OXFORD

Briefings in Bioinformatics, 2022, **23(3)**, 1–22
https://doi.org/10.1093/bib/bbac072
Problem Solving Protocol

HyMM: hybrid method for disease-gene prediction by integrating multiscale module structure

Ju Xiang, Xiangmao Meng, Yichao Zhao, Fang-Xiang Wu and Min Li

Corresponding author. Min Li, Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China. Tel.: +86-731-88830212; Fax: +86-731-88830212; E-mail: limin@mail.csu.edu.cn

Abstract

Identifying disease-related genes is an important issue in computational biology. Module structure widely exists in biomolecule networks, and complex diseases are usually thought to be caused by perturbations of local neighborhoods in the networks, which can provide useful insights for the study of disease-related genes. However, the mining and effective utilization of the module structure is still challenging in such issues as a disease gene prediction.

We propose a hybrid disease-gene prediction method integrating multiscale module structure (HyMM), which can utilize multiscale information from local to global structure to more effectively predict disease-related genes. HyMM extracts module partitions from local to global scales by multiscale modularity optimization with exponential sampling, and estimates the disease relatedness of genes in partitions by the abundance of disease-related genes within modules. Then, a probabilistic model for integration of gene rankings is designed in order to integrate multiple predictions derived from multiscale module partitions and network propagation, and a parameter estimation strategy based on functional information is proposed to further enhance HyMM's predictive power. By a series of experiments, we reveal the importance of module partitions at different scales, and verify the stable and good performance of HyMM compared with eight other state-of-the-arts and its further performance improvement derived from the parameter estimation. The results confirm that HyMM is an effective framework for integrating multiscale module structure to enhance the ability to predict disease-related genes, which may provide useful insights for the study of the multiscale module structure and its application in such issues as a disease-gene prediction.

Keywords: disease-gene prediction, association prediction, complex networks, biological networks, multiscale module structure, ranking methods

Introduction

The progress of human disease gene discovery has promoted the understanding of the underlying molecular basis of human diseases, but genes known to be associated with diseases only account for a very small proportion of the incidences [1–4]. Traditional approaches such as linkage analysis and genome-wide association studies (GWAS) often provide a long list of candidate genes, requiring expensive and time-consuming experimental identification [5, 6]. Therefore, with the accumulation of biomedical data [7–10], developing computational algorithms for predicting disease-related candidate genes is indispensable to accelerate the discovery of disease-related genes [3, 11, 12].

Organism as a complex biological system is composed of a large number of biomolecules (e.g. genes and proteins) with complex relationships (physical interactions or functional associations), forming a complex biomolecule network system, where the biomolecules exert biological functions through intermolecular synergy while rarely function alone. Human complex diseases can be viewed as the consequences of perturbations or functional abnormalities of associated synergistic biomolecules in the complex network system [13]. Therefore, it is very necessary to study complex diseases and relevant biological phenomena from the perspective of system biology, and biological networks provide an important means for the research of system biology [14–18]. Especially, network-based algorithms have been a popular strategy for the study of disease-related genes [3, 19–28], since genes associated with the same or similar diseases are more similar functionally

Ju Xiang is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. He is an Associate Professor with Changsha Medical University, Hunan, China. His research interests include complex networks, bioinformatics, machine learning and deep learning.
Xiangmao Meng is currently postdoctoral in the School of Computer Science and Engineering, Central South University, China. His current research interests include bioinformatics, complex network analysis and data mining.
Yichao Zhao is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. His current research interests include bioinformatics and system biology.
Fang-Xiang Wu is a Professor in the Division of Biomedical Engineering, Department of Computer Science, Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada. He is a senior member of IEEE. His current research interests include bioinformatics and artificial intelligence.
Min Li is currently the vice dean and a Professor at the School of Computer Science and Engineering, Central South University, China. Her main research interests include bioinformatics and system biology.

Received: October 20, 2021. **Revised:** January 18, 2022. **Accepted:** February 13, 2022

and their products tend to be highly interconnected in biomolecule networks [4, 29] (see next section). However, how to mine the characteristics of biomolecule networks so as to more effectively explore disease-related genes and related issues is still under continuous exploration due to the inherent complexity of biomolecule networks and the limitation of existing knowledge (e.g. the incompleteness of protein interactome) [15, 17, 30–34].

As we know, module structure as a common property of complex networks is ubiquitous in biomolecule networks, and the modular nature of human diseases can provide useful insights for the study of diseases, but it has not been fully explored in disease-gene prediction [35–37]. Generally, the genes and their products of disease tend to form a disease module due to their high interconnectivity in biomolecule networks [4, 29], but they are usually found to be distributed in multiple modules/subnetworks due to the intrinsic definition of a specific algorithm and the existence of multiscale module structure in the networks [4, 38–41]. The multiscale structure is indeed widespread in biological networks. For example, a module in a protein network may contain several submodules, e.g. some protein complexes (such as SAGA) contain several secondary complexes; most of the biological information (e.g. in Gene Ontology) is organized in the form of hierarchical structure. Many algorithms with a flexible resolution parameter have been proposed and applied to mine multiscale module structures in biological networks [42–45], where the resolution parameter can adjust the size or scale of identified modules (see next section). This can provide richer information for studying complex systems such as biomolecule networks, but there are still many challenging issues, such as how to effectively identify the multiscale modules from a network and how to mine the valuable information hidden in the multiscale structure.

To make use of multiscale module structure to more effectively predict disease-related genes, we therefore propose a hybrid method integrating the information of multiscale modules (HyMM) (Figure 1). HyMM extracts a series of module partitions from local to global scales by multiscale modularity optimization (MO) with exponential sampling, and estimates the disease relatedness of genes in partitions by the abundance of disease-related genes in modules. Then, a probabilistic model for integration of gene rankings is designed so as to integrate multiple predictions derived from multiscale module partitions and network propagation, and a parameter estimation strategy based on functional information for Gene Ontology (GO) annotations, pathways or disease genes is proposed to further enhance HyMM's predictive power.

The rest of the paper is organized as follows. Firstly, we introduce some related work in this study (including the identification of module structure and the prediction of disease-related genes). Secondly, we present the datasets and details of HyMM, as well as evaluation methods. Thirdly, we study the effectiveness of multiscale module information in disease-gene prediction by combining the functional analysis of multiscale modules; then, by a series of experimental tests, we verify the good performance of HyMM and study the effects of various factors including the definition of conditional probability, multiscale module extraction, parameter estimation based on functional information, sampling of multiscale module partitions and random shuffling of disease-gene associations. Furthermore, we apply HyMM to other datasets as well as specific diseases [e.g. Alzheimer's disease (AD)] to further demonstrate the effectiveness of HyMM. These results confirm that HyMM can enhance the ability to predict disease-related genes by integrating a multiscale module structure. It is a protocol for disease-gene prediction integrating multiscale module structure, which may become a very useful computational tool for the study of disease-related genes.

Related work

In this study, we focus on the mining of information hidden in the multiscale structure to enhance the ability of disease-gene prediction, which involves two main aspects: disease-gene prediction and (multiscale) module identification (also called community detection or community mining in the field of complex networks [46–48]).

Disease-gene prediction is not only an important issue in computational biology but also is an important field of network medicine/biology [4, 14, 16, 17, 19, 43–45]. Numerous network-based methods for disease-gene prediction have been proposed, based on various approaches, e.g. from homogeneous (HO) network to heterogeneous (HE) network and from single-layer network to multi-layer network [15, 17, 34, 49–52]. For HO network model, for example, Köhler *et al.* [53] predicted disease-associated genes by the random walk with restart (RWR) on a protein interactome; Chen *et al.* [54] prioritized disease candidate genes by the k-step markov (KS) method; Hsu *et al.* [55] developed the gene interconnectedness-based method to rank candidate genes by evaluating the network closeness of them to seeds (known disease-related genes); Zhu *et al.* [56] proposed the vertex similarity-based (VS) method to discover disease-associated genes. For HE network model, for example, Li *et al.* [57] proposed the RWR on a disease-gene HE network to infer disease-gene associations; Wu *et al.* [58] proposed the network-based global inference method called CIPHER to predict human disease-related genes; Xie *et al.* [59] proposed the bi-random walk (BiRW) to predict disease-gene associations; Singh-Blom *et al.* [60] predicted disease-gene associations by developing the KATZ measure on a HE network, inspired by social network analyses. For more sophistic network models or techniques, for example, Valdeolivas et *al.* proposed the RWR on multiplex and HE networks [27]; Xiang *et al.* [34] proposed the network impulsive dynamics on the multiplex network for disease-gene prediction; Liu *et al.* [33] proposed a new network embedded representation algorithm to infer pathogenic genes; Xiang *et al.* [61]

