case, node can only be moved to that is close to the node in the layer, or may be taken back to the node where the node began. In light of this, we refer to a diagonal matrix as $H_{\hat{D}}$ and $\left(H_{\hat{D}}\right)_{i,i} = \sum_j (\hat{D})_{j,j}$. In a two-layer heterogeneous network, the following approaches is used to get the final transfer matrix T of the network propagation process,

$$T = \hat{D} \cdot H_{\hat{D}}^{-1}. \tag{11}$$

The model of a random walk with a reset can be described by,

$$q_{t+1} = (1-\alpha)T \cdot q_t + \alpha q_0. \tag{12}$$

The initial probability vector of the random walk is characterized by $q_0 = \left(q_{0D}^T, q_{0G}^T\right)^T$, where q_{0D}^T is the initial probability vector of the disease network. The initial probability vector of the gene network is denoted by q_{0G}^T, where parameter $\alpha \in (0,1)$ is the restart probability, which means that the random walk algorithm will have probability α to return to the initial position. After a certain number of steps, the acquired probability will eventually arrive at a state of stability. The genes can be sequenced in accordance with the probability which will stabilize, and then the genes can be predicted which are associated with the diseases.

3. Results

3.1. Evaluation parameters

In this section, we utilize the disease gene association network that is included inside DisGeNet as a standard dataset to assess the

effectiveness of PSNE. In addition, a number of traditional well-performance disease gene prediction methods, such as dgn2vec, PRINCE (Vanunu et al., 2010), CIPHER (Wu et al., 2008), BiRW (Xie et al., 2015), RWR (Köhler et al., 2008) and DK (Köhler et al., 2008) are used as comparison.

In the comparison of performance, all other genes are used as control group except for genes related to diseases in training and testing groups, which are called genome-wide control set. AUROC, AUPRC and top-K Recall and Precision then are used as evaluation parameters (Vihinen, 2012). AUROC is defined as the area under the receiver's operating characteristic curve and AUPRC is defined as the area under the precision-recall curve, the two parameters are able to use the intuitive indications of the data in order to assess the performance of the classifiers, and effective reflection of predictive sequencing of disease-associated genes is possible. The greater score means that the performance of the classifier is better. For the disease d in the disease set D, T_D represents the set of genes in the test set. The disease gene prediction algorithm will provide a ranking list of candidate genes for disease d. We use $R_D(k)$ to represent the collection of the first k candidate genes in the ranking list. Then, Recall in the top-k ranking list is set as $\text{Recall} = \left|T_d \cap R_d(k)\right| / |T_d|$, and Precision in the top-k ranking list is set as $\text{Precision} = \left|T_d \cap R_d(k)\right| / \left|R_d(k)\right|$.

3.2. Overall comparison

Figure 2 shows the AUROC and AUPRC values of PSNE, dgn2vec, RWRH, PRINCE, CIPHER, BiRW, RWR and DK where AUROC values are 0.868, 0.867, 0.856, 0.821, 0.628, 0.768, 0.653, 0.641; AUPRC values are 0.125, 0.121, 0.078, 0.032, 0.016, 0.046, 0.031, and 0.032, respectively. Compared with other advanced methods, PSNE has different degrees of improvement, with a maximum increase of 27.5% (AUROC) and 74.8% (AUPRC). Table 1 shows that the top-k Recall rates for PSNE are 0.078, 0.176, 0.227, 0.287, 0.359, 0.418, and 0.499. In the top-k Recall rate, PSNE is better than all comparison methods.

