Integrate multiscale module kernel for disease-gene discovery in biological networks

Ju Xiang, Xiangmao Meng, Ruiqing Zheng, Min Li*

Abstract—Biomedical data mining is very important for the research of complex diseases, and disease-gene discovery is one of the most representative topics in this field. Multiscale module structure (MMS) that widely exists in biological networks can provide useful insight for disease research. However, how to effectively mine information in MMS to enhance the ability of disease-gene discovery is challenging. Thus, we propose a type of novel hybrid methods (HyMSMK) for disease-gene discovery by integrating multiscale module kernel (MSMK) derived from multiscale module profile (MSMP). We extract MSMP with local to global structural information from comprehensive human protein interactome by multiscale modularity optimization with exponential sampling, and construct MSMK by using the MSMP as a feature matrix, combining with the relative information content of features and kernel sparsification. Then, we present several fusion strategies integrating MSMK, including a probabilistic model for rank aggregation. By a series of experiments, we study the effect of the fusion strategies and kernel sparsification on HyMSMK, and demonstrate that HyMSMK outperforms the state-of-art network-based algorithms. These results confirm that MSMK is particularly helpful for disease-gene discovery, and the kernel sparsification can improve HyMSMK in storage space and computing speed. This may provide useful insights for the study and application of MMS.

Index Terms—Complex networks, Community detection, Multiscale module structure, Biological networks, Disease-gene discovery

1 Introduction

BIOMEDICAL data mining is very important for the research of biomedicine, and the rapid accumulation of networked biomedical data (e.g., protein-protein interactions [1], disease-gene associations [2], and disease-phenotype\symptom associations [3, 4]) promotes the development of network medicine [5-7]. Many network-based mining strategies have been proposed to explore unknown knowledge from the networked biomedical data, and the discovery of disease-related genes (also known as disease-gene prediction) is one of the most representative topics in this field [8-12].

Many diseases are closely related to the mutations and dysfunctions of relevant genes, and they are usually considered as the consequence of local neighborhood perturbation in biomolecular networks [13]. Many computational methods for mining disease-associated genes have been developed, and network-based methods are one of the best popular strategies [12, 14-17]. Genes associated with the same disease are often more similar in function, and they usually have more physical interactions or functional correlations, which are encoded in complex biomolecular networks [18]. It is a natural and effective approach to mine unknown disease-associated genes by direct interactions, network distance or closeness between candidates and seeds (i.e., genes known to be associated with specific disease) [19, 20]. Network propagation, such as random walk with restart, can more effectively explore the neighborhood topology around seeds in biological networks [19, 21, 22], so it has been widely applied to the discovery of disease-related genes as well as related issues [10, 14, 19, 23, 24], and it also has been extended to more sophisticated network models such as multigraph model, multiple network model and multiplex network model in homogeneous or heterogeneous networks, which can integrate multiple types of biological networks [14, 17, 24, 25]. However, partly due to data noises in biological networks, there is still room in these methods, and they can be further improved, e.g., by such approaches as network enhancement [11, 26].

The existence of module structure (or say community structure) in biomolecular networks can provide insights for human disease research [27-31]. Some researchers have conducted the analysis of network modules to find relevant enlightenment for disease research [28, 32-35], but network module mining or say community detection is still an open issue [36-44]. Many algorithms for network module mining have intrinsic limitations such as the resolution limit [45-47]. As a result, some methods may generate large-scale modules, which may contain many genes unrelated to the disease. Or they may also produce many fragmented disease-related modules, which may lead to the loss of some candidate genes that are possibly related to the disease. Some complex diseases may indeed involve multiple different network modules, since multiple genes related to the diseases may possess distinct functions, playing distinct roles in the diseases [48], but these modules may be of relatedness at a higher level of functions while not completely independent. Modules at distinct scales may provide relevant information at distinct levels. Therefore, multi-scale analysis of modules may provide more comprehensive perspective for the

- *J. Xiang, X. Meng, R. Zheng and M. Li are with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China. Email: {xiangju@csu.edu.cn, mxmanhui@csu.edu.cn, rqzheng@csu.edu.cn, limin@mail.csu.edu.cn}.*
- *J. Xiang is also with Department of Basic Medical Sciences, Changsha Medical University, Changsha, 410219, China. Email: xiang.ju@foxmail.com.*
- **Corresponding author: Min Li.*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

research of complex diseases and their genes.

Multiscale module structure (MMS) is widely present in biomolecular networks, which are composed of subnetworks that can be gradually decomposed into smaller subnetworks in a deep hierarchical structure [49]. Typically, some complexes may consist of multiple secondary complexes, and the hierarchical structure of gene function annotations also implies the existence of functional modules at multiple scales [50-52]. For example, Lewis et al have studied the correlation between modules at different scales and protein functions, showing the dependence of different important scales of modules on proteins and processes under study [53]; Wang et al showed that the hierarchical organization of uncovered functional modules by a fast hierarchical clustering algorithm approximately corresponded to the GO hierarchy [54]. Overall, the MMS can provide more abundant information from local to global structure for disease research, but how to make use of the MMS information is still a challenging topic in disease-gene discovery.

In order to utilize the MMS to improve the ability of mining disease-associated genes, we propose a type of novel hybrid methods (HyMSMK) by integrating multiscale module kernel (MSMK) derived from multiscale module profile (MSMP), where MSMP contains information from local to global structure that is extracted from a comprehensive human protein interactome by multiscale modularity optimization with exponential sampling (MMOES), and MSMK is constructed by using the MSMP as a feature matrix, combining with the relative information content (IC) of features.

In the following, we will first introduce several datasets for three types of association data (disease-associated genes, gene-gene associations and disease-disease associations) as well as corresponding computational methods. Then, we present the details of HyMSMK, including the method for mining multiscale module profile (MSMP), the method for constructing multiscale module kernel (MSMK) from MSMP, and several funsion strategies integrating MSMK in disease-gene discovery. In section 3, the experimental settings are first introduced. Then, we study the effect of different integration strategies and kernel sparseness of MSMK on performance, and verify the performance of HyMSMK, by a series of experiments including cross-validations and independent tests. Furthermore, we conduct case studies for several diseases (e.g., Alzheimer's disease and Breast cancer), which provide sets of candidate genes for these diseases and verify the disease relatedness of the genes. Finally, we come to the conclusion.

2 Matrieals and Methods

We will introduce three types of association network data, including computational methods. Then, we present the details of HyMSMK (see **Fig. 1**), including several variants of HyMSMK (see **TABLE 1**).

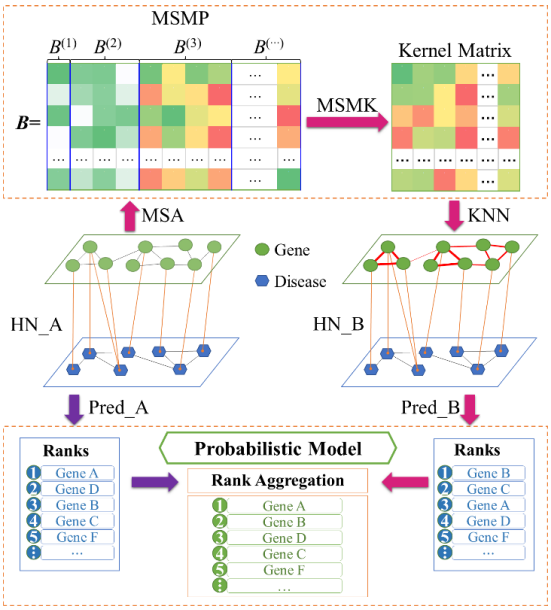

Fig. 1. Workflow of HyMSMK for integrating multiscale module kernel by rank aggregation.

2.1 Datasets for disease-gene associations

We will use three datasets for disease-gene associations (DGA) in our experimental study. (1) Dataset1: the first one is an integrated dataset of disease-associated genes retrieved from GWAS and OMIM [55, 56], in which MeSH was used to integrate the disease nomenclatures of different sources and only diseases with ≥ 20 associated genes were retained in the original literature [35, 57]. (2) Dataset2: the second dataset is obtained just from the OMIM database, an authoritative database of human genetic diseases/traits and causal genes, containing a large amount of knowledge about human genes and genetic diseases/disorders [56]. (3) Dataset3: the third dataset is obtained from DisGeNet, which is one of the largest disease-associated gene collections publicly available [2].

2.2 Datasets for gene-gene associations from multiple sources

The protein-protein interactions are important sources for revealing intrinsic gene-gene associations (GGA). Due to the incompleteness and noises of data in single sources, we utilize a comprehensive protein interactome that were derived from the several data sources [57], such as BioGRID [1], MINT [58], IntAct [59], and HPRD [60].

2.3 Estimation of disease-disease associations

(1) Disease-disease similarity measurement based on co-occurrence of diseases and symptoms

Symptoms represent the actually observed high-level manifestations of a disease, which are important for clinical diagnosis and treatment; and the symptom-based disease similarity has been showed to be closely associated to their underlying molecular interactions [4]. So, we will make use of the disease-disease associations that are estimated based on the disease-symptom associations derived from text mining.