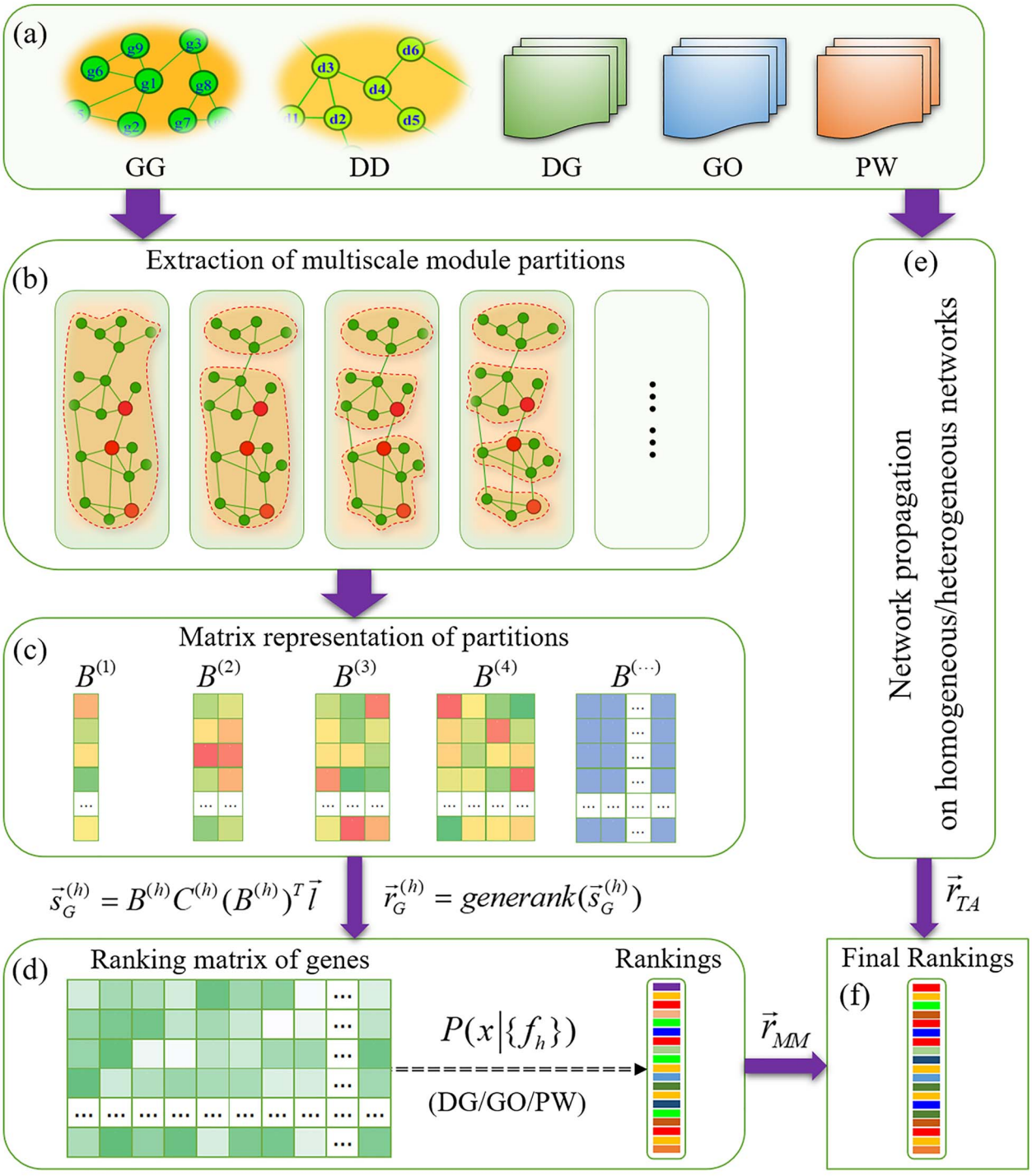

Figure 1. Workflow of the HyMM method integrating multiscale module structure. (**a**) Datasets: GG, DD and DG denote the gene–gene, disease–disease and disease–gene associations, respectively; GO and PW denote the GO annotations and pathways of genes, respectively. (**b**) Extract multiscale module partitions by multiscale algorithm. (**c**) The extracted module partitions are transformed into matrix representation. (**d**) A ranking list of genes is generated for each partition matrix; these ranking lists are organized into a ranking matrix of genes, and then, an integrated ranking list of genes based on this ranking matrix is generated by a probabilistic model along with a parameter estimation based on functional information. (**e**) Generate a ranking list of genes by network propagation. (**f**) Final rankings of genes are generated by integrating the ranking lists of genes from multiscale modules and network propagation.

proposed a disease-gene-prediction method based on fast network embedding, which can effectively use information in a multi-source HE network constructed by integrating multiple types of association data. Moreover, some module-based algorithms are also applied to the analysis of disease-related genes/modules as well as related issues [42, 62–65]. See references [2, 3, 15, 19, 66–68] for related reviews.

The existing methods in literature have promoted the progress of disease-gene prediction, while the 'guilt-by-

association' becomes a top-down central principle for predicting disease-related genes in the networks [69]. Evaluating the closeness or distance between candidates and known disease-related genes is a direct strategy to infer disease-related genes in the networks [20, 55, 56], while network propagation (e.g. RWR, KS, RWRH and BiRW) can effectively make use of more information in the whole network to mine potential disease-gene associations [26, 53, 54, 70, 71]. Network propagation (especially the RWR) shows excellent performance in many scenarios [19], so it has been widely applied in the study of bio-entity associations including disease-gene prediction [70].

Biological networks such as protein–protein interaction networks are an important basis for network-based methods in disease-gene prediction, and the mining of biological networks can be helpful for understanding the characteristics of networks and promoting the study of relevant issues. The existence of module structure is an important property of biological networks [72–74], and the research of module structure has been an important topic in the study of complex networks including biological networks. In the past decades, a large number of different types of algorithms based on various approaches (e.g. MO [75], dynamics [76] and statistical inference [77]) were proposed to identify modules (or say communities) in networks, which involve various types of modules (e.g. from single scale to multiple scales, and from non-overlapping to overlapping) and various types of networks (e.g. from unweighted to weighted networks, from undirected to directed networks, and from unsigned to signed networks) [46, 47]. Many of the module identification algorithms, especially MO-based algorithms, have been applied to the study of biological networks, e.g. functional module mining [72, 74, 78], protein complex detection [79–81], and disease module identification [42].

As mentioned above, genes/proteins associated with the same disease tend to form relevant disease modules in a biomolecule network [4, 29], but these genes are usually distributed in multiple modules by specific algorithms [4, 38]. There are several possible reasons for this phenomenon. (i) Complex diseases usually involve functional abnormalities of multiple genes, and these genes may have different functions, playing different roles in the development of complex diseases [82]. (ii) The existing biomolecule networks such as protein–protein interactions are still incomplete [30, 83]. This may cause the detected network modules to be broken and incomplete. (iii) Detected modules in networks are often algorithm-specific, because specific definitions of modules are different for different algorithms [46]. Some algorithms may split a large module into several small submodules in a network, or aggregate several small modules into a large one, because of the existence of a resolution limit that is related to the intrinsic definition or mechanism of algorithms [39–41]. This also implies the existence of a multiscale structure in the network.

In fact, multiscale structure widely exists in various natural and artificial complex networks, including biological networks [84, 85]. In this case, algorithms with flexible resolution parameters, e.g. multiscale MO [86, 87], may more effectively mine the module structure of networks at different scales, where the resolution parameter can be used to tune the scale or size of identified modules [48, 86, 88]. For example, multiscale MO can find relatively large modules in a network when the resolution parameter is small, while it can identify relatively small modules when the resolution parameter is large. This is similar to observing an object from a local to a global scale by a microscope with adjustable resolution parameters. Modules at different scales from local to global ones can be identified by adjusting the resolution parameter in continuous real number space. To study modules at different scales, one generally extracts a set of module partitions corresponding to a set of values sampled from the space of the resolution parameter by a suitable strategy (e.g. exponential sampling).

Multiscale module identification is important for studying biomolecule networks. Dunn *et al.* [89] have used edge-betweenness clustering to separate protein interaction networks into modules correlating to annotated gene functions, where modules of different sizes can be identified by removing different numbers of edges. Lewis *et al.* [90] investigated the correlation between the functions of sets of proteins and network module/-community structure at multiple resolutions/scales, and they showed that there exist different important scales of module/community structure depending on studied proteins and processes. Wang *et al.* [91] proposed a fast hierarchical clustering (HC) algorithm using the local metric of edge clustering value, which can uncover the hierarchical organization of functional modules that approximately corresponds to the hierarchical structure of GO annotations. Extended (multiscale) MO was used to identify disease modules [42]. More recently, Zheng *et al.* developed a multiscale approach called HiDeF to identify robust structures at all scales by integrating the concept of persistent homology with existing community detection algorithms (e.g. multiscale MO).

The mining of multiscale module structure can reveal the features of biological networks at multiple scales, reflecting the correlations of network nodes (e.g. genes) at different levels. This can provide more abundant information for the relevant research involving biological networks, such as network-based disease–gene prediction and protein–protein interaction prediction. Therefore, in this study, we will explore methods for disease-gene prediction by integrating a multiscale module structure.

Materials and methods

Here, we introduce the datasets in this study, and then propose the details of HyMM (including multiscale MO with exponential sampling, disease-relatedness estimation of genes based on multiscale modules, a probabilistic model for integration of multiple gene rankings as well as parameter estimation based on

functional information) and evaluation methods. See Figure 1 and Supplementary Note 1 (see Supplementary Data available online at https://academic.oup.com/bib) for the workflow of HyMM.

Datasets

To investigate the predictive ability of algorithms, we employ the disease-gene associations, gene–gene associations and disease-disease associations. In order to conduct functional analysis of modules and parameter estimation based on functional information, we adopt three types of functional groups: GO annotations, PW, and disease-gene sets. See Supplementary Note 2 (see Supplementary Data available online at https://academic.oup.com/bib) for details of datasets.

Disease–gene associations

We use three disease-gene association datasets. (i) The first dataset is an integrated disease-gene dataset [30, 92] retrieved from GWAS and Online Mendelian Inheritance in Man [93]. It is denoted as the Medical Subject Headings Ontology (MeSH) dataset, since MeSH is used to combine the different disease nomenclatures of the two sources into a single standard vocabulary. (ii) The second one is obtained from the DISEASES database, which is a weekly updated web resource for disease-gene associations [94]. (iii) The third one is obtained from the DisGeNet database (https://www.disgenet.org/), which is known as a platform that contains one of the largest publicly available collections of disease-related genes [95]. The UMLS (Unified Medical Language System) diseases in the dataset are mapped into MeSH diseases.

Gene–gene associations

Genes and their products mainly perform their biological functions through their direct or indirect interactions, forming a complex gene–gene association network [83, 96–99]. The gene–gene associations are very important for the study of disease research, since complex diseases are usually considered to be caused by local disturbances of complex biomolecule networks [17, 100, 101]. Here, the gene–gene associations are derived from protein–protein interactions (PPIs). Because single-source protein–protein networks are often incomplete and there exist data noises in existing protein networks, we adopt a comprehensive protein interactome that consists of multiple sources of protein–protein interactions: regulatory interactions, binary interactions from several yeast two-hybrid high-throughput and literature-curated datasets, literature-curated interactions derived mostly from low-throughput experiments, metabolic enzyme-coupled interactions, protein complexes, kinase-substrate pairs and signaling interactions [30]. The network data considers only physical protein interactions with experimental support. The identifiers of proteins are mapped into gene symbols.

These gene–gene associations form a HO network of genes. Furthermore, we construct a disease-gene HE network by integrating gene–gene associations, disease-gene associations and disease-disease associations mentioned above. Note that only the disease-gene associations in the training set are used in the construction of the HE network.