3.3. Comparison of the results and case study for Alzheimer's disease

Tables 2, 3 show that the top-k Recall and Precision rates of PSNE and other methods. It is also obvious from the figure that the performance of PSNE is superior to other methods. To further illustrate the performance of PSNE, we used PSNE method to predict and analyze the genes that may cause AD. Table 4 shows the top 20 Alzheimer's-related genes predicted by PSNE and the corresponding literature support. Progranin encoded by GRN gene, plays a key role in the development, survival, function and maintenance of neurons and microglia in mammalian brain. GRN functional deletion mutations cause neuronal waxy lipofuscinosis or frontotemporal dementia-GRN (FTD-GRN) in a dose-dependent manner. Mutations that lower PGRN levels increase the risk of AD (Rhinn et al., 2022). The implication of genome-wide significant differential methylation of OXT, encoding oxytocin, in two independent cohorts indicates it is a promising target for future studies on early biomarkers and novel therapeutic strategies in AD (Lardenoije et al., 2019). Ma et al. (2015) suggested that targeting TYROBP might provide a new opportunity for the treatment of AD based on its potential protective role in the pathogenesis of AD. The report from Marioni et al. (2019) evidence that the association of SNPs in the TOMM40 gene with AD is potentially mediated by both gene expression and DNA methylation in the prefrontal cortex.

3.4. Comparison of the results and case study of Parkinson's disease

Tables 5, 6 show that the top-k Recall and Precision rates of PSNE and other methods. Except that Recall-1 and Precision-1 is not higher than dgn2vec and RWRH, the effect of PSNE is better than other methods. Generally speaking, PSNE also has advantage in the prediction of Parkinson's-related genes.

Likewise, we used the PSNE method to predict and analyze the genes that may cause PD. Table 7 shows the top 20 Parkinson's-related genes predicted by PSNE and the corresponding literature support. In late-onset disease which is most reminiscent of idiopathic PD subtle deficits in endosomal receptor-sorting/recycling are highlighted by the discovery of pathogenic mutations DNAJC13. Molecular deficits in these processes are genetically linked to the phenotypic spectrum of Parkinsonism associated with Lewy body pathology (Vilariño-Güell et al., 2014). Nitric oxide synthase (NOS) genes (NOS1, NOS2A, and NOS3) may create excess nitric oxide that contributes to neurodegeneration in Parkinson's disease (PD). NOS genes might also interact with one another or with environmental factors in PD (Hancock et al., 2008). The protein product of the nuclear-encoded POLG gene plays a key role in the maintenance of mitochondrial DNA replication, and its failure causes multi-system diseases with varying severity. It is known that mitochondrial dysfunction in Parkinson's disease plays a key role in the loss of dopaminergic neurons in the substantia nigra. Therefore, changes in the POLG gene may influence the development of various hereditary neurodegenerative diseases, including monogenic Parkinsonism (Illés et al., 2020).

4. Conclusion

In the context of the globalization of AD and PD, it is crucial to identify and predict the pathogenic genes of AD and PD for disease prevention and treatment. In this paper, we first combined a disease-gene network, disease-disease network, and human protein network to build a heterogeneous network model, used a network embedding algorithm to achieve low dimensional vector representation of the network. In network embedding algorithm, nodes tended to walk along the edge with the highest transition probability to their neighbors. Assuming that a node had n neighbors, it took $O(N)$ time to find the exit edge with the highest transition probability. Then, a new two-layer heterogeneous network of diseases and genes was constructed. Finally, the network propagation algorithm was used to predict the disease genes. Unlike previous methods of referring to the

topological features of heterogeneous protein networks or learning features from gene sequences, this method represented nodes in heterogeneous networks as potential feature vectors. It used network embedding to maximize cross-relationships. We applied the network embedding algorithm to the constructed dataset, and the results showed that our method can achieve better prediction performance. At the same time, we used this method to predict the candidate genes related to AD and PD and carried out literature verification through the PubMed website. We confirmed that most of the predicted candidate genes correlate with AD and PD. In addition, a small number of candidate genes had not been proven on the PubMed website, but at the same time, there was no objection. Perhaps these tiny numbers of genes without examples could provide helpful ideas for the medical research of AD and PD.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/TianTianTian14/PSNE.