Here, the diseases-symptom associations (DSA) are

quantified based on co-occurrence of MeSH terms, i.e., the number of PubMed literature in which MeSH terms occur together [56]. Each disease can be described by a vector of symptoms,

$$d_x = (w_{x,1}, w_{x,2}, \ldots, w_{x,j}, \ldots, w_{x,n})^T, \quad (1)$$

where $w_{x,j}$ denotes the strength of DSA association between disease x and symptom j. The symptom overlaps of diseases can itself reflect the associations between diseases, but different symptoms, such as pain and stupor, generally have different prevalence in diseases, and there may be publication biases towards some diseases. To deal with the heterogeneity, the association strength between disease x and symptom j is often estimated by the term frequency-inverse document frequency (TF-IDF) [61],

$$w_{xj} = C_{x,j}\, log(\, x/x_j), \quad (2)$$

where C_{xj} is their absolute co-occurrence; N denotes the number of diseases and $N_j \geq 1$ denotes the number of diseases with symptom j. Finally, the association scores between diseases are quantified by the cosine similarity between the normalized symptom vectors of diseases,

$$S_D(d_x, d_y) = \langle d_x, d_y \rangle / \sqrt{\langle d_x, d_x \rangle \langle d_y, d_y \rangle}, \quad (3)$$

where $\langle d_x, d_y \rangle$ denotes the dot product of two vectors. Note that the above process does not consider the intrinsic correlations between different symptoms.

(2) Disease-disease similarity measurement based on human phenotype ontology

Each of human diseases is often associated to a set of phenotypic abnormalities. The Human Phenotype Ontology (HPO) database provides human phenotype annotations (HPA) for diseases, corresponding to a set of HPO terms, and the HPO terms are organized as a directed acyclic graph (DAG) [62]. Different from the above co-occurrence between diseases and symptoms, this provides the more important information of phenotypic abnormalities involved in human diseases, suggesting another possible way to uncover the associations between diseases.

For the disease-gene dataset from OMIM [56], we evaluate the associations between diseases by using the data in the HPO database [63], since it provides human phenotype annotations for disease entries in the OMIM database. The phenotypic similarity between a pair of diseases can be calculated by the number of their shared HPO terms, but some HPO terms are not as informative compared to others. Two diseases sharing a very rare HPO term should be considered as more similar than those sharing a very common HPO term. Therefore, the relevance of each HPO term in the HPO database is measured by the relative information content (IC) of HPO terms [64],

$$IC(t) = -\, log\, f_t, \quad (4)$$

where f_t denotes the frequency of the HPO term t within the HPA set. The similarity between two HPO terms (t_1 and t_2) (HPS) is given by,

$$HPS(t_1, t_2) = \max_{v \in anc(t_1) \cap anc(t_2)} IC(v), \quad (5)$$

where $anc(t_1)$ denotes the ancestor terms of phenotype t_1 on the HPO structure. Finally, the phenotypic similarity between two diseases (d_x and d_y) that are represented by two sets of HPO terms, T_x and T_y, is measured by,

$$S_D(d_x, d_y) = \frac{1}{2}\Big(\frac{1}{|T_x|}\sum_{t_1 \in T_x} \max_{t_2 \in T_y} HPS(t_1, t_2) + \frac{1}{|T_y|}\sum_{t_2 \in T_y} \max_{t_1 \in T_x} HPS(t_1, t_2)\Big). \quad (6)$$

Phenotypic similarity scores between a pair of diseases are determined by the overall rareness of the pair's shared terms. This disease-disease similarity measurement not only uses the HPA annotations, but also makes use of the intrinsic relations between HPO terms in the DAG of the HPO database.

2.4 Mining of multiscale module profile using multiscale modularity optimization with exponential sampling

Multiscale module structure (MMS) widely exists in biological networks and the mining of MMS is an important topic [36, 65]. We will mine multiscale module profiles by using multiscale modularity optimization with exponential sampling (MMOES) (see **Fig. 1**) [66, 67].

Multiscale modularity optimization (MMO) extracts module partitions in a network by searching for optimal solution of modularity Q, which is a widely used target function for module/community detection [66-68]. Given a module partition in a network, the mathematical form of general modularity Q can be expressed as,

$$Q = \sum_s (e_{ss} - \gamma a_s^2), \quad (7)$$

where e_{st} denotes the fraction of links between modules s and t,

$$e_{st} = \frac{1}{2m}\sum_{i,j} A_{ij}\, \delta(c_i, s)\, \delta(c_j, t)\ , \quad (8)$$

while $e_{ss} = \frac{k_s^{in}}{2M}$ (k_s^{in} is the inner degree of module s),

$$a_s = \frac{1}{2m}\sum_{ij} A_{ij}\, \delta(c_i, s) = \frac{k_s}{2m}, \quad (9)$$

m denotes the number of links in the network; A is the adjacent matrix of the network; c_i denotes the module label of node i; γ is the resolution parameter; the sum over all modules in the network. Modules at different scales can be detected by using effective algorithms to optimize the above modularity Q with different values of resolution parameter γ [65, 69, 70]. Small-size modules can be discovered if γ-value is large; large-size modules can be discovered if γ-value is small. A group of single-node modules can be generated when γ-value is large enough, e.g., $\gamma > 2m/k_{min}^2$, where k_{min} denotes minimum degree of nodes in the network [71].

Then, to cover all possible module scales, we define a semi-empirical interval of $\gamma \in [\gamma_{min}, \gamma_{max}]$, and use exponential sampling (ES) to generate a set of γ-values from continuous resolution space, since it can sample different meaningful scales in a network reasonably, according to previous work [67, 69]. The exponential sampling will produce a set of γ-values that are equally spaced on a logarithmic scale.

By MMOES with the set of sampled γ-values, we will obtain a series of meaningful module partitions from local to global scales in a network, which contain information at different scales that may be helpful for mining gene-gene and disease-gene associations.

To effectively organize the module partitions from local

to global scales generated by MMOES, we firstly construct a partition matrix $B^{(h)}$ for module partition at each scale, with $B_{js}^{(h)}$ denoting if gene j belongs to module s in the module partition, and then we define a multiscale module profile (MSMP) by integrating all the partition matrices,

$$B = (B^{(1)}, B^{(2)}, \dots, B^{(\hbar)}, \dots), \tag{10}$$

which contains multiscale information hidden in a network(see **Fig. 1**).

2.5 Construction of multiscale module kernel

(1) Multiscale module kernel

MSMP can be regarded as a multiscale feature matrix about module structure in a network, where a module at certain scale can be viewed as a feature at this scale, and each column of MSMP records the information of this feature.

Features of MSMP at different scales contain different levels of information helpful for revealing gene-gene associations, but, similarly to the above DSA associations and HPA annotations in form, some features of MSMP are not as informative as others. If two genes can share very rare MSMP features, they should be considered to be more similar than those sharing only very common MSMP features. In fact, the rarer MSMP features generally correspond to local scale modules at higher resolution, which generally have higher link densities and higher functional consistency, and thus genes within these modules may be more similar than genes within other modules. We therefore measure the relevance of MSMP features by the relative information content (IC) [64], $IC(s) = -log\, f_s$, where $f_s = n_s/N$ denotes the ratio of nodes belonging to module s in a network. Similarly, a similarity matrix of genes can be generated in the way mentioned above, but we will use the kernel function to express the gene-gene associations derived from MSMP.

For convenience of calculation, we first define a diagonal matrix D_{IC} of the IC values of MSMP features in a network by,

$$(D_{\mathrm{IC}})_{ss} = -\log\left(\sum_j I\left(B_{j,s}\right)/N\right), \tag{11}$$

where $(D_{IC})_{ss}$ is the IC value of module s; $I(x)$ is a counting function: $I(x) = 1$ if $x > 0$; $I(x) = 0$ otherwise. Then, a multiscale module kernel (MSMK) can be constructed naturally by,

$$K_{MSMK} = \hat{B}\hat{B}^T = BD_{IC}^2B^T, \tag{12}$$

where $\hat{B} = BD_{IC}$ can be viewed as a normalized multiscale module profile (nMSMP). Further, to normalize the multiscale module kernel, we define $\hat{K}_{MSMK} = D_{NM}K_GD_{NM}$, where D_{NM} is a diagonal matrix with diagonal elements $(D_{NM})_{ii} = 1/\sqrt{(K_{MSMK})_{ii}}$, so that its elements are between 0 and 1, corresponding to the similarity scores between genes.

(2) Process of kernel sparseness

The multiscale module kernel contains gene-gene relations discovered from MSMP, but it, as a full matrix, is expensive in storage and computing. So, we will transform it into a sparse matrix by retaining k-nearest neignbors of genes, that is, by retaining only the connections between each gene and its k-nearest neighbors (i.e., its k neighbors with highest scores). This is helpful for filtering out data noise while retaining the critical information of gene-gene associations derived from MSMP. The degree of kernel sparseness is inversely correlated with the number of the k-nearest neighbors (knn) in kernel sparseness process.