Disease–disease associations

The disease-disease associations can provide useful knowledge for the discovery of disease-related genes. Here, the disease–disease association network is constructed by using the associations between symptoms and diseases. The strengths of these associations between a symptom s and a disease d are quantified through the co-occurrence ($C_{d,s}$) of their MeSH terms in literature, i.e. the number of PubMed identifiers where two MeSH terms occur together, and then they are normalized as $w_{d,s} = C_{d,s}\log(n/n_s)$ by the term frequency-inverse document frequency, where n denotes the number of diseases and $n_s \geq 1$ denotes the number of diseases with symptoms [102]. Finally, the association score between two diseases is quantified by the cosine similarity scores of their normalized symptom vectors, $\mathbf{Score}(V_d, V_b) = \langle V_d, V_b\rangle / \sqrt{\langle V_d, V_d\rangle \langle V_b, V_b\rangle}$, where $V_d = (w_{d,1}, w_{d,2}, \ldots, w_{d,s}, \ldots, w_{d,n})^T$ denotes the normalized symptom vector of disease d and $\langle \cdot, \cdot \rangle$ denotes the scalar product of two vectors.

Three types of functional groups

(i) The GO annotations are downloaded from the Molecular Signatures Databases (MSigDB) [103, 104], which omits GO terms with fewer than 5 genes or in very broad categories; (ii) the pathway-gene sets (PW) are also obtained from MSigDB, which were curated from several online pathway databases (such as KEGG and Reactome), publications in PubMed and knowledge of domain experts [105, 106] and (iii) the disease-gene sets (DG) are obtained as mentioned above.

Multiscale MO with exponential sampling

Identification of module/community structure itself is an important issue in the research of networked systems [46, 48, 107]. We here extract module structure from local to global scales by MO with exponential sampling. Moreover, we also consider two other multiscale methods: asymptotic surprise (AS) [88] and fast HC [91]. All of them have flexible resolution parameters to adjust the scale or size of modules. Given a set of reasonably sampled resolution-parameter values, they can generate a set of network module partitions that contain important information of network structure. (see Supplementary Note 3, see Supplementary Data available online at https://academic.oup.com/bib, for details).

Multiscale MO

MO detects module structure via optimizing modularity Q—a widely used quality function of module structure [108]. Original modularity can just generate a single-scale module structure due to its fixed resolution [39]. Therefore, its multiscale variants have been widely studied. For

example, the self-loop rescaling strategy can naturally transform the original single-scale modularity into a multiscale one, and the original optimization algorithms can be applied directly without the need for any other modification [87]. Given a module partition of a network, the general definition of modularity Q can be written as

$$Q = \sum_s \left(\frac{k_s^{in}}{2M} - \gamma \left(\frac{k_s}{2M} \right)^2 \right), \quad (1)$$

where γ is the resolution parameter; M is the total number of edges in the network; k_s^{in} the inner degree of module s; k_s the total degree of module s; $\sum$ the sum over all modules in the network.

MO can detect module structure at different scales by varying the resolution parameter γ. It can find global-scale modules (i.e. relatively large modules) when the γ-value is small; it can find local-scale modules (i.e. relatively small modules) when the γ-value is large. It will decompose a network into a set of single-node modules when γ is large enough, e.g. $\gamma > 2M/k_{min}^2$, where k_{min} is the minimum node degree in the network [109]. This is similar to what happens when we observe objects by a microscope with adjustable resolution. When the resolution of the microscope is high, we can see very local and subtle areas of the object in the field of vision of the microscope; when the resolution is low, we can see its relatively macro and coarse-grained areas, and even its global appearance in the field of vision of the microscope. There is no strict threshold to distinguish between local and global modules, but the limit of 'local scale' is that the network is split into a set of single-node modules, and the limit of 'global scale' is that the whole network is considered as a large module.

MO needs to be realized with the help of effective optimization algorithms. Here, the Louvain algorithm is applied, because it is a very effective and widely used strategy for optimizing objective functions of module structure in networks, and we have shown that it can be further improved by an effective initialization process and refining process [48, 75, 88].

Exponential sampling of multiscale module partitions

To extract a set of meaningful module partitions from local to global scales $\{\Psi_h \mid h = 1 \sim H\}$, where Ψ_h denotes the h-th module partition and H denotes the number of extracted module partitions, we first define a meaningful range of the resolution parameter $\gamma \in [\gamma_{min}, \gamma_{max}]$, which covers all possible sampled resolution-parameter values. In a network with N nodes, γ_{min} and γ_{max} can theoretically be defined as $\gamma_{min} = \mathbf{max}\{\gamma | \#[\Psi_h] = 1\}$ and $\gamma_{max} = \mathbf{min}\{\gamma | \#[\Psi_h] = N]\}$, where $\#[\Psi_h]$ denotes the number of modules in the partition Ψ_h, but it is usually not easy to obtain accurate interval boundaries. Therefore, we use semi-empirical boundaries according to previous research [87, 110].

The resolution parameter belongs to a continuous real number space. We sample γ-values from γ_{min} to γ_{max} by exponential sampling method, because it can give a reasonable coverage to different scales in the network, according to previous research [87, 88]. The exponential sampling generates a set of γ -values that are equally spaced on a logarithmic scale, i.e.

$$\log \gamma_{h+1} - \log \gamma_h \equiv \Delta \log \gamma. \quad (2)$$

According to the set of sampled γ-values, the above multiscale algorithms can extract a set of corresponding module partitions.

Disease-relatedness estimation of genes based on multiscale module structure

Given the set of extracted module partitions $\{\Psi_h \mid h = 1 \sim H\}$, known disease-gene associations as well as disease-disease associations, we calculate the disease-relatedness scores of modules and genes in each module partition by the abundance of disease-related genes within modules, and then generate a disease-relatedness scoring/ranking matrix of genes (see Figure 1). The basic hypothesis of estimating the disease relatedness of modules/genes in a module is that the larger the abundance of specific disease-related genes in the module, the more likely the module and its genes are related to the disease.

Definition 1. A vector indicating association scores between N genes and a disease under study is defined as

$$\vec{l} = (l_1, l_2, \cdots, l_i, \cdots l_N)^T \in R^{N \times 1}, \quad (3)$$

where l_i=1 if the ith node is a known disease-related gene and l_i=0 otherwise, e.g. in a gene–gene network.

Definition 2. A partition matrix $B^{(h)}$ is defined to indicate the hth module partition, where $B_{i,j}^{(h)}$ indicates whether gene i belongs to module j in this partition (see Figure 1).

Definition 3. A diagonal matrix $C^{(h)}$ for each partition is defined as

$$C^{(h)} = \left(\mathbf{diag} \left(\vec{e}^{\,T} B^{(h)} \right) \right)^{-1}, \quad (4)$$

where $\vec{e} = (1, 1, \cdots, 1)^T \in R^{N \times 1}$.

Definition 4. The disease relatedness scorings of modules in the hth module partition are defined as

$$\vec{s}_M^{(h)} = \left(\vec{l}^{\,T} B^{(h)} C^{(h)} \right)^T. \quad (5)$$

Definition 5. The disease-relatedness scores of genes in the hth module partition are defined as

$$\vec{s}_G^{(h)} = B^{(h)} \vec{s}_M^{(h)}. \quad (6)$$

We introduce a union matrix of gene scorings, $S_G = (\vec{s}_G^{(1)}, \vec{s}_G^{(2)}, \cdots, \vec{s}_G^{(h)}, \cdots, \vec{s}_G^{(H)})$, to store the gene scoring lists for all module partitions. For the sake of computation, we construct a union matrix B of all partition matrices by $B = (B^{(1)}, B^{(2)}, \ldots, B^{(H)})$, and define a block matrix S_M of disease-related scores of modules by

$$S_M = \begin{pmatrix} \vec{s}_M^{(1)} & 0 & \cdots & 0 \\ 0 & \vec{s}_M^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \vec{s}_M^{(H)} \end{pmatrix}. \quad (7)$$

Then, the union matrix of gene scorings can be calculated by $S_G = BS_M$. Finally, we generate a union matrix of gene rankings, $R_G = (\vec{r}_G^{(1)}, \vec{r}_G^{(2)}, \cdots, \vec{r}_G^{(h)}, \cdots, \vec{r}_G^{(H)})$, by the decreasing order of genes' scores, $R_G = generank(S_G)$. Note that the mean value of ranking is used for genes with the same scores.

According to the above gene scoring/ranking strategy, genes within the same module have the same disease-relatedness scorings/rankings. Therefore, these scoring/ranking lists of genes contain the information of disease relatedness as well as the information of multiscale module structure from low to high resolutions. For example, if a module partition consists of two modules: one contains disease-related genes while not for another, gene scorings/rankings will have two values/levels. If the module with disease-related genes is further split into two sub-modules with disease-related genes, then gene scorings/rankings will have three values/levels if there is no degeneracy of module scorings. The number of values/levels in the scoring/ranking list of genes is closely related to the number of disease-related modules in a module partition. As the resolution increases, we can get the gene scoring/ranking lists with more levels/values, thereby revealing different levels of disease-related information in the network.

From the perspective of kernel function, a kernel matrix for each module partition Ψ_h can be defined as

$$K^{(h)} = \tilde{B}^{(h)} \left(\tilde{B}^{(h)}\right)^T, \quad (8)$$

where $\tilde{B}^{(h)} = B^{(h)}\tilde{C}^{(h)}$ is a normalized partition matrix, $\tilde{C}^{(h)} = (\mathbf{diag}(\vec{e}^T B^{(h)}))^{-1/2}$ and $\vec{e} = (1, 1, \cdots, 1)^T \in R^{N\times 1}$. This kernel matrix indicates the relationships between genes at the hth partition; that is, the information extracted from this partition has been contained in the kernel matrix. As a result, all the information extracted from multiscale module partitions can be recorded in the set of module kernel matrices $\{K^{(h)} \mid h = 1 \sim H\}$. And then, the disease-relatedness scores of all genes for module partition Ψ_h can be calculated by

$$\vec{s}_G^{(h)} = K^{(h)} \vec{1}. \quad (9)$$

This provides another possible way to understand the scorings based on multiscale module partitions.