Author contributions

JM and JX conceived, designed, managed, and reviewed the manuscript. TQ performed the experiments and drafted the manuscript. All authors approved the final manuscript.

Funding

This work was supported by the Training Program for Excellent Young Innovators of Changsha (Grant Nos. kq2206056, kq1802024, kq1905045, kq2009093, and kq2106075), Hunan Provincial Natural Science Foundation of China (Grant No. 2018JJ3568), the National Natural Science Foundation of China under Grant (No. 71871233), and Science and Technology Project of Hebei Education Department under Grant (No. ZD2022031).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ascherio, A., and Schwarzschild, M. A. (2016). The epidemiology of Parkinson's disease: risk factors and prevention. *Lancet Neurol.* 15, 1257–1272. doi: 10.1016/S1474-4422(16)30230-7

Ata, S. K., Wu, M., Fang, Y., Le, O. Y., and Li, X. L. (2021). Recent advances in network-based methods for disease gene prediction. *Brief. Bioinform.* 22, bbaa303. doi: 10.1093/bib/bbaa303

Dai, W., Chang, Q., Peng, W., Zhong, J., and Li, Y. (2019). "Identifying human essential genes by network embedding protein-protein interaction network," in *Bioinformatics Research and Applications: 15th International Symposium, ISBRA 2019* (Barcelona, Spain), 127–137. doi: 10.1007/978-3-030-20242-2_11

Grover, A., and Leskovec, J. (2016). *node2vec: Scalable Feature Learning for Networks*. New York, NY: ACM. doi: 10.1145/2939672.2939754

Hancock, D. B., Martin, E. R., Vance, J. M., and Scott, W. K. (2008). Nitric oxide synthase genes and their interactions with environmental factors in Parkinson's disease. *Neurogenetics* 9, 249–262. doi: 10.1007/s10048-008-0137-1

Hohmann, S. (2010). Unicellsys-understanding the cell's functional organization. *J. Biotechnol.* 150, 545. doi: 10.1016/j.jbiotec.2010.09.902

Hu, K., Hu, J. B., Tang, L., Xiang, J., Ma, J. L., Gao, Y. Y., et al. (2018). Predicting disease-related genes by path structure and community structure in protein-protein networks. *J. Stat. Mech. Theory Exp.* 2018, 100001. doi: 10.1088/1742-5468/aae02b

Illés, A., Balicza, P., Gál, A., Pentelényi, K., Csabán, D., Gézsi, A., et al. (2020). Hereditary Parkinson's disease as a new clinical manifestation of the damaged polg gene. *Orvosi Hetilap.* 161, 821–828. doi: 10.1556/650.2020.31724

Jevtic, S., Sengar, A. S., Salter, M. W., and McLaurin, J. (2017). The role of the immune system in Alzheimer's disease: etiology and treatment. *Ageing Res. Rev.* 40, 84–94. doi: 10.1016/j.arr.2017.08.005

Joe, E., and Ringman, J. M. (2019). Cognitive symptoms of Alzheimer's disease: clinical management and prevention. *BMJ* 367, L6217. doi: 10.1136/bmj.l6217

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Human Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Lardenoije, R., Roubroeks, J. A. Y., Pishva, E., Leber, M., Wagner, H., Iatrou, A., et al. (2019). Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood. *Clin. Epigenet.* 11, 164. doi: 10.1186/s13148-019-0755-5

Li, S., Qu, L., Wang, X., and Kong, L. (2021). Novel insights into ripk1 as a promising target for future Alzheimer's disease treatment. *Pharmacol. Therap.* 231, 107979. doi: 10.1016/j.pharmthera.2021.107979

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btq108

Liu, Y., Guo, Y., Liu, X., Wang, C., and Guo, M. (2021). Pathogenic gene prediction based on network embedding. *Brief. Bioinform.* 22, bbaa353. doi: 10.1093/bib/bbaa353