This process of kernel sparseness can be described as follows. For each gene i, we first generate a ranking vector of neighbor genes,

$$\vec{r}_i = (r_{i1}, r_{i2}, \dots, r_{ij}, \dots)^T, \tag{13}$$

by the descending order of the similarity scores between this gene and its neighbor genes, where r_{ij} is the ranking index of neighbor gene j of gene i. Note that node i itself is given a largest ranking index. Then, a ranking matrix can be constructed by the union of the ranking vectors about all genes, $R = (\vec{r}_1, \vec{r}_2, \dots, \vec{r}_i, \dots)$. Finally, we can generate a new network $G = (V, E, W)$ by using the multiscale module kernel matrix and the ranking matrix, where V, E and W denote the sets of genes, edges and edge weights in the network, respectively,

$$E = \{(i,j) |\, R_{ij} \le k \text{ or } R_{ji} \le k\ \}, \tag{14}$$

$$W = \{S_{ij} | (i,j) \in E\}, \tag{15}$$

k is the number of the nearest neighbors and $S_{ij} = (\hat{K}_{MSMK})_{ij}$ denotes the similarity score between genes i and j. The adjacent matrix of G denotes as M_{MSMK}, whose elements indicate the weights of edges in G.

TABLE 1

THREE TYPES OF DISEASE-GENE-DISCOVERY METHODS THAT INTEGRATE INFORMATION IN MULTISCALE MODULE KERNEL.

Method	Description
HyMSMK_CPF (CPF1/CPF2/CPF3)	Two heterogeneous networks are constructed, one of which contains MSMK, and two predictions from them are integrated by a probabilistic model for rank aggregation.
HyMSMK_FS	A heterogeneous network that contains a fusion network of GGA and MSMK is used.
HyMSMK_MG	A multigraph heterogeneous network that contains a multigraph of GGA and MSMK is used.

2.6 Methods integrating multiscale module kernel

Gene-gene associations are critical to network-based methods for disease-gene discovery. The MSMK above provides a type of new gene-gene associations revealed from the view of MMS in a network. It may contain complementary information that is helpful for the discovery of disease genes. Here, we propose three types of methods that integrate the information from multiscale module kernel (see **TABLE 1**): (1) three hybrid methods integrating MSMK by rank aggregation, (2) a hybrid method integrating MSMK by a fusion network model, as well as (3) a hybrid method integrating MSMK by a multigraph model. Especially, we focus on the hybrid methods based on rank aggregation.

(1) Hybrid methods integrating MSMK by a probabilistic model for rank aggregation (HyMSMK_CPF). We construct

two gene-disease heterogeneous networks: HN_A and HN_B. In HN_A, original GGA is connected with DDA by DGA, and in HN_B, MSMK is connected with DDA by DGA. Then, we generate two prediction lists of genes for specific disease by applying the random walk process to HN_A and HN_B, respectively. Finally, the two prediction lists of genes are integrated by a probabilistic model for rank aggregation. See **Fig. 1** and **Algorithm 1** for details.

To perform the random walk process on a heterogeneous network, we construct a transition matrix of the heterogeneous network by using two intra-layer transition matrices and two inter-layer transition matrices between diseases and genes. The details are as follows.

Intra-layer transition matrices. The intra-layer transition matrix at the gene-gene layer is defined by

$$T_G = M_G D_G^{-1}, \tag{16}$$

where M_G is the adjacent matrix of GGA (or MSMK), and D_G is a diagonal matrix with diagonal elements $M_G^T\mathbf{1}$ (note that $\mathbf{1}$ is an all-one vector).

Similarly, the intra-layer transition matrix at the disease-disease layer is defined by $T_D = M_D D_D^{-1}$, where M_D is the adjacent matrix of DDA, and D_D is a diagonal matrix with diagonal elements $M_D^T\mathbf{1}$.

Inter-layer transition matrices. The inter-layer transition matrix from diseases to genes is defined by

$$T_{GD} = M_{GD}D_{GD}^{-1}, \tag{17}$$

where M_{GD} is the gene-disease association matrix: $M_{GD}(j,i) = 1$ if gene j is associated with disease i, and $M_{GD}(j,i) = 0$ otherwise; D_{GD} is a diagonal matrix with diagonal elements $M_{GD}^T\mathbf{1}$.

Similarly, the inter-layer transition matrix from genes to diseases is defined by $T_{DG} = M_{DG}D_{DG}^{-1}$, where $M_{DG} = M_{GD}^T$ is the disease-gene association matrix, and D_{DG} is a diagonal matrix with diagonal elements $M_{DG}^T\mathbf{1}$.

Transition matrix of heterogeneous network. Finally, we calculate the transition matrix of the heterogeneous network by

$$T_H = M_H D_H^{-1}, \tag{18}$$

where D_H is a diagonal matrix with diagonal elements $M_H^T\mathbf{1}$, and M_H denotes the raw transition matrix of the heterogeneous network that is constructed by,

$$M_H = \begin{pmatrix} (1-\beta)T_G & \beta T_{GD} \\ \beta T_{DG} & (1-\beta)T_D \end{pmatrix}, \tag{19}$$

where β is the jumping probability between network laryers.

Based on the transition matrix T_H, we will calculate stationary probability vectors of random walker reaching different nodes (genes and diseases) for the random walk process on the heterogeneous network. Generally, the closed solution of the stationary probability vectors in the random walk process can be written as,

$$\begin{pmatrix} p_G \\ p_D \end{pmatrix} = \alpha\left[I - (1-\alpha)T_H\right]^{-1}\begin{pmatrix} p_G^{(0)} \\ p_D^{(0)} \end{pmatrix}, \tag{20}$$

where $p_G^{(0)}$ and $p_D^{(0)}$ are the initial probability vectors of random walker reaching different gene nodes and disease nodes, respectively; the element of $p_D^{(0)}$ for studied disease is equal to 1, while zeros for other diseases; α is the restart probability. To avoid the inverse of matrix, we can obtain the stationary probability vectors by an iterative process,

$$\begin{pmatrix} p_G^{(t)} \\ p_D^{(t)} \end{pmatrix} = (1-\alpha)T_H\begin{pmatrix} p_G^{(t-1)} \\ p_D^{(t-1)} \end{pmatrix} + \alpha\begin{pmatrix} p_G^{(0)} \\ p_D^{(0)} \end{pmatrix}, \tag{21}$$

where $p_G^{(t)}$ and $p_D^{(t)}$ are the probability vector of random walker reaching different genes and diseases at the t-th step. The stationary probability scores of genes can be rewritten as $p_G = \lim_{t\to\infty}(p_G^{(t)})$. By a descending order of genes' stationary probability p_G, a ranking list of genes will be generated to prioritize genes associated to specific disease.

Algorithm 1 HyMSMK(M_G,M_D,M_{GD},$p_G^{(0)}$,$p_D^{(0)}$,knn,PMRA)

Input:
M_G, M_D, M_{GD}: gene-gene, disease-disease and gene-disease matrices;
$p_G^{(0)}$, $p_D^{(0)}$: initial probabilities of genes and diseases;
knn: # of the k-nearest neighbors;
PMRA: probabilistic model for rank aggregation.

Output: r_G: list of gene rankings

```
//*****Construct MSMK*****
B ← get MSMP from interactome by MMOES
K_MSMK ← get MSMK matrix by B
M_MSMK ← get sparse MSMK by K_MSMK and knn
//*****Construct transition matrices*****
T_MSMK ← intra-layer transition matrix of M_MSMK
T_G ← intra-layer transition matrix of M_G
T_D ← intra-layer transition matrix of M_D
T_GD ← inter-layer transition matrix of M_GD
T_DG ← inter-layer transition matrix of M_GD^T
//*****HN_A*****
T_HNA ← normalized matrix of [(1-β)T_G  βT_GD; βT_DG  (1-β)T_D]
p_A ← α[I-(1-α)T_HNA]^-1 [p_G^(0); p_D^(0)]
r_A ← get gene rankings from the descending order of p_A
//*****HN_B*****
T_HNB ← normalized matrix of [(1-β)T_MSMK  βT_GD; βT_DG  (1-β)T_D]
p_B ← α[I-(1-α)T_HNB]^-1 [p_G^(0); p_D^(0)]
r_B ← get gene rankings from the descending order of p_B
//*****Aggregation*****
r_G ← get fusion rankings of r_A and r_B by PMRA
return r_G
```

Model for rank aggregation. After generating two prediction lists of genes from HN_A and HN_B, we aggregate the results by the rank-aggregation strategy. Rank aggregation itself is an important topic [72]. Supposing that two prediction lists above are derived from two features $\{f_A, f_B\}$, respectively, we introduce a probabilistic model for rank aggregation based on Bayes theorem (PMRA). It can estimate the probability that gene i is related to a disease by,

$$P(x_i|f_A, f_B) = P(x_i)P\left(f_A, f_B|x_g\right)/P(f_A, f_B) \propto \prod_{h\in\{A,B\}} P(x_i|f_h) \tag{22}$$

when the prior probabilities $P(x_i)$ and $P(f_A, f_B)$ are assumed to be unknown. To estimate the conditional probabilities for two features (CPF), $P(x_i|f_A)$ and $P(x_i|f_B)$, we introduce the definitions of CPF based on gene rankings, since, for each prediction list of genes, a high ranking of a

gene usually means a high probability that it is associated with a disease. Here, we study three definitions of CPF based on gene rankings:

$$P(x_i|f_A) \propto (N - r_{Ai})/N,\ 1/r_{Ai},\ \text{and}\ \exp(-r_{Ai}), \quad (23)$$

which are denoted as CPF1, CPF2 and CPF3, respectively, where r_{Ai} is the ranking of gene i from HN_A. These definitions are also applied to $P(x_i|f_B)$. HyMSMK using CPF1, CPF2 and CPF3 are denoted as HyMSMK_CPF1/CPF2/CPF3, respectively.