Probabilistic model for integration of multiple gene rankings

Integration of multiple gene rankings is an important way to fuse information from multiple features [111–113]. The Bayesian theory provides a usefully theoretical framework for integrating multi-feature information. Here, we introduce a probabilistic model for the integration of multiple gene rankings based on the Bayesian theory. By considering the set of above module partitions as a set of features $\{f_h = \Psi_h | h = 1 \sim H\}$, the comprehensive conditional probability of a candidate gene g being related to a disease can be expressed as,

$$\begin{aligned} P\left(x_g \left| \{f_h\}\right.\right) &= P\left(x_g\right) P\left(\{f_h\} \left| x_g\right.\right) / P\left(\{f_h\}\right) \\ &= P\left(x_g\right) \prod_{h=1}^{H} P\left(f_h \left| x_g\right.\right) / P\left(\{f_h\}\right) \end{aligned}, \quad (10)$$

where $P(x_g)$ denotes the prior probability of a gene being at state x_g (see Supplementary Note 4, see Supplementary Data available online at https://academic.oup.com/bib). $P(f_h|x_g) = P(f_h)P(x_g|f_h)/P(x_g)$, and the prior probabilities $P(\{f_h\})$ and $P(f_h)$ are independent of specific genes, so the conditional probability about all features can be rewritten as $P(x_g|\{f_h\}) \propto P(x_g)\prod_{h=1}^{H} P(x_g|f_h)/P(x_g)$. Here, it can be regarded as the likelihood ratio of a posteriori probability to a priori probability. For convenience, the comprehensive scores of genes can be evaluated by the logarithm of the conditional probability above,

$$\vec{s}_G = \log\left(P\left(x_g\right)\right) + \sum_{h=1}^{H} \log\left(P\left(x_g \left| f_h\right.\right) / P\left(x_g\right)\right). \quad (11)$$

The prior knowledge $P(x_g)$ can contribute to the evaluation of the genes' scores if it is available, or it can be set as a constant if no prior knowledge is available for specific genes.

In order to calculate the final scores of candidate genes, it is necessary to provide an explicit mathematical form of the above conditional probability function (CPF) $P(x_g|f_h)$ about each feature (denoted as CPF). For each module partition, we have calculated the scoring list $\vec{s}_G^{(h)}$ of candidate genes being related to a disease, and get the ranking list $\vec{r}^{(h)}$ of the genes. The higher ranking of a gene generally means its larger probability of disease relatedness, which provides a possible way to the definition of CPF. In this study, we design three explicit forms of CPF (denoted as CPF1, CPF2 and CPF3), which correspond to three variants of HyMM.

Without loss of generality, given a ranking list of genes $\overrightarrow{r}_G^{(h)} = (r_1^{(h)}, r_2^{(h)}, \ldots, r_g^{(h)}, \ldots, r_N^{(h)})^T$ based on the decreasing order of disease-relatedness of genes, CPF can be defined as follows,

CPF1: $P(x_g|f_h) \propto 1 - r_g^{(h)}/N$;

CPF2: $P(x_g|f_h) \propto 1/r_g^{(h)}$;

CPF3: $P(x_g|f_h) \propto \exp(-\beta_h r_g^{(h)})$,

where $\beta = \{\beta_h\}$ is a set of tunable parameters (see Supplementary Note 4, see Supplementary Data available online at https://academic.oup.com/bib, for details).

For CPF1, $P(x_g|f_h)$ linearly varies with the value of gene ranking; for CPF2, $P(x_g|f_h)$ is inversely proportional to the value of gene ranking; for CPF3, $P(x_g|f_h)$ exponentially decreases with the value of gene ranking. CPF2 and CPF3 are typically nonlinear functions, which decay more strongly than CPF1 (Figure S1, see Supplementary Data available online at https://academic.oup.com/bib). As a result, CPF2 and CPF3 prefer genes with higher rankings, i.e. genes at the top of the ranking list, because the higher-ranked genes will be assigned the relatively higher possibility values than other lower-ranked genes (Figure S1, see Supplementary Data available online at https://academic.oup.com/bib). This will be conducive to the mining of disease-related genes.

Given the above union matrix of gene rankings from multiscale modules, a comprehensive scoring list $\overrightarrow{s}_G$ of genes can be generated by the above integration strategy. Then, we get the ranking list of genes $\overrightarrow{r}_{MM} = generank(\overrightarrow{s}_G)$ by the decreasing order of genes' scores (see Figure 1). The above process can be regarded as a raw algorithm for disease-gene prediction (denoted as MM for simplicity).

The scorings/rankings from multiscale modules may provide useful and complementary information for disease-gene prediction, which is different from that of many other algorithms based on various principles, e.g. network propagation. So, we further integrate the ranking list $\overrightarrow{r}_{MM}$ with that (denoted by $\overrightarrow{r}_{TA}$) of network propagation, e.g. the random walk with a restart in HO/HE networks (see Figure 1). The final scores/rankings of genes will be used to prioritize candidate genes. We will show that this integration can very effectively enhance the ability to predict disease-related genes due to information complementarity.

Parameter estimation based on functional information

Because of the good performance of CPF3, it will be used as the default form of $P(x_g|f_h)$ in this study. This integration strategy for multiple gene rankings provides a possible theoretical explanation for classical rank aggregation methods such as Borda count [114], since it degenerates into the arithmetic mean of rankings about multiple features when the parameters $\beta \equiv 1$. However, different from the classical Borda count, the optimization of $\beta = \{\beta_h\}$ can further improve the ability to disease-gene prediction, e.g. by using functional information such as GO annotations and PW.

Module partitions at different scales provide different levels of information, which have different degrees of importance for problems such as disease-gene prediction. To study the statistical properties of module partitions at different scales and their importance in disease-gene prediction, we therefore define functional consistency metrics of network modules and module partitions.

Definition 6. We firstly introduce the functional consistency of a module M_m with respect to a functional group G_f, by

$$C_{f,m} = \left|G_f \cap M_m\right| / \left|M_m\right|, \tag{12}$$

where a functional group denotes a set of genes with common characteristics and $|*|$ denotes the number of elements in the group.

Definition 7. For a set of functional groups, the functional consistency of a module is defined as the maximal functional consistency over all the functional groups

$$C_m = \max_f C_{f,m}. \tag{13}$$

Definition 8. The functional consistency of a module partition Ψ_h is defined as the weighted arithmetic mean of the functional consistency scores of related modules in the module partition

$$C^{(h)} = \sum\nolimits_m \omega_m C_m / \sum\nolimits_m \omega_m, \tag{14}$$

where ω_m is a parameter for module M_m, which can be a constant or proportional to the module size.

We will use the functional consistency metrics to quantify the functional relevance of network modules and module partitions, based on the set of functional groups for GO/PW/DG. This may provide insights for parameter estimation of $\beta = \{\beta_h\}$ so as to improve the ability of HyMM to disease-gene prediction.

Evaluation methods

We implement the above procedure of HyMM (including the multiscale algorithms) by Matlab (2016 version). To evaluate the performance of algorithms in disease-gene prediction, we use two evaluation strategies: traditional 5-fold cross-validation (5FCV) and independent test (IndTest). (i) For 5FCV, known disease-related genes for each disease are randomly split into five subsets. In each realization, one of the subsets is treated as a test set, while the rest is treated as a training set. (ii) For IndTest, the disease-gene associations in the MeSH dataset are used as a training set, and the disease-gene associations that only belong to the DisGeNet dataset are used as a test set.

To construct the candidate set of genes, which consists of a test set of genes and a control set of genes, we will construct two kinds of control sets: artificial linkage-interval control set (ALICS) and whole-genome control set (WGCS). (i) For ALICS, each test gene selects 99 control genes from genes closest to this test gene on the same chromosome. This simulates the scenario with disease-related mutation locations (e.g. derived from the genome-wide association study or linkage analysis) [53]. (ii) For WGCS, all unknown genes outside the training and test sets are used as a control set. This simulates the scenario without the information of disease-related mutation locations.

Then, based on the ranking list of candidate genes, several standard evaluation metrics (AUPRC, Recall, and Precision) are used to quantify the performance of prediction algorithms. (i) AUPRC denotes the area under the Precision-Recall curve (PRC), where the PRC curve has Recall on the x-axis and Precision on the y-axis. This is a widely used metric to comprehensively evaluate the performance of algorithms. (ii) Recall measures the ratio of known disease-related genes found in the top-k ranking list compared to the test set, which focuses on how many disease-related genes in the test set have been retrieved. (iii) Precision (Prec) measures the probability of discovering known disease-related genes in the top-k ranking list. Recall and Precision as a function of k-value can provide an intuitive comparison for the local performance of prediction algorithms.

Experimental results

In this section, we first study the effectiveness of multiscale modules and display the good performance of the HyMM framework in disease-gene prediction by a series of experimental tests, including the effect of various factors such as the CPF, multiscale algorithm, parameter estimation based on functional information, and sampling of multiscale module partitions.

Effectiveness of multiscale modules in disease-gene prediction

Here, we study the predictive performance of each module partition being used independently (Figure 2; Figures S2–S4, and Supplementary Note 5, see Supplementary Data available online at https://academic.oup.com/bib). As a whole, the predictive power of the algorithm based on single-scale module partition first has a large upward trend and then a downward trend with the increase of resolution, and the downward trend appears earlier and is more obvious in the HO network. This clearly indicates that different scale module partitions have different levels of importance in disease-gene prediction.

The main reason behind the above phenomenon is that as the resolution increases, modules become more and more fine-grained, and the relevance between nodes in the same module is getting higher and higher. Gradually splitting modules into more fine-grained ones can filter low-relevance nodes while retaining high-relevance nodes in a module, but this will also lead to the loss of information, resulting in the decline of prediction power, because modules without disease-related information are trivial for our scoring strategy.

In order to verify the relationship between the above-mentioned functional relevance of nodes inside modules and the studied scale (resolution), we calculate the functional consistency of module partitions at different scales by using the GO, PW and DG functional groups. The results show that the functional consistency scores of module partitions increase with the increase of resolution (Figure S5, see Supplementary Data available online at https://academic.oup.com/bib). This means that the ratio of similar genes within the modules is increasing with the resolution: these genes are more likely to have the same GO annotations or belong to the same pathway or disease-gene set. This is because the edge density in the modules becomes higher with the increase of resolution. Genes in the modules are more likely to tend to interact with each other and thus have the same or similar functions or participate in a common biological process. Therefore, the module partitions at different scales can provide different levels of information about the relationship between genes. This may provide a more comprehensive understanding of genes and their functions.

Further, with the increase of resolution, disease-related genes also tend to be dispersed into more modules with smaller sizes but stronger functional relevance (e.g. for GO, PW or DG) (Figure S6, see Supplementary Data available online at https://academic.oup.com/bib). As a whole, candidate genes in these modules will be more likely to be disease-related. So, the predictive power of the algorithm based on single-scale module partition gradually increases with the increase of resolution, but the decline of predictive power may appear due to information loss caused by over-filtering of low-relevance nodes, especially when the network is divided into extremely broken modules.