Ma, J., Jiang, T., Tan, L., and Yu, J.-T. (2015). Tyrobp in Alzheimer's disease. *Mol. Neurobiol.* 51, 820–826. doi: 10.1007/s12035-014-8811-9

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., et al. (2019). Correction: GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* 9, 161. doi: 10.1038/s41398-019-0498-2

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347, 1257601. doi: 10.1126/science.1257601

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Adv. Neural Inform. Process. Syst.* 26, 2–3. doi: 10.48550/arXiv.1310.4546

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. doi: 10.1145/2623330.2623732

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., et al. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* 30, 1591–1601. doi: 10.1002/mds.26424

Rhinn, H., Tatton, N., Mccaughey, S., Kurnellas, M., and Rosenthal, A. (2022). Progranulin as a therapeutic target in neurodegenerative diseases. *Trends Pharmacol. Sci.* 43, 641–652. doi: 10.1016/j.tips.2021.11.015

Ruan, P., and Wang, S. (2021). Disnep: a disease-specific gene network enhancement to improve prioritizing candidate disease genes. *Brief. Bioinform*. 22, bbaa241. doi: 10.1093/bib/bbaa241

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "Line: Large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*, 1067–1077. doi: 10.1145/2736277.2741093

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R., and Wasserman, W. W. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*. 6, e1000641. doi: 10.1371/journal.pcbi.1000641

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genom*. 13(Suppl 4), S2. doi: 10.1186/1471-2164-13-S4-S2

Vilariño-Güell, C., Rajput, A., Milnerwood, A. J., Shah, B., Szu-Tu, C., Trinh, J., et al. (2014). DNAJC13 mutations in Parkinson disease. *Human Mol. Genet*. 23, 1794–11801. doi: 10.1093/hmg/ddt570

Wang, Y., Yao, Y., Tong, H., Xu, F., and Lu, J. (2019). A brief review of network embedding. *Big Data Mining Anal*. 2, 35–47. doi: 10.26599/BDMA.2018.9020029

Wirdefeldt, K., Adami, H.-O., Cole, P., Trichopoulos, D., and Mandel, J. (2011). Epidemiology and etiology of Parkinson's disease: a review of the evidence. *Eur. J. Epidemiol*. 26, s1–s58. doi: 10.1007/s10654-011-9581-6

Wu, X., Jiang, R., and Zhang, M. Q. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol*. 4, 189. doi: 10.1038/msb.2008.27

Xiang, J., Meng, X., Zhao, Y., Wu, F. X., and Li, M. (2022a). HYMM: hybrid method for disease-gene prediction by integrating multiscale module structure. *Brief. Bioinform*. 23, bbac072. doi: 10.1093/bib/bbac072

Xiang, J., Zhang, J., Zhao, Y., Wu, F. X., and Li, M. (2022b). Biomedical data, computational methods and tools for evaluating disease associations. *Brief. Bioinform*. 23, bbac006. doi: 10.1093/bib/bbac006

Xiang, J., Zhang, J., Zheng, R., Li, X., and Li, M. (2021a). NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Brief. Bioinform*. 22, bbab080. doi: 10.1093/bib/bbab080

Xiang, J., Zhang, N. R., Zhang, J. S., Lv, X. Y., and Li, M. (2021b). Prgefne: predicting disease-related genes by fast network embedding. *Methods* 192, 3–12. doi: 10.1016/j.ymeth.2020.06.015

Xie, M. Q., Xu, Y. J., Zhang, Y. G., Hwang, T. H., and Kuang, R. (2015). Network-based phenome-genome association prediction by bi-random walk. *PLoS ONE* 10, e0125138. doi: 10.1371/journal.pone.0125138

Zeng, X., Zhang, X., Liao, Y., and Pan, L. (2016). Prediction and validation of association between microRNAs and diseases by multipath methods. *Biochim. Biophys. Acta* 1860, 2735–2739. doi: 10.1016/j.bbagen.2016.03.016