(2) Hybrid method based on a fusion network (HyMSMK_FS). We first construct a fusion network (FS) by the union of gene-gene associations from GGA and MSMK, and connect the fusion gene-gene network with DDA by DGA to construct a gene-disease heterogeneous network. Then, we generate a prediction list of genes for specific disease by applying the above random walk process to the heterogeneous network.

(3) Hybrid method based on multigraph model (HyMSMK_MG). First, a gene-gene multigraph (MG) is constructed by using MSMK and GGA [25], and MG is connected with DDA by DGA to construct a multigraph heterogeneous network (MGHE). Then a random walk process on MGHE is applied to generate a prediction list of genes for specific disease.

To perform the random walk process on MGHE, we first generate the intra-layer transition matrix T_G of GGA by Eq.(16), and the intra-layer transition matrix of MSMK by,

$$T_{MSMK} = M_{MSMK} D_{MSMK}^{-1}, \quad (24)$$

where M_{MSMK} is the adjacent matrix of $MSMK$, and D_{MSMK} is a diagonal matrix with diagonal elements $M_{MSMK}^T \mathbf{1}$. Then, we generate a transition matrix of the multigraph for GGA and MSMK by,

$$\hat{T}_G = \left(T_G + T_{MSMK}\right)\hat{D}_G^{-1}, \quad (25)$$

where $\hat{D}_G$ is a diagonal matrix and its diagonal element $(\hat{D}_G)_{ii} = 2$ if node i is not isolated in both the two networks, or $(\hat{D}_G)_{ii} = 1$ if node i is isolated in either of the networks. Finally, we substitute T_G by $\hat{T}_G$ in Eq.(19) and use Eq.(18) to construct the probability transition matrix of the multigraph heterogeneous network. After the iterative process in Eq. (21) is conducted, stationary probability scores of genes are generated to prioritize genes associated to specific disease.

3 Experimental Results

We have proposed three types of HyMSMK methods based on different strategies: HyMSMK_CPF (CPF1/CPF2/CPF3), HyMSMK_FS and HyMSMK_MG. We will evaluate these methods in this section, and HyMSMK_CPF2 will be used as default, due to its excellent performance in many cases.

In the following, we will introduce the experimental settings. Then, by tranditional cross-validation experiments, we will study the effect of several fusion strategies and kernel sparsity on HyMSMK, and compare HyMSMK to the state-of-art network-based algorithms on two disease-gene datasets. Further, we apply HyMSMK to external dataset to exhibit the predictive performance of HyMSMK. Finally, we conduct several case studies, which show the disease relevance of predicted candidate genes by our method.

3.1 Experimental settings

To systematically evaluate our algorithms, we will use two evaluation strategies, two kinds of control sets, and three standard evaluation metrics.

(1) Two kinds of evaluation strategies. We will apply classical 5-fold cross-validation (5FCV) and independent test (IndTest). Given Dataset1 or Dataset2, 5FCV will randomly split known genes associated with a disease into 5 subsets. Each of these subsets in turn is used as a test set once, while the remaining subsets are combined into a training set. IndTest will use the disease genes in Dataset1 or Dataset2 as a training set, and the disease genes only belonging to Dataset3 as a test set. Moreover, we also show the performance of our method by case studies.

(2) Two kinds of control sets. A candidate set of genes composes of a test set and a control set. We apply two kinds of control sets of candidate genes: artificial linkage-interval control set (ALICS) and whole-genome control set (WGCS). ALICS will select ninety-nine genes adjacent to each test gene at the same chromosome [19]. WGCS will use all unknown genes as the control set, excluding the training set and test sets.

(3) Three kinds of evaluation metrics. According to a scoring/ranking list of candidate genes by algorithms, we will quantify the performance of the algorithms by three standard evaluation metrics (Recall, Precision and AUPRC). (a) Recall evaluates the ratio of found disease genes in the top-k ranking list of candidate genes relative to the test set. (b) Precision evaluates the probability of finding known disease genes in the top-k candidate list. Recall and Precision will vary with the length of the top-k list, providing an intuitive comparison for local performance of algorithms. (c) To comprehensively evaluate performance of algorithms, we will compute the area under the Precision-Recall curve (PRC) (AUPRC), where PRC curve has Precision and Recall on the y-axis and x-axis, respectively.

3.2 Performance analysis of HyMSMK under different fusion strategies

We have designed different fusion strategies to integrate MSMK, resulting in several variants of HyMSMK. Here, we compare the effects of these fusion strategies on predictive performance in different scenarios (**Fig. 2**). For Dataset1, the fusion strategy CPF2 is the best one in terms of any of three evaluation metrics: AUPRC, Recall and Precison, under both the ALICS and WGCS control sets. For Dataset2, CPF2 also can generate better or comparable results compared with other strategies. So, CPF2 is used as a default fusion strategy in this study.

Furthermore, we can see that the fusion strategies based on rank aggregation (RA) are superior to MG and FS in most cases. The possible reason is that FS and MG may lead to strong adverse interference in the fusion process of original GGA and MSMK, while RA treats each network as an independent system, preventing them from interfering with each other. For MG, the walkers randomly walk in two networks with equal probability at each step and the scores in two networks will be directly fused after each step. For FS, two networks have been integrated into one.

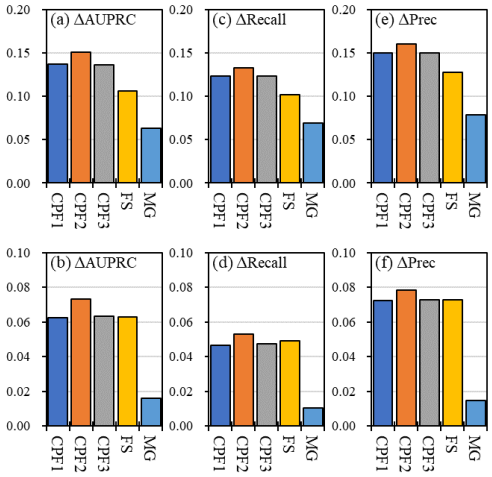

Fig. 2. In Dataset1 (MeSH), relative performance (Δ) of HyMSMK when using different fusion strategies for integrating multiscale module kernel, compared to the randon walk algorithm in HN_A. (a)-(b) Δ AUPRC, (c)-(d) Δ Recall, (e)-(f) Δ Prec. The first and second lines are under the ALICS and WGCS control sets, respectively.

3.3 Performance analysis of HyMSMK under different kernel sparseness

In the above experiments, to improve the storage/computing of MSMK and filter out its data noises, we have already performed a kernel sparseness process for MSMK by default. Here, we further study the effect of the kernel sparseness on HyMSMK in different scenarios.

In Dataset1 (**Fig. 3**), all the results under the two control sets show an obvious consistent growth trend: the predictive performance of HyMSMK gradually improves as the kernel sparseness decreases, that is, as the k-nearest neighbors of genes (knn) increase in the kernel sparseness process. When knn is small, the performance improvement of HyMSMK is more sensitive to the change of knn. This may be because the increase of knn starting from a small knn value makes the strongly correlated information increase rapidly, thereby bringing about a faster performance improvement. When knn is very large, the increase of knn will make each gene only add some neighbors with relatively low similarity to the gene. So, the growth of predictive performance becomes relatively flat.

According to the trend of performance curves in **Fig. 3**, knn=1000 is a good default choice for the tradeoff between performance and speed, since it can be seen as a turning point for performance growth. Moreover, MSMK has a greater improvement in performance than HyMSMK, with the increase of knn, under the WGCS control set.

In Dataset2, the performance of HyMSMK also improves rapidly with knn and reaches a relatively stable and high level. In this dataset, knn=1000 is also a good choice for HyMSMK, though the Recall has a slight decline at this point, and its AUPRC still has a certain growth with the increase of knn.

Moreover, note that the results of HyMSMK with knn=0 should be the same as that of the reference method in both datasets, and HyMSMK with knn=1000 outperforms the reference method.

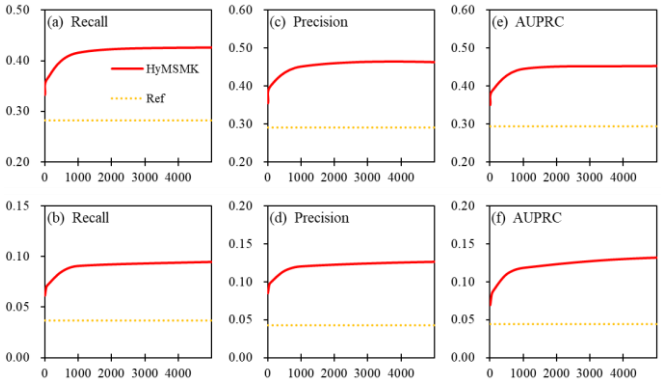

Fig. 3. In Dataset1 (MeSH), effect of kernel sparsity (knn) on performance of HyMSMK. (a)-(b) Recall, (c)-(d) Prec, (e)-(f) AUPRC. The first and second lines are under the ALICS and WGCS control sets, respectively.