HyMM effectively enhances ability to disease-gene prediction

HyMM outperforms baseline algorithms

Here, we evaluate the performance of HyMM/MM when default setting is used, by comparing to eight baseline algorithms: RWR (Random Walk with Restart) [53], KS (K-Step Markov) [54], VS (Vertex Similarity) [56], ICN (Interconnectedness) [55], RWRH (Random Walk with Restart on Heterogeneous network) [57], CIPHER (Correlating protein Interaction network and PHEnotype network to pRedict disease genes) [58], BiRW (Bi-Random Walk) [59] and KATZ [60] (also see Supplementary Note 6). MM denotes the algorithm that uses only multiscale modules to generate predictive scores. For simplicity, we here define the ratio of performance improvement as

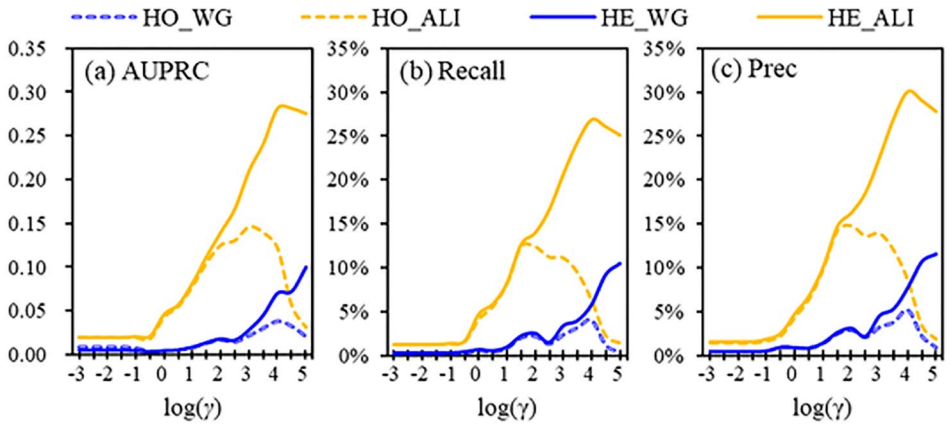

Figure 2. Predictive ability of MO-based module partitions at different scales in disease-gene prediction, as a function of resolution parameter, in HO and HE networks, under different control sets (WGCS and ALICS).

$(x - y)/y$,where x and y denote the results of HyMM and the best baseline algorithm(s), respectively.

The experimental results show that HyMM outperforms all these baseline algorithms, in both the HO and HE networks, under both the ALICS and WGCS control sets (see the results of AUPRC/Recall/Prec in Figure 3; Figures S7 and S10, see Supplementary Data available online at https://academic.oup.com/bib). Specifically, under the ALICS control set, HyMM in the HO network exceeds the best baseline algorithm by 7, 4 and 7% in AUPRC, Recall and Prec metrics, respectively; HyMM in the HE networks exceeds the best baseline algorithm by 27, 25 and 23% in AUPRC, Recall and Prec metrics, respectively. Under the WGCS control set, HyMM in HO exceeds the best baseline algorithms by 9, 21 and 31% in AUPRC, Recall and Prec, respectively; HyMM in HE exceeds the best baseline algorithm by 28, 33 and 32% in AUPRC, Recall and Prec, respectively. The results of top-k Recall/Prec curves have again confirmed the performance of HyMM (Figure 4; Figures S8 and S11, see Supplementary Data available online at https://academic.oup.com/bib).

HyMM provides a useful framework to enhance ability to disease-gene prediction

We systematically test the performance of the HyMM framework by integrating it with other baseline algorithms. For simplicity, we here define the ratio of performance improvement due to the HyMM framework as $(x-y)/y$,where x and y denote the results of the improved and original algorithms, respectively. The results show that the HyMM framework can improve the performance of these algorithms in various test scenarios (Figure 5; Figures S9 and S12, see Supplementary Data available online at https://academic.oup.com/bib). For example, under the ALICS control set, the AUPRC, Recall and Prec of CIPHER improve by 87, 88 and 100%, respectively; the AUPRC, Recall and Prec of KATZ improve by 64, 50 and 61%, respectively. Under the WGCS control set, the AUPRC, Recall and Prec of CIPHER improve by 125, 100 and 157%, respectively; the AUPRC, Recall and Prec of KATZ improve by 140, 145 and 158%, respectively. All these results indicate that HyMM is a very effective framework for integrating multiscale modules to promote the ability to disease-gene prediction.

Moreover, ALICS and WGCS simulate the scenarios with and without disease-related mutation locations, respectively. ALICS has a smaller set of candidate genes than WGCS, due to its more information (about mutation locations). Thus, ALICS generally has relatively larger values of evaluation metrics (AUPRC, Recall, and Precision). In fact, this can also be understood from a random point of view, since the probability of randomly selecting correct genes in a smaller candidate set is usually greater.

Comparison of different CPFs

We have compared the performance of HyMM using different CPFs (CPF1/CPF2/CPF3) (Figures S13 and S14, in Supplementary Note 7, see Supplementary Data available online at https://academic.oup.com/bib). In HE, CPF3 can obtain the best performance under both the ALICS and WGCS control sets. In HO, HyMM using CPF3 has comparable or better performance than that using CPF1/CPF2 under the ALICS and WGCS control sets. So, CPF3 is used as the default form of CPF.

Moreover, it is interesting that the nonlinear forms of CPF (e.g. CPF2/CPF3) are better than the linear form (e.g. CPF1). This means that it is beneficial to give more preference to high-ranking genes in this probabilistic framework.

Comparison of different multiscale algorithms

We have compared the performance of HyMM using different multiscale algorithms (MO/AS/HC). Under the ALICS control set, HyMM using MO outperforms HyMM using AS and HC in most cases, while the Recall- and Prec-values of AS in HE are slightly higher than those of MO (Figure 6; Figure S15, in Supplementary Note 8, see Supplementary Data available online at https://academic.oup.com/bib). Under the WGCS control set, HyMM using MO has good performance,

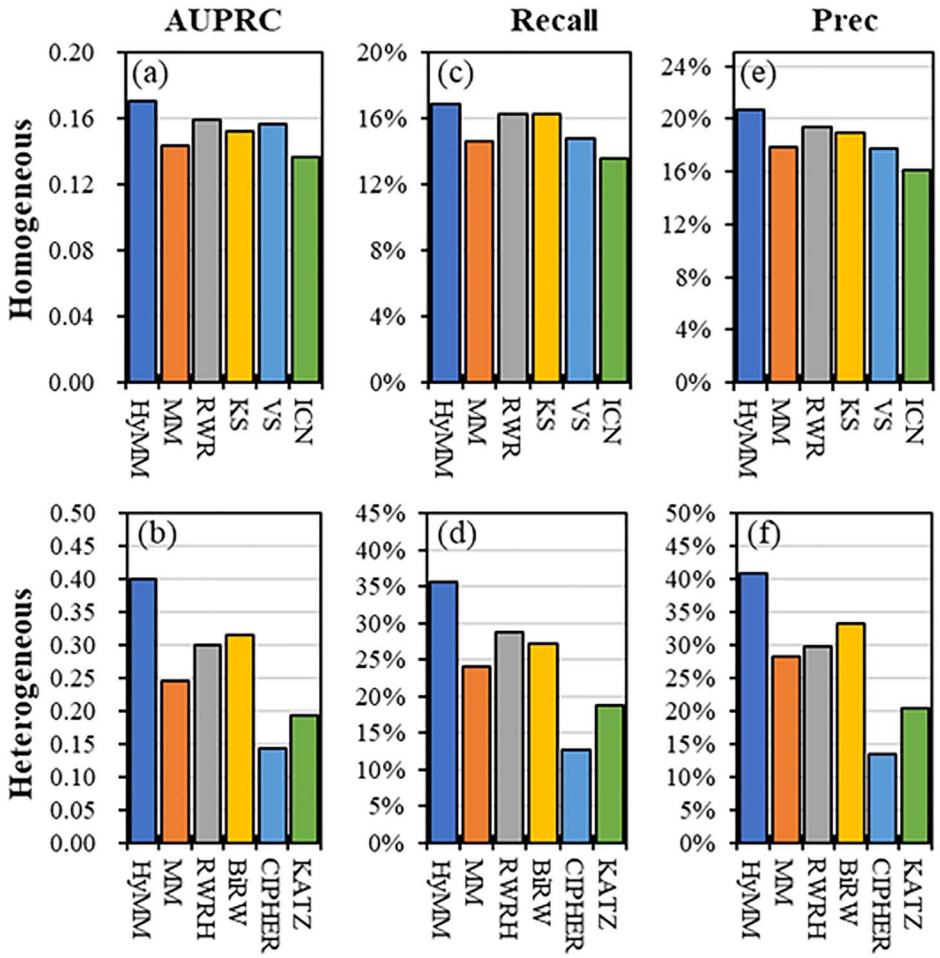

Figure 3. Performance comparison of HyMM/MM to different baseline algorithms under the ALICS control set. (**a** and **b**) AUPRC, (**c** and **d**) Recall and (**e** and **f**) Precision (Prec) in the HO/HE networks.

although HC is the best in HO, and AS is the best in HE (Figure S16, see Supplementary Data available online at https://academic.oup.com/bib). Moreover, MM using MO is better than that using AS and HC in all the cases. Overall, MO can robustly produce better or comparable performance in various tests, so MO is used as the default choice.

Performance improvement through parameter estimation based on functional information

Since module partitions at different scales are of different importance for disease-gene prediction, the optimization of β may further enhance the ability of HyMM, although the default setting has been capable of producing good performance in disease-gene prediction. Due to the close correlation between the functional consistency and predictive power of module partitions, we further use the functional consistency scores of GO/PW/DG as parameter estimation of β. The results show that the parameter estimation can indeed improve the ability of HyMM/MM (using MO, AS or HC) comprehensively (Figure 7; Figures S17 and S18, in Supplementary Note 9, see Supplementary Data available online at https://academic.oup.com/bib).

Stability to sampling of multiscale module partitions

Sampling of multiscale module partitions is closely related to the number of multiscale module partitions and the amount of information extracted from the network. To study the effect of sampling of multiscale module partitions on prediction performance, we evaluate the performance of HyMM for different values of resolution interval $\Delta log\gamma$. The results show that HyMM using MO/AS is very stable to module partition sampling in various scenarios, while HyMM using HC has relatively large fluctuations and declines (Figure 8; Figures S19 and S20, in Supplementary Note 10, see Supplementary Data available online at https://academic.oup.com/bib).

Effect of random shuffling of disease-gene associations

Further, we study the effect of random shuffling of disease-gene associations on the predictive performance of algorithms. We generate a series of disease-gene

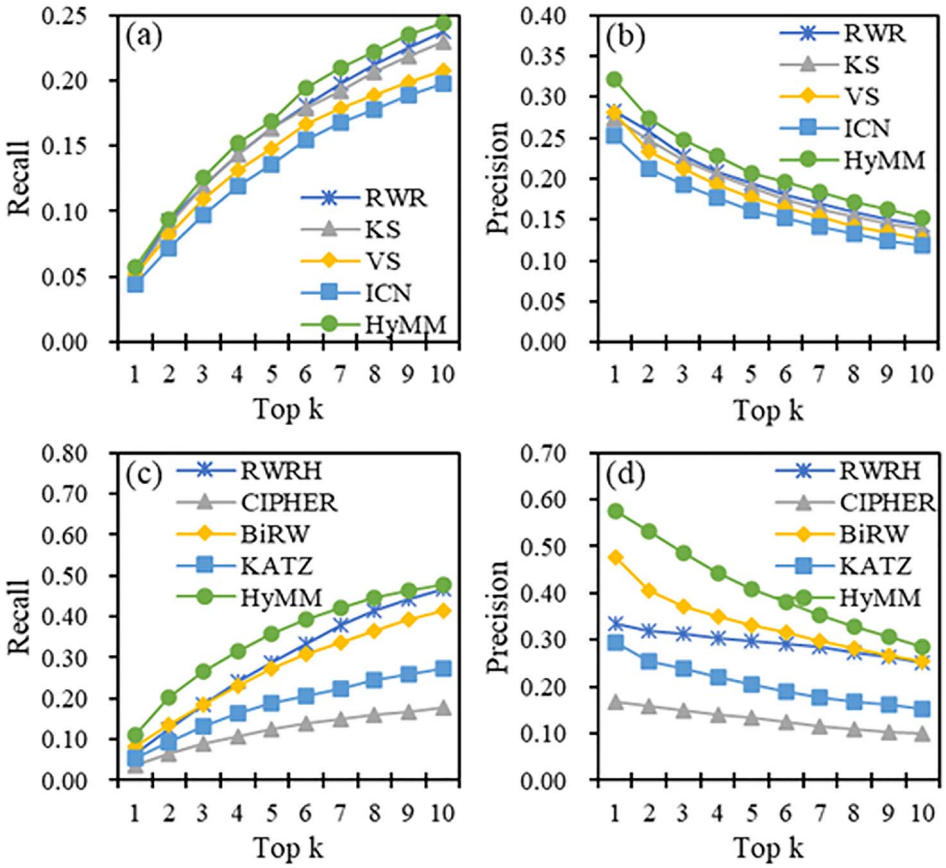

Figure 4. Comparison of local performance of HyMM to different baseline algorithms under the ALICS control set. (**a** and **b**) Top-k Recall and Precision in the HO network; (**c and d**) top-k Recall and Precision in the HE network.

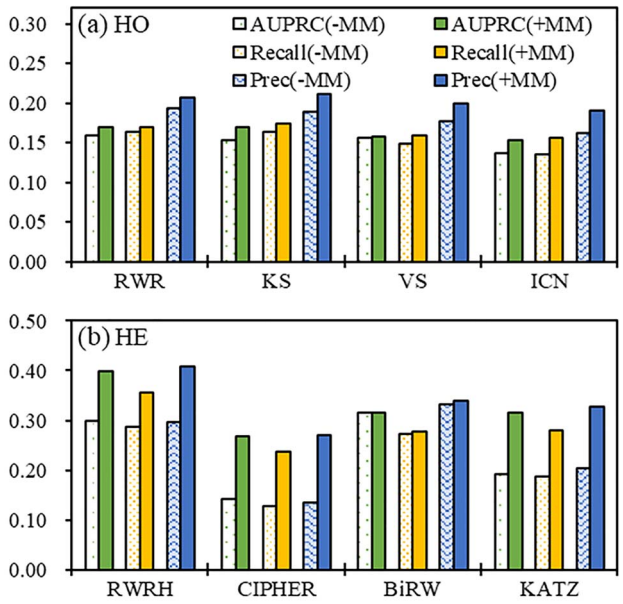

Figure 5. Performance improvement of different baseline algorithms due to the use of multiscale module information in (**a**) HO and (**b**) HE networks, under the ALICS control set. −MM and + MM denote the non-use and use of multiscale module information, respectively.

datasets with different degrees of randomization from no randomization to complete randomization by randomly replacing a certain ratio of known disease-gene associations in a dataset with randomly sampled unknown associations; and then we test the performance of algorithms on these datasets (see Supplementary Note 11, see Supplementary Data available online at https://academic.oup.com/bib, for details).

The results show that HyMM consistently outperforms other baseline algorithms with an increasing degree of randomization (Figure 9, Figures S21 and S22, see Supplementary Data available online at https://academic.oup.com/bib). This again confirms the stable and good performance of HyMM in disease-gene prediction. Moreover, as expected, the performance of all algorithms obviously degrades as the degree of randomization increases. This means that HyMM on real datasets is far superior to that on random datasets, and the known disease-gene associations are critical for effectively inferring disease-gene associations. The predictive power of existing prediction algorithms is extremely dependent on the accumulation of confirmed and reliable disease-gene associations, which is the solid basis for the development of disease-gene-prediction algorithms.

Applications to other datasets

In the above sections, we have demonstrated that HyMM has stable and good performance in disease-gene prediction. Here, we further apply HyMM to

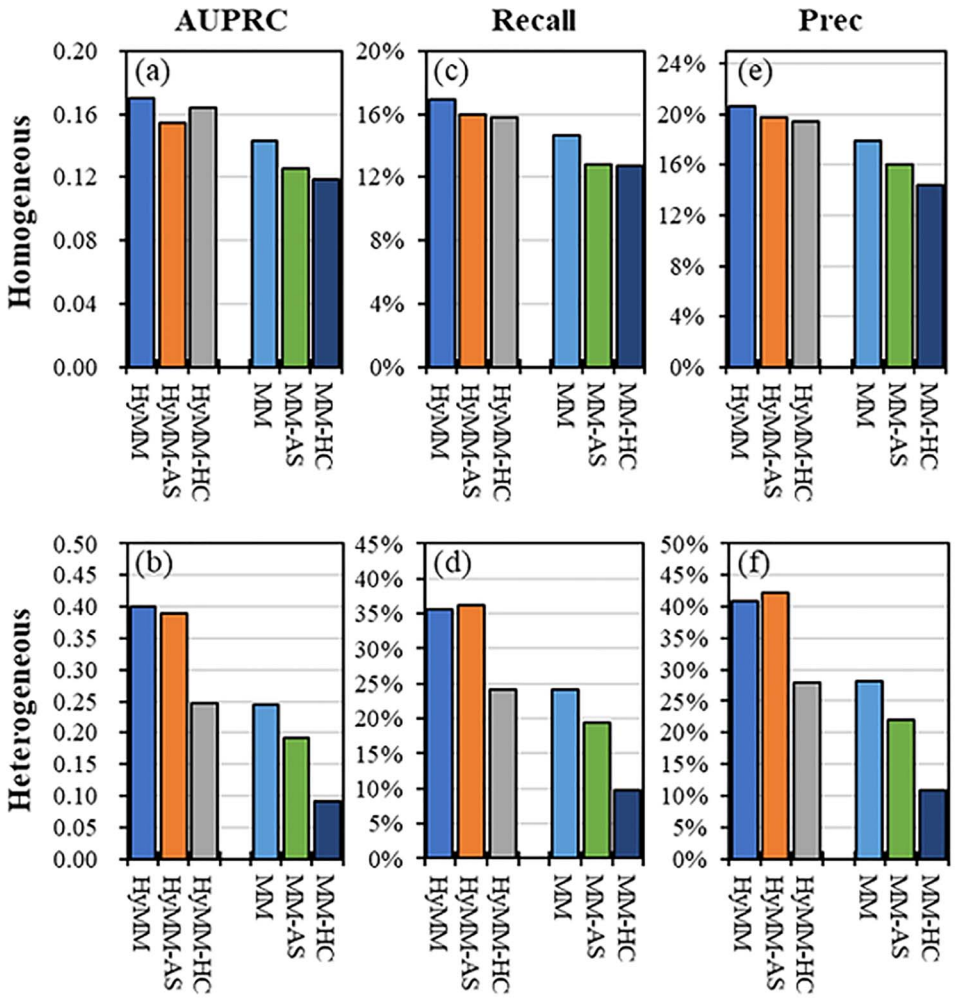

Figure 6. For HyMM/MM, comparison of different multiscale algorithms (MO, AS and HC) under the ALICS control set. (**a** and **b**) AUPRC, (**c** and **d**) Recall and (**e** and **f**) Precision (Prec), in HO and HE networks. HyMM/MM denotes the default algorithms using MO; HyMM-AS/MM-AS and HyMM-HC/MM-HC denote the algorithms using AS and HC, respectively.

the other two datasets: the DISEASES and DisGeNET datasets, e.g. by cross-validation and independent test (see Supplementary Note 12, see Supplementary Data available online at https://academic.oup.com/bib).

Performance evaluation in cross-validation

For the DISEASES dataset, the disease terminology in Disease Ontology (DO) database is used, and thus the similarity scores between diseases are calculated based on the DO database by the DOSE package [115]. For the DisGeNet dataset, the UMLS diseases are mapped into the MeSH diseases according to the disease mappings in the DisGeNet database, and thus the MeSH symptom-based disease similarity scores are still used. Here, the disease-gene associations in the two datasets will be used as a benchmark in turn, and we have tested the performance of HyMM in the two datasets by cross-validation experiments (Figures S23 and S24, see Supplementary Data available online at https://academic.oup.com/bib). In the DISEASES dataset, HyMM has higher or comparable values of AUPRC/Recall/Prec than the best baseline algorithm(s) under both the ALICS control set, and HyMM has good overall performance under the WGCS control set (see Supplementary Note 12, see Supplementary Data available online at https://academic.oup.com/bib). In the DisGeNet dataset, HyMM consistently has higher values of AUPRC/Recall/Prec than the best baseline algorithm(s) under both the control sets. These results show that, as in the MeSH dataset, HyMM also has good performance when applied to the datasets, further confirming the effectiveness of HyMM in disease-gene prediction.

Performance evaluation on external dataset

Furthermore, we evaluate HyMM by experimental test on the external dataset (also denoted as IndTest). We calculate the scores of candidate genes by using the MeSH dataset of disease-gene associations as a training set and then evaluate the prediction performance by using the disease-gene associations belonging to DisGeNet (excluding the training set) as a test set, since DisGeNet is one of the largest publicly available datasets of disease-related genes. In this test, HyMM shows higher Recall- and Prec-values than the baseline algorithms

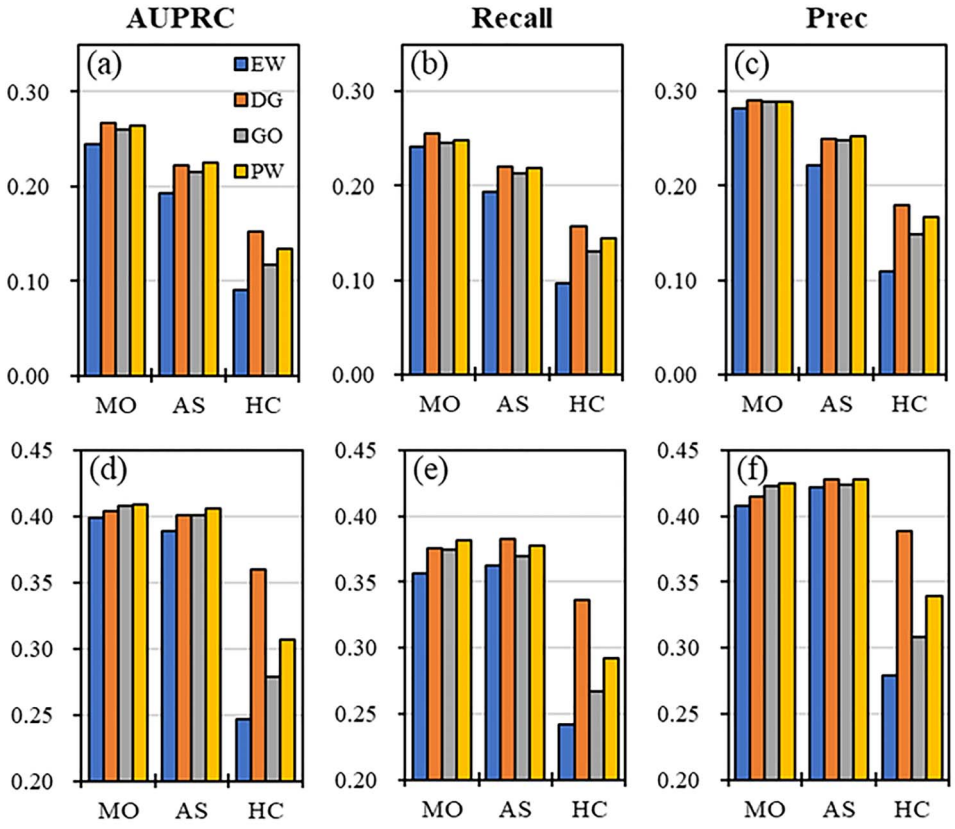

Figure 7. Due to the use of functional information of DG/GO/PW, performance improvement of (**a**–**c**) MM and (**d**–**f**) HyMM (HE) (using MO, AS and HC), under the ALICS control set. EW denotes that equal weight is used.

under both the ALICS and WGCS control sets (Figure S25, see Supplementary Data available online at https://academic.oup.com/bib). Especially, the Prec of HyMM is obviously better than the baseline algorithms.

Applications to specific diseases

In the above sections, we have confirmed the ability of HyMM to a disease-gene prediction by its average performance in different datasets. Here, we further display the predictive ability of HyMM for specific diseases.

Disease-specific performance evaluation

We first study the effectiveness of the HyMM framework in enhancing the ability of predicting specific disease-related genes by using the MeSH dataset as a benchmark. The results of AUPRC/Recall/Prec show that HyMM can improve the ability of predicting disease-related genes for many diseases such as AD (see Figures S26–S28, see Supplementary Data available online at https://academic.oup.com/bib).

Then, we further study the performance of HyMM for AD and some related diseases. Figure 10 and Figures S29 and S30 (see Supplementary Data available online at https://academic.oup.com/bib) display the results of AD, Huntington disease (HD), Parkinson disease (PD), Lewy Body disease (LBD), Frontotemporal Lobar Degeneration (FLD), Anxiety Disorder (Anxiety), Major Depressive Disorder (MDD), and Depressive Disorder (DD). The results show that, under both ALICS and WGCS control sets, for AD and some related diseases (e.g. HD, PD and LBD), HyMM has consistently better performance of AUPRC than the best baseline algorithm(s), except for the results of FLD. Under ALICS, for most diseases (e.g. AD, Anxiety, MDD, DD, HD, PD and LBD), HyMM has comparable or higher values of Recall/Prec compared to the best baselines. Especially for AD, Anxiety, MDD, PD and LBD, HyMM has obviously higher values of Prec. Under WGCS, for most diseases (e.g. AD, Anxiety, MDD, DD, HD and LBD), HyMM has higher values of Recall/Prec compared to the best baselines, except for the results of PD and FLD. Especially for AD, MDD, DD and HD, HyMM has obviously higher values of Recall/Prec. see Supplementary Note 13 (see Supplementary Data available online at https://academic.oup.com/bib) for details.

Overall, HyMM has good performance for AD and many related diseases. Especially for AD and some related diseases, HyMM has better performance than the best baseline algorithm(s), though it is not specially designed for these diseases.

Case study

AD is a progressive neurodegenerative disease and the most common dementia. Its prevalence is increasing in our aging population, resulting in a huge socio-economic burden [116–118]. AD involves specific onset and course of age-related cognitive and functional decline, as well as

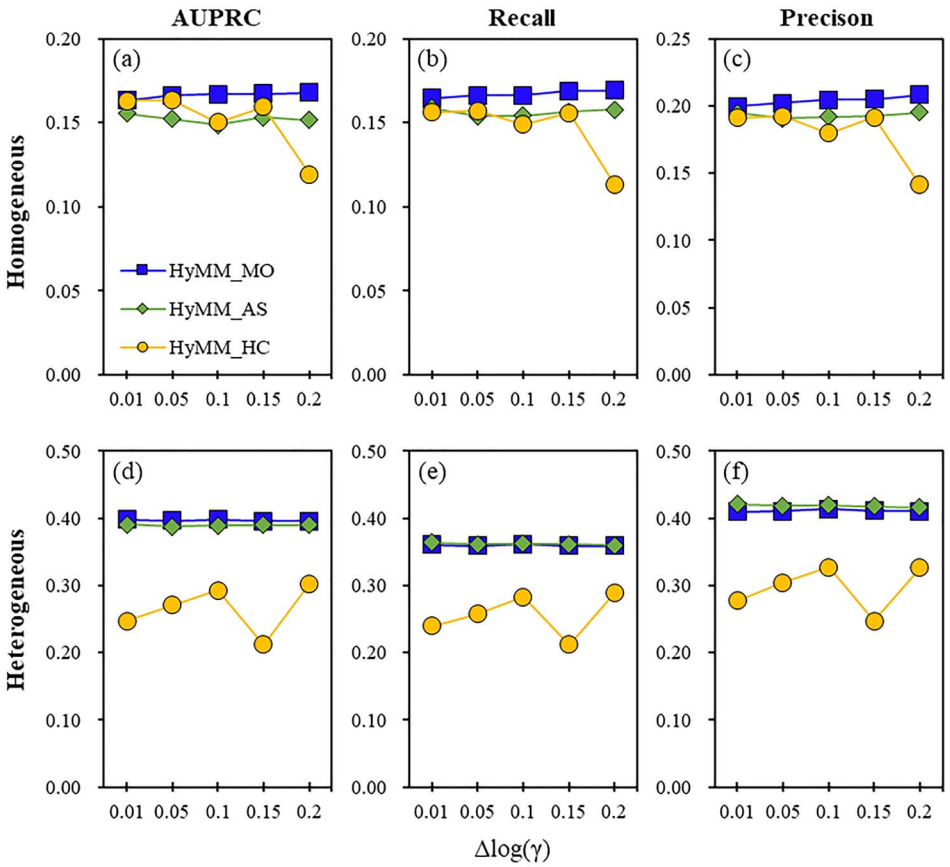

Figure 8. Performance stability of HyMM using MO/AS/HC (i.e. HyMM_MO, HyMM_AS and HyMM_HC) under the ALICS control set, as a function of resolution interval: (**a**) AUPRC, (**b**) Recall and (**c**) Precision in the HO network; (**d**) AUPRC, (**e**) Recall and (**f**) Precision in the HE network.