We further test the space requirement and time consumption of HyMSMK for different kernel sparseness of MSMK in both sparse matrix representation mode (sMRM) and full matrix representation mode (fMRM). The results show that the memory usage of HyMSMK in sMRM increases approximately linearly with the increase of knn, and the time consumption has the same trend, which increases with the increase of space consumption. Especially, when knn is very small (e.g., knn≤1000), the memory usage in sMRM is obviously less than that in fMRM.

In conclusion, the kernel sparseness process for MSMK can not only effectively reduce the memory usages and time consumption of HyMSMK, speeding up its storage and computing, but also HyMSMK using sparsified MSMK (e.g., when knn=1000) can still achieve stable and good performance in disease-gene discovery.

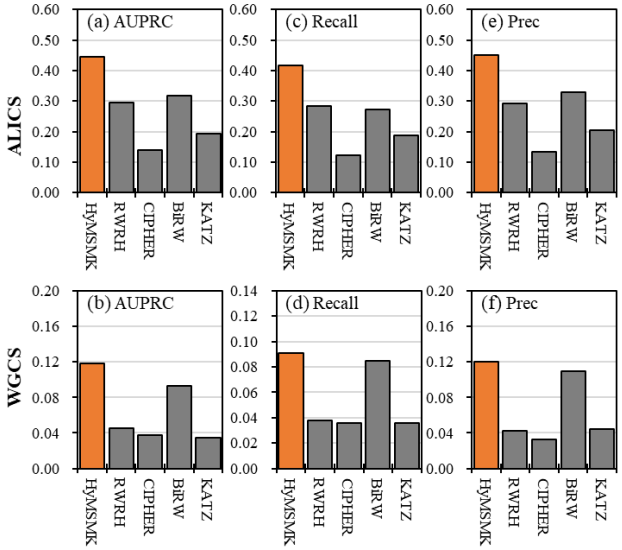

Fig. 4. In Dataset1 (MeSH), performance comparison of HyMSMK to different baseline algorithms: (a)-(b) AUPRC, (c)-(d) Recall, (e)-(f) Prec. The first and second lines are under the ALICS and WGCS control sets, respectively.

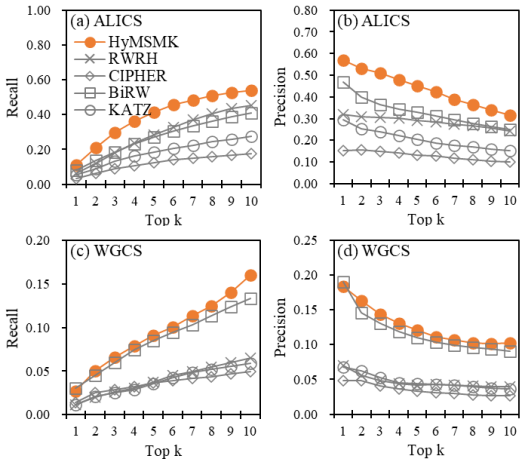

Fig. 5. In Dataset1 (MeSH), comparison of local performance of HyMSMK to baseline algorithms. (a)-(b) top-k Recall and Precision under the ALICS control set; (c)-(d) top-k Recall and Precision under the WGCS control set.

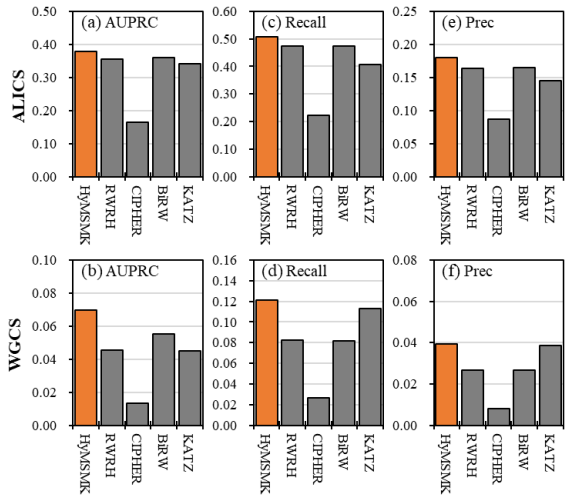

Fig. 6. In Dataset2(OMIM), performance comparison of HyMSMK to different baseline algorithms: (a)-(b) AUPRC, (c)-(d) Recall, (e)-(f) Prec. The first and second lines are under the ALICS and WGCS control sets, respectively.

3.4 Comparison between HyMSMK and baseline algorithms in cross-validation experiments

Here, we use the 5FCV experiments to evaluate the performance of HyMSMK (default), accompanied by comparison with classical network-based baseline algorithms: RWRH [22], CIPHER [73], BiRW [74] and KATZ [75].

(1) Cross-validation experiments in Dataset1. The results in Dataset1 (see **Fig. 4**) show that HyMSMK is superior to all the compared baseline algorithms under both control sets: ALICS and WGCS. Specifically, compared to the best results of the baseline algorithms, the AUPRC-, Recall- and Prec-values of HyMSMK under ALICS are improved by 41%, 47% and 37%, respectively; the AUPRC-, Recall- and Prec-values under WGCS are improved by 27%, 7% and 10%, respectively. The top-k Recall and Precision in **Fig. 5** again verify the good local performance of HyMSMK in Dataset1.

(2) Cross-validation experiments in Dataset2. In Dataset2, the results also show that HyMSMK outperforms all the baseline algorithms under both control sets (**Fig. 6**). Specifically, compared to the best results of baseline algorithms, the AUPRC-, Recall- and Prec-values of HyMSMK are improved by 5%, 7% and 9% under ALICS, and 27%, 47% and 48% under WGCS, respectively. **Fig. 7** also further verifies the good local performance of HyMSMK in Dataset2. So, HyMSMK is a robustly good choice under both control sets in Dataset2.

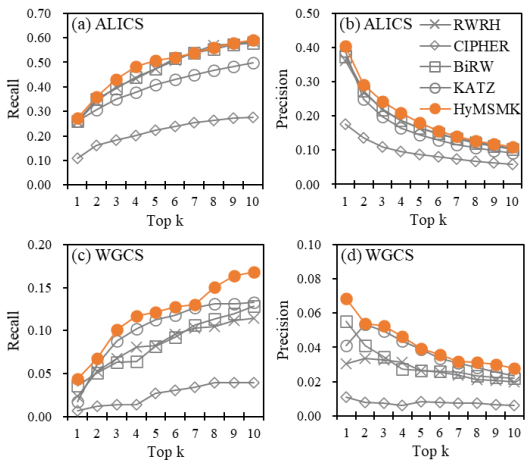

Fig. 7. In Dataset2 (OMIM), comparison of local performance of HyMSMK to baseline algorithms. (a)-(b) top-k Recall and Precision under the ALICS control set; (c)-(d) top-k Recall and Precision under the WGCS control set.

3.5 Performance evaluation of HyMSMK on external dataset

Here, we use Dataset1/Dataset2 as training set to compute the predictive scorces of candidates and use Dataset3 as test set to evaluate the predictive performance of algorithms (**Fig. 8** and **Fig. 9**). Note that there is no overlap between training set and test set.

When Dataset1 is used as test set (**Fig. 8**), HyMSMK under ALICS has higher Precision (e.g., from top-2 to 5) than the baselines; and it has higher Recall for top-3, 4, 5 and larger k-values. Under WGCS, the Precision of HyMSMK is larger than or comparable to that of the baselines in many cases (e.g., top-1 Precision); and its Recall is larger than others except BiRW.

When Dataset2 is used as test set (**Fig. 9**), under both

control sets, HyMSMK also has higher top-k Precision (e.g., from top-1 to top-5) than the baselines, while the Recall of HyMSMK is better than or comparable to the baselines in many cases. For example, under WGCS, its Recall from top-1 to top-5 is larger than all baselines.

These results show that HyMSMK also has good performance on the test of external dataset. Especially, HyMSMK has obviously higher Precision when Dataset2 is used as traning set.

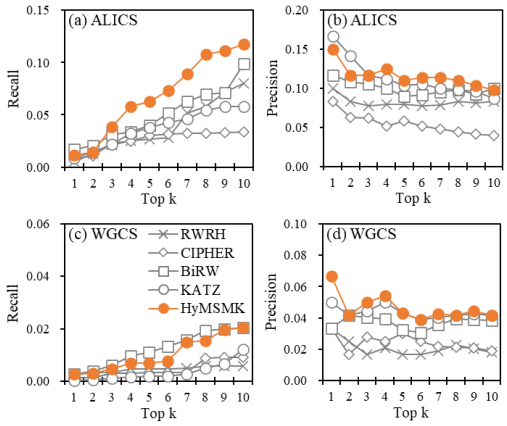

Fig. 8. In Dataset1 (MeSH), performance of HyMSMK on external dataset. (a)-(b) top-k Recall and Precision under the ALICS control set; (c)-(d) top-k Recall and Precision under the WGCS control set.

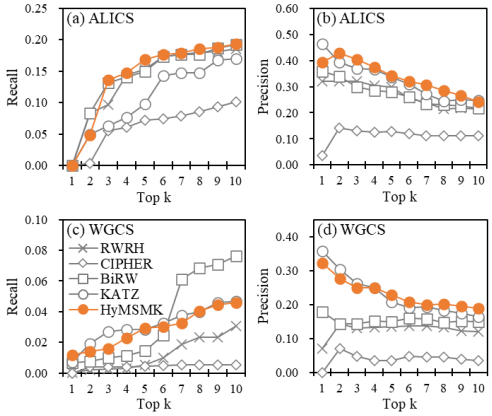

Fig. 9. In Dataset2 (OMIM), performance of HyMSMK on external dataset. (a)-(b) top-k Recall and Precision under the ALICS control set; (c)-(d) top-k Recall and Precision under the WGCS control set.

3.6 Case studies

Here, we conduct several case studies for specific diseases to further show the effectiveness of HyMSMK. We first generate several lists of candidate genes for specific diseases by using the OMIM database as a training set, and then show the disease relevance of the genes by literature validation and differential expression analysis (see **TABLE 2-3**). Note that "Y" denotes that there exist records in DisGeNet.

TABLE 2 lists the predicted top-k candidate genes for Alzheimer's disease (AD, #104300). In this table, we find 13 genes with AD-related literature records, 5 genes with AD-related records of the DisGeNet database (APOE, TF, PSEN1, MTHFR and MAOB), and 7 genes with AD-related records in literature while not in DisGeNet (EIF4G1, SNCA, SNCB, MAOA, MTHFD1, CP, AOC3). For example, Florentinus-Mefailoski showed that the increase of mean precursor strength of peptides from several plasma proteins (such as EIF4G1) was related to AD [76]. Bergström observed the increased levels of synuclein beta (SNCB) in cerebrospinal fluid (CSF) from AD samples compared with controls [77]. Quartey et al found that the increase of MAO-A activity occurred in AD cortical while not hippocampal samples, but the increase of MAO-B activity occurred in both regions [78].

Furthermore, we also study the top-k candidate genes for other diseases, showing their correlations with specific diseases by differential expression analysis of genes [79]. **TABLE 3** lists the top-20 candidate genes for breast cancer (BC, #114480), in which 12 genes are significantly differentially expressed in BC, and 8 genes have BC-related records in DisGeNet. This means that these candidate genes are very likely to be related to breast cancer. See supplementary materials for analysis of other diseases (e.g., lung cancer and prostate cancer).

TABLE 2
ANALYSIS OF CANDIDATE GENES FOR ALZHEIMER'S DISEASE.

Symbol	DisGeNet	Druggable (Science 2017)	DGIdb	Literature Validation
TFR2	-	-	0	-
APOE	Y	Tier 3A	20	Yes
TF	Y	Tier 3A	12	Yes
PSEN1	Y	Tier 1	10	Yes
EIF4G1	-	-	1	Yes
SNCA	-	Tier 2	33	Yes
GBA	-	Tier 1	32	-
SNCB	-	-	0	Yes
MAOA	-	Tier 1	47	Yes
MTHFR	Y	-	26	Yes
MTHFD1	-	-	3	Yes
SPACA3	-	Tier 3B	0	-
MTR	-	-	9	-
AOC1	-	Tier 2	3	-
CP	-	Tier 3A	20	Yes
MTRR	-	-	3	-
MAOB	Y	Tier 1	68	Yes
AOC3	-	Tier 1	6	Yes
IFIT3	-	-	0	-
AOC2	-	Tier 2	0	-

TABLE 3
ANALYSIS OF CANDIDATE GENES FOR BREAST CANCER.

Symbol	Differential Expression			DisGeNet	Druggable (Science 2017)	DGIdb
	Sig	log2FC	Trend			#Drug
BRCA1	Y	0.49	Up	Y	Tier 2	97
RAD51C	Y	0.25	Up	Y	-	0
NTHL1	Y	0.57	Up	-	-	0
SEC23B	Y	0.33	Up	-	-	0
MLH1	-	-	-	Y	-	0
SMAD7	-	-	-	-	-	2
MSH2	Y	0.23	Up	Y	-	1
RBBP8	-	-	-	-	-	0
XRCC2	Y	0.8	Up	Y	-	0
PMS2	Y	0.15	Up	Y	-	2
RAD51D	-	-	-	Y	-	0
CTNNB1	-	-	-	-	Tier 2	20
TOPBP1	-	0.19	Up	-	-	0
PATZ1	Y	0.29	Up	-	-	0
PRKN	-	-	-	-	-	0
TPX2	Y	1.64	Up	-	-	0
OPCML	-	0.41	Up	-	Tier 3A	1
MSH6	Y	0.27	Up	Y	-	2
NRAS	Y	0.21	Up	-	-	48
BLM	Y	0.98	Up	-	Tier 1	100

5 Conclusion

Disease-gene discovery is an important topic in biomedical data mining due to high cost and time consumption of biomedical experiment. The module nature of biomolecular networks may provide insights for disease research, and there exist different levels of valuable information hidden in MMS. However, it is challenging to make use of the information of MMS to more effectively mine disease-associated genes. Therefore, we propose the HyMSMK method by integrating multiscale module kernel (MSMK); MSMK is constructed by using multiscale module profile (MSMP) as a feature matrix and combining with the relative information content (IC) of features; and the MSMP containing the structural information from local to global scales is extracted from a comprehensive protein interactome by multiscale modularity optimization with exponential sampling.

By a series of experiments: cross-validation and independent test, we have systematically studied HyMSMK under both the ALICS and WGCS control sets, including the effects of different fusion strategies (CPF1/CPF2/CPF3, FS and MG) and kernel sparseness of MSMK on HyMSMK. The experimental results in different scenarios have showed the good performance of HyMSMK and confirmed the usefulness of MSMK for disease-gene discovery. Specifically, HyMSMK using CPF2 is recommended because it has consistently good performance in different scenarios (different datasets and control sets).

Furthermore, the case studies for several diseases (e.g., Alzheimer's disease and Breast cancer) verified the disease relevance of the candiadate genes obtained by HyMSMK for specific diseases, by several strategies such as literature validation and differential expression analysis, further showing the effectiveness of HyMSMK.

In summary, this work not only provides an effective framework of integrating MSMK for disease-gene discovery, but also may provide useful insights for the study and application of multiscale module structure in the mining of bio-entity associations.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (Grant No. 2019YFA0706202), and the Training Program for Excellent Young Innovators of Changsha (Grant No. kq2009093).

References

[1] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatraryamontri, K. Dolinski, and M. Tyers, "The BioGRID interaction database: 2019 update," *Nucleic Acids Research,* vol. 47, no. D1, pp. D529-D541, 2019.

[2] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L.I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research,* vol. 45, no. D1, pp. D833-D839, 2017.

[3] S. Köhler, M. Gargano, N. Matentzoglu, L.C. Carmody, D. Lewis-Smith, N.A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A.M. Brower, T.J. Callahan, C.G. Chute, J.L. Est, P.D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N.L. Harris, Michael J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J.A. McMurry, J.A. Miller, Monica C. Munoz-Torres, R.L. Peters, C.K. Rapp, A.M. Rath, S.A. Rind, Avi Z. Rosenberg, M.M. Segal, M.G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S.A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C.J. Mungall, M.A. Haendel, and P.N. Robinson, "The Human Phenotype Ontology in 2021," *Nucleic Acids Research,* vol. 49, no. D1, pp. D1207-D1217, 2020.

[4] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature Communications,* vol. 5, pp. 4212, 2014.

[5] L.Y.-H. Lee, and J. Loscalzo, "Network Medicine in Pathobiology," *The American Journal of Pathology,* vol. 189, no. 7, pp. 1311-1326, 2019.

[6] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nat Rev Genet,* vol. 12, no. 1, pp. 56-68, 2011.

[7] I.F. do Valle, H.G. Roweth, M.W. Malloy, S. Moco, D. Barron, E. Battinelli, J. Loscalzo, and A.-L. Barabási, "Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols," *Nature Food,* vol. 2, no. 3, pp. 143-155, 2021.

[8] Y. Moreau, and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nat Rev Genet,* vol. 13, no. 8, pp. 523-536, 2012.

[9] S.K. Ata, M. Wu, Y. Fang, L. Ou-Yang, C.K. Kwoh, and X.-L. Li, "Recent advances in network-based methods for disease gene prediction," *Briefings in Bioinformatics,* vol. 22, no. 4, pp. DOI:10.1093/bib/bbaa303, 2020.

[10] H. Zhang, A. Ferguson, G. Robertson, M. Jiang, T. Zhang, C. Sudlow, K. Smith, K. Rannikmae, and H. Wu, "Benchmarking network-based gene prioritization methods for cerebral small vessel disease," *Briefings in Bioinformatics*, 2021.

[11] P. Ruan, and S. Wang, "DiSNEP: a Disease-Specific gene Network Enhancement to improve Prioritizing candidate disease genes," *Briefings in Bioinformatics*, 2020.

[12] Y. Liu, Y. Guo, X. Liu, C. Wang, and M. Guo, "Pathogenic gene prediction based on network embedding," *Briefings in Bioinformatics*, pp. DOI:10.1093/bib/bbaa353, 2020.

[13] A. del Sol, R. Balling, L. Hood, and D. Galas, "Diseases as network perturbations," *Current Opinion in Biotechnology,* vol. 21, no. 4, pp. 566-571, 2010.

[14] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot, "Random walk with restart on multiplex and heterogeneous biological networks," *Bioinformatics,* vol. 35, no. 3, pp. 497–505, 2018.