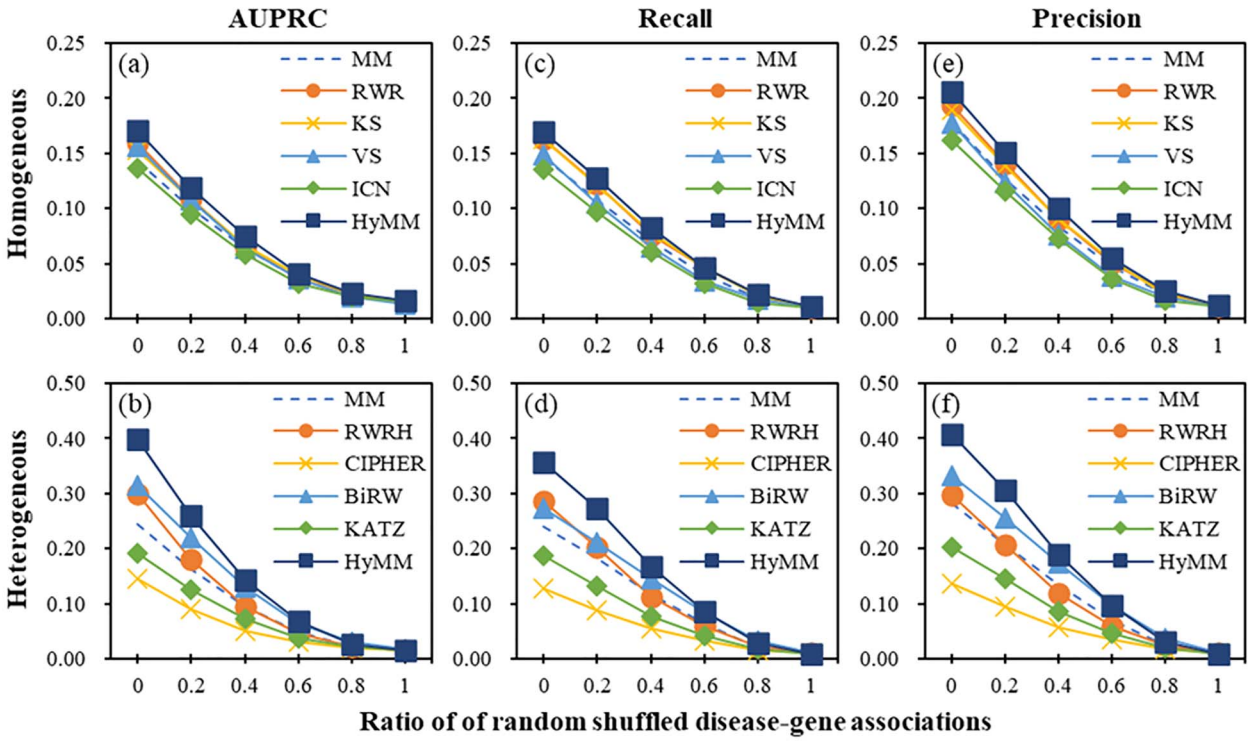

Figure 9. Effect of random shuffling of disease-gene associations on predictive performance under the ALICS control set, as a function of ratio of shuffled disease-gene associations: (**a** and **b**) AUPRC, (**c** and **d**) Recall and (**e** and **f**) Precision, in the HO and HE networks, respectively.

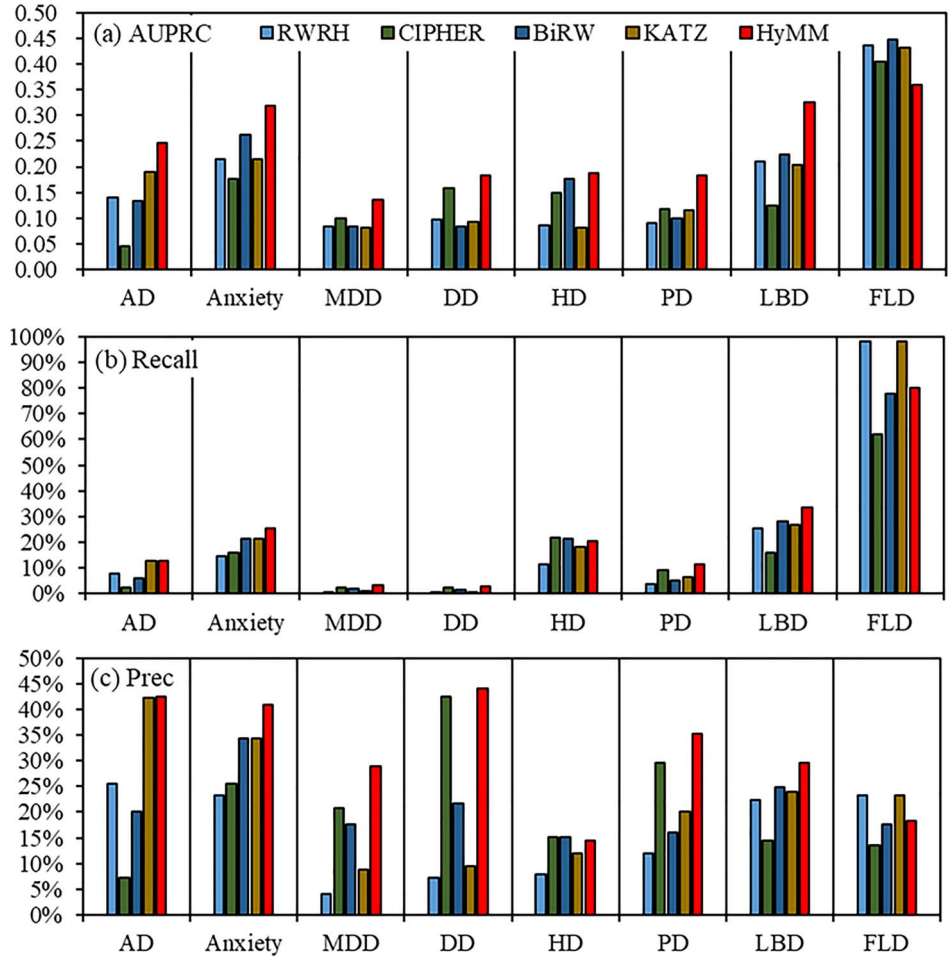

Figure 10. Performance of HyMM and other methods for specific diseases in DisGeNet dataset under ALICS control set: (**a**) AUPRC, (**b**) Recall and (**c**) Prec.

specific neuropathology. It is a highly hereditary disease with high complexity, and identifying AD-related genes is of great significance for determining its therapeutic targets [118]. Here, we calculate the scores of candidate genes related to AD by using the known disease-gene associations in the MeSH dataset as a training set and then obtain the top-20 genes from the ranking list of candidate genes based on the decreasing order of the scores (Table S1, see Supplementary Data available online at https://academic.oup.com/bib).

Through literature verification, we find that some biomedical studies have implied the associations between AD and many genes in this list of candidate genes [119–125]. For example, Park *et al.* [119] showed that ALK was important to the tau-mediated AD pathology; Annunziata *et al.* [120] showed that the deficiency of NEU1 caused the occurrence of an AD-like amyloidogenic process; Qi *et al.* [126] showed that GAA promoted Aβ clearance by promoting autophagy via the Axl/Pak1 signaling pathway in microglial cells and improved cognitive deficiency in a mouse model; Michele *et al.* [123] observed a statistically significant increase of CNVs for C4B in AD patients, suggesting a possible role for C4A CNVs in the risk of AD; Pichiah *et al.* [124] showed that C4B was differentially expressed in AD; Lian *et al.* [121] showed that the dysregulation of neuron–glia interaction through NFκB/C3/C3aR signaling might lead to synaptic dysfunction in AD; Rasmussen *et al.* [122] confirmed that the low baseline levels of complement C3 were associated with a high risk of AD; Stoye *et al.* [125] demonstrated that APOA1 might be a key factor within intestine altered in AD-like pathology. Rai *et al.* [127] showed that the MTHFR C677T polymorphism was associated with an increased risk of AD; Feng *et al.* showed that the autophagosome-lysosome fusion could be repressed by the AD-like MAPT accumulation, showing a vicious cycle of MAPT accumulation and autophagy deficit in the chronic course of AD [128]. MTHFR and MAPT have been recorded as related to AD in DisGeNet.

By the enrichment analysis of the above genes, we obtain the most relevant KEGG pathways and GO terms (Tables S2 and S3, see Supplementary Data available online at https://academic.oup.com/bib), many of which are known to be related to AD, such as the pathways (Lysosome, Metabolic pathways, Oxidative phosphorylation and PD) and the GO annotations (myeloid leukocyte activation, leukocyte mediated immunity, regulated

exocytosis, oxidation–reduction process, glycosphingolipid metabolic process, mitochondrial respiratory chain complex I assembly, neutrophil degranulation, energy derivation by oxidation of organic compounds, mitochondrion organization, small molecule metabolic process). As we know, lysosomes are the main digestive compartments in cells that degrade extracellular and intracellular substances by a series of processes (e.g. autophagy, endocytosis and phagocytosis), and the dysfunction of lysosomes leads to the accumulation of undigested substances [129]. For example, pathological aggregates of proteins Aβ and τ can result in AD [130]. Autophagy-lysosome defects appear in the early stage of AD and are considered to be an important factor in the AD process [131]. Removing these aggregates by autophagy and degrading them in lysosomes may be a promising treatment. Many evidences showed that AD is a widespread metabolic disorder that is related to the dysregulation of multiple biochemical pathways [132, 133]. Understanding the metabolic perturbations related to AD is essential to identify new therapeutic targets. The reduction of oxidative phosphorylation enzyme activities may be related to β- Amyloid accumulation or other neurodegenerative processes, which may play a critical role in the pathology of AD [134, 135]. Many neurodegenerative disorders are closely related [136, 137]. For example, iron plays an important role in maintaining the normal physiological function of the brain, and the iron metabolism dysregulation associated with cell injury and oxidative stress often co-occurs in several neurodegenerative diseases such as AD and PD [136].

In addition, we analyze the druggability of the candidate genes (Table S1, see Supplementary Data available online at https://academic.oup.com/bib) and find that there are many genes corresponding to protein targets of approved or clinical trial-phase drug candidates [138, 139], and many genes have a large number of interacting drugs [140], which may be potential therapeutic agents.

Conclusions and discussions

Identifying disease-related genes is important for the study of human diseases. Network-based algorithms for disease-gene prediction are very popular, because human complex diseases are usually considered to be caused by the perturbations or functional abnormalities of biomolecule networks. Multiscale module structure widely exists in the biomolecule networks, but it is not fully utilized in the analysis and prediction of disease-related genes. Therefore, we proposed the hybrid method called HyMM that integrates the information of multiscale modules to more effectively predict disease-related genes. HyMM consists of several key components: the multiscale MO with exponential sampling for extracting multiscale module structure, the disease-relatedness estimation of genes based on multiscale modules and the probabilistic model for integration of multiple gene rankings, along with the parameter estimation based on functional information.

We first revealed the importance of module partitions at different scales in disease-gene prediction by the partition-by-partition analysis of multiscale modules (e.g. by MO, AS and HC). Then, by a series of experimental tests, we verified the good performance of HyMM, and showed the effect of different conditional probability forms and different multiscale module extraction algorithms. Next, we confirmed the performance improvement derived from parameter estimations based on functional information (DG/PW/GO), and the stability of HyMM to multiscale module partition sampling and random shuffling of disease-gene associations. Finally, the applications of HyMM to other datasets as well as specific diseases further demonstrated the effectiveness of HyMM. Overall, HyMM provides an effective framework for integrating multiscale module structure to enhance the ability to