[15] V.D. Tran, A. Sperduti, R. Backofen, and F. Costa, "Heterogeneous networks integration for disease–gene prioritization with node kernels," *Bioinformatics,* vol. 36, no. 9, pp. 2649-2656, 2020.

[16] K. Yang, K. Lu, Y. Wu, J. Yu, B. Liu, Y. Zhao, J. Chen, and X. Zhou, "A network-based machine-learning framework to identify both functional modules and disease genes," *Human Genetics,* vol. 140, no. 6, pp. 897-913, 2021.

[17] J. Xiang, J. Zhang, R. Zheng, X. Li, and M. Li, "NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction," *Briefings in Bioinformatics,* vol. 22, no. 5, pp.

bbab080, 2021.

[18] L.I. Furlong, "Human diseases through the lens of network biology," *Trends in Genetics,* vol. 29, no. 3, pp. 150-159, 2013.

[19] S. Köhler, S. Bauer, D. Horn, and P.N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," *The American Journal of Human Genetics,* vol. 82, no. 4, pp. 949-958, 2008.

[20] K. Hu, J.-B. Hu, L. Tang, J. Xiang, J.-L. Ma, Y.-Y. Gao, H.-J. Li, and Y. Zhang, "Predicting disease-related genes by path structure and community structure in protein–protein networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2018, no. 10, pp. 100001, 2018.

[21] J. Chen, E.E. Bardes, B.J. Aronow, and A.G. Jegga, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Research,* vol. 37, no. suppl_2, pp. W305-W311, 2009.

[22] Y. Li, and J.C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics,* vol. 26, no. 9, pp. 1219-1224, 2010.

[23] L. Cowen, T. Ideker, B.J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics,* vol. 18, pp. 551, 2017.

[24] R. Jiang, "Walking on multiple disease-gene networks to prioritize candidate genes," *Journal of Molecular Cell Biology,* vol. 7, no. 3, pp. 214-230, 2015.

[25] Y. Li, and J. Li, "Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data," *BMC Genomics,* vol. 13, no. 7, pp. S27, 2012.

[26] J. Xiang, N.-R. Zhang, J.-S. Zhang, X.-Y. Lv, and M. Li, "PrGeFNE: Predicting disease-related genes by fast network embedding," *Methods,* vol. 192, pp. 3-12, 2021.

[27] M. Oti, and H. Brunner, "The modular nature of genetic diseases," *Clinical Genetics,* vol. 71, no. 1, pp. 1-11, 2007.

[28] S. Choobdar, M.E. Ahsen, J. Crawford, M. Tomasoni, T. Fang, D. Lamparter, J. Lin, B. Hescott, X. Hu, J. Mercer, T. Natoli, R. Narayan, F. Aicheler, N. Amoroso, A. Arenas, K. Azhagesan, A. Baker, M. Banf, S. Batzoglou, A. Baudot, R. Bellotti, S. Bergmann, K.A. Boroevich, C. Brun, S. Cai, M. Caldera, A. Calderone, G. Cesareni, W. Chen, C. Chichester, S. Choobdar, L. Cowen, J. Crawford, H. Cui, P. Dao, M. De Domenico, A. Dhroso, G. Didier, M. Divine, A. del Sol, T. Fang, X. Feng, J.C. Flores-Canales, S. Fortunato, A. Gitter, A. Gorska, Y. Guan, A. Guénoche, S. Gómez, H. Hamza, A. Hartmann, S. He, A. Heijs, J. Heinrich, B. Hescott, X. Hu, Y. Hu, X. Huang, V.K. Hughitt, M. Jeon, L. Jeub, N.T. Johnson, K. Joo, I. Joung, S. Jung, S.G. Kalko, P.J. Kamola, J. Kang, B. Kaveelerdpotjana, M. Kim, Y.-A. Kim, O. Kohlbacher, D. Korkin, K. Krzysztof, K. Kunji, Z. Kutalik, K. Lage, D. Lamparter, S. Lang-Brown, T.D. Le, J. Lee, S. Lee, J. Lee, D. Li, J. Li, J. Lin, L. Liu, A. Loizou, Z. Luo, A. Lysenko, T. Ma, R. Mall, D. Marbach, T. Mattia, M. Medvedovic, J. Menche, J. Mercer, E. Micarelli, A. Monaco, F. Müller, R. Narayan, O. Narykov, T. Natoli, T. Norman, S. Park, L. Perfetto, D. Perrin, S. Pirrò, T.M. Przytycka, X. Qian, K. Raman, D. Ramazzotti, E. Ramsahai, B. Ravindran, P. Rennert, J. Saez-Rodriguez, C. Schärfe, R. Sharan, N. Shi, W. Shin, H. Shu, H. Sinha, D.K. Slonim, L. Spinelli, S. Srinivasan, A. Subramanian, C. Suver, D. Szklarczyk, S. Tangaro, S. Thiagarajan, L. Tichit, T. Tiede, B. Tripathi, A. Tsherniak, T. Tsunoda, D. Türei, E. Ullah, G. Vahedi, A. Valdeolivas, J. Vivek, C. von Mering, A. Waagmeester, B. Wang, Y. Wang, B.A. Weir, S. White, S. Winkler, K. Xu, T. Xu, C. Yan, L. Yang, K. Yu, X. Yu, G. Zaffaroni, M. Zaslavskiy, T. Zeng, J.D. Zhang, L. Zhang, W. Zhang, L. Zhang, X. Zhang, J. Zhang, X. Zhou, J. Zhou, H. Zhu, J. Zhu, G. Zuccon, A. Subramanian, J.D. Zhang, G. Stolovitzky, Z. Kutalik, K. Lage, D.K. Slonim, J. Saez-Rodriguez, L.J. Cowen, S. Bergmann, D. Marbach, and D.M.I.C.C. The, "Assessment of network module identification across complex diseases," *Nature Methods,* vol. 16, no. 9, pp. 843-852, 2019.

[29] X. Meng, W. Li, X. Peng, Y. Li, and M. Li, "Protein interaction networks: centrality, modularity, dynamics, and applications," *Frontiers of Computer Science,* vol. 15, no. 6, pp. 156902, 2021.

[30] S.S. Bhowmick, and B.S. Seah, "Clustering and Summarizing Protein-Protein Interaction Networks: A Survey," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 3, pp. 638-658, 2016.

[31] J. Ji, A. Zhang, C. Liu, X. Quan, and Z. Liu, "Survey: Functional Module Detection from Protein-Protein Interaction Networks," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 2, pp. 261-277, 2014.

[32] M. Kitsak, A. Sharma, J. Menche, E. Guney, S.D. Ghiassian, J. Loscalzo, and A.-L. Barabási, "Tissue Specificity of Human Disease Module," *Scientific Reports,* vol. 6, pp. 35241, 2016.

[33] Y. Liu, M. Brossard, D. Roqueiro, P. Margaritte-Jeannin, C. Sarnowski, E. Bouzigon, and F. Demenais, "SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network," *Bioinformatics,* vol. 33, no. 10, pp. 1536-1544, 2017.

[34] S.K. Dwivedi, A. Tjärnberg, J. Tegnér, and M. Gustafsson, "Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder," *Nature Communications,* vol. 11, no. 1, pp. 856, 2020.

[35] S.D. Ghiassian, J. Menche, and A.-L. Barabási, "A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome," *PLOS Computational Biology,* vol. 11, no. 4, pp. e1004120, 2015.

[36] S. Fortunato, and D. Hric, "Community detection in networks: A user guide," *Physics Reports,* vol. 659, pp. 1-44, 2016.

[37] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, P. Yu, and W. Zhang, "A Survey of Community Detection Approaches: From Statistical Modeling to Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-1, 2021.

[38] Z. He, H. Liang, Z. Chen, C. Zhao, and Y. Liu, "Detecting Statistically Significant Communities," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-1, 2020.

[39] W. Luo, D. Zhang, L. Ni, and N. Lu, "Multiscale Local Community Detection in Social Networks," *IEEE Transactions on Knowledge and Data Engineering,* vol. 33, no. 3, pp. 1102-1112, 2021.

[40] X. Liu, H.M. Cheng, and Z.Y. Zhang, "Evaluation of Community Detection Methods," *IEEE Transactions on Knowledge and Data Engineering,* vol. 32, no. 9, pp. 1736-1746, 2020.

[41] C. Tu, X. Zeng, H. Wang, Z. Zhang, Z. Liu, M. Sun, B. Zhang, and L. Lin, "A Unified Framework for Community Detection and Network Representation Learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 31, no. 6, pp. 1051-1065, 2019.

[42] Y. Dong, L.U.O. M, J. Li, D. Cai, and Q. Zheng, "LookCom: Learning Optimal Network for Community Detection," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-1, 2020.

[43] R. Sun, C. Chen, X. Wang, Y. Zhang, and X. Wang, "Stable Community Detection in Signed Social Networks," *IEEE*

Transactions on Knowledge and Data Engineering, pp. 1-1, 2020.

[44] Z. Wang, C. Wang, X. Li, C. Gao, X. Li, and J. Zhu, "Evolutionary Markov Dynamics for Network Community Detection," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-1, 2020.

[45] S. Fortunato, and M. Barthélemy, "Resolution limit in community detection," *Proc. Natl. Acad. Sci. USA,* vol. 104, no. 1, pp. 36-41, 2007.

[46] J. Xiang, Y.-N. Tang, Y.-Y. Gao, L. Liu, Y. Hao, J.-M. Li, Y. Zhang, and S. Chen, "Phase transition of Surprise optimization in community detection," *Physica A: Statistical Mechanics and its Applications,* vol. 491, no. C, pp. 693-707, 2018.

[47] J. Xiang, Z.-Z. Wang, H.-J. Li, Y. Zhang, F. Li, L.-P. Dong, J.-M. Li, and L.-J. Guo, "Community detection based on significance optimization in complex networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2017, no. 5, pp. 053213, 2017.

[48] A.-L. Barabasi, and Z.N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet,* vol. 5, no. 2, pp. 101-113, 2004.

[49] L.V. Schaffer, and T. Ideker, "Mapping the multiscale structure of biological systems," *Cell Systems,* vol. 12, no. 6, pp. 622-635, 2021.

[50] T.G.O. Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research,* vol. 47, no. D1, pp. D330-D338, 2018.

[51] M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp, "CORUM: the comprehensive resource of mammalian protein complexes—2019," *Nucleic Acids Research,* vol. 47, no. D1, pp. D559–D563, 2018.

[52] X. Meng, J. Xiang, R. Zheng, F.-X. Wu, and M. Li, "DPCMNE: detecting protein complexes from protein-protein interaction networks via multi-level network embedding," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. DOI:10.1109/TCBB.2021.3050102, 2021.

[53] A. Lewis, N. Jones, M. Porter, and C. Deane, "The function of communities in protein interaction networks at multiple scales," *BMC Systems Biology,* vol. 4, no. 1, pp. 100, 2010.

[54] J. Wang, M. Li, J. Chen, and Y. Pan, "A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 8, no. 3, pp. 607 - 620, 2011.

[55] E.M. Ramos, D. Hoffman, H.A. Junkins, D. Maglott, L. Phan, S.T. Sherry, M. Feolo, and L.A. Hindorff, "Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources," *European Journal of Human Genetics,* vol. 22, no. 1, pp. 144-147, 2014.

[56] J.S. Amberger, C.A. Bocchini, A.F. Scott, and A. Hamosh, "OMIM.org: leveraging knowledge across phenotype–gene relationships," *Nucleic Acids Research,* vol. 47, no. D1, pp. D1038-D1043, 2018.

[57] J. Menche, A. Sharma, M. Kitsak, S.D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science,* vol. 347, no. 6224, pp. 841, 2015.

[58] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A.P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research,* vol. 40, no. D1, pp. D857-D861, 2011.

[59] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N.H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R.C. Lovering, B. Meldal, A.N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob, "The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic acids research,* vol. 42, no. Database issue, pp. D358-63, 2014.

[60] T.S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D.S. Somanathan, A. Sebastian, S. Rani, S. Ray, C.J. Harrys Kishore, S. Kanth, M. Ahmed, M.K. Kashyap, R. Mohmood, Y.L. Ramachandra, V. Krishna, B.A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database—2009 update," *Nucleic Acids Research,* vol. 37, no. Database issue, pp. D767-D772, 2009.

[61] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," *Commun. ACM,* vol. 18, no. 11, pp. 613–620, 1975.

[62] S. Köhler, L. Carmody, N. Vasilevsky, J.O B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N.L. Harris, N. Matentzoglu, J.A. McMurry, D. Osumi-Sutherland, V. Cipriani, J.P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A.C. Jansen, A. Muaz, W.H. Chang, J. Bergerson, S.J F. Laulederkind, Z. Yüksel, S. Beltran, A.F. Freeman, P.I. Sergouniotis, D. Durkin, A.L. Storm, M. Hanauer, M. Brudno, S.M. Bello, M. Sincan, K. Rageth, M.T. Wheeler, R. Oegema, H. Lourghi, M.G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R.C. Lovering, C. Hajek, A. Olry, L. Notarangelo, M. Similuk, X.A. Zhang, D. Gómez-Andrés, H. Lochmüller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J.D. Milner, D. Leroux, C.F. Boerkoel, A. Klion, M.C. Carter, T. Groza, D. Smedley, M.A. Haendel, C. Mungall, and P.N. Robinson, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," *Nucleic Acids Research,* vol. 47, no. D1, pp. D1018-D1027, 2018.

[63] S.K. Westbury, E. Turro, D. Greene, C. Lentaigne, A.M. Kelly, T.K. Bariana, I. Simeoni, X. Pillois, A. Attwood, S. Austin, S.B. Jansen, T. Bakchoul, A. Crisp-Hihn, W.N. Erber, R. Favier, N. Foad, M. Gattens, J.D. Jolley, R. Liesner, S. Meacham, C.M. Millar, A.T. Nurden, K. Peerlinck, D.J. Perry, P. Poudel, S. Schulman, H. Schulze, J.C. Stephens, B. Furie, P.N. Robinson, C. van Geet, A. Rendon, K. Gomez, M.A. Laffan, M.P. Lambert, P. Nurden, W.H. Ouwehand, S. Richardson, A.D. Mumford, K. Freson, and o.b.o.t.B.-B. Consortium, "Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders," *Genome Medicine,* vol. 7, no. 1, pp. 36, 2015.

[64] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of artificial intelligence research* vol. 11 pp. 95-130, 1999.

[65] A. Singhal, S. Cao, C. Churas, D. Pratt, S. Fortunato, F. Zheng, and T. Ideker, "Multiscale community detection in Cytoscape," *PLOS Computational Biology,* vol. 16, no. 10, pp. e1008239, 2020.

[66] M.E. Newman, and M. Girvan, “Finding and evaluating community structure in networks,” *Phys Rev E* vol. 69, 2004.

[67] J. Xiang, Y.-N. Tang, Y.-Y. Gao, Y. Zhang, K. Deng, X.-K. Xu, and K. Hu, “Multi-resolution community detection based on generalized self-loop rescaling strategy,” *Physica A: Statistical Mechanics and its Applications,* vol. 432, pp. 127-139, 2015.

[68] S. Rahiminejad, M.R. Maurya, and S. Subramaniam, “Topological and functional comparison of community detection algorithms in biological networks,” *BMC Bioinformatics,* vol. 20, no. 1, pp. 212, 2019.

[69] J. Xiang, Y. Zhang, J.-M. Li, H.-J. Li, and M. Li, “Identifying multi-scale communities in networks by asymptotic surprise,” *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2019, no. 3, pp. 033403, 2019.

[70] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, no. 10, pp. P10008, 2008.

[71] J. Xiang, X.-G. Hu, X.-Y. Zhang, J.-F. Fan, X.-L. Zeng, G.-Y. Fu, K. Deng, and K. Hu, “Multi-resolution modularity methods and their limitations in community detection,” *European Physical Journal B,* vol. 85, no. 10, pp. 1-10, 2012.

[72] X. Li, X. Wang, and G. Xiao, “A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications,” *Briefings in Bioinformatics,* vol. 20, no. 1, pp. 178-189, 2017.

[73] X. Wu, R. Jiang, M.Q. Zhang, and S. Li, “Network-based global inference of human disease genes,” *Molecular Systems Biology,* vol. 4, pp. 189-189, 2008.

[74] M. Xie, Y. Xu, Y. Zhang, T. Hwang, and R. Kuang, “Network-based Phenome-Genome Association Prediction by Bi-Random Walk,” *Plos One,* vol. 10, no. 5, pp. e0125138, 2015.

[75] U.M. Singh-Blom, N. Natarajan, A. Tewari, J.O. Woods, I.S. Dhillon, and E.M. Marcotte, “Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses,” *Plos One,* vol. 8, no. 5, pp. e58977, 2013.

[76] A. Florentinus-Mefailoski, P. Bowden, P. Scheltens, J. Killestein, C. Teunissen, and J.G. Marshall, “The plasma peptides of Alzheimer’s disease,” *Clinical Proteomics,* vol. 18, no. 1, pp. 17, 2021.

[77] S. Bergström, J. Remnestål, J. Yousef, J. Olofsson, I. Markaki, S. Carvalho, J.C. Corvol, K. Kultima, L. Kilander, M. Löwenmark, M. Ingelsson, K. Blennow, H. Zetterberg, B. Nellgård, F. Brosseron, M.T. Heneka, B. Bosch, R. Sanchez-Valle, A. Månberg, P. Svenningsson, and P. Nilsson, “Multi-cohort profiling reveals elevated CSF levels of brain-enriched proteins in Alzheimer's disease,” *Ann Clin Transl Neurol*, 2021.

[78] M.O. Quartey, J.N.K. Nyarko, P.R. Pennington, R.M. Heistad, P.C. Klassen, G.B. Baker, and D.D. Mousseau, “Alzheimer Disease and Selected Risk Factors Disrupt a Co-regulation of Monoamine Oxidase-A/B in the Hippocampus, but Not in the Cortex,” *Front Neurosci,* vol. 12, pp. 419, 2018.

[79] H.M. Dingerdissen, J. Torcivia-Rodriguez, Y. Hu, T.-C. Chang, R. Mazumder, and R. Kahsay, “BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery,” *Nucleic Acids Research,* vol. 46, no. D1, pp. D1128-D1136, 2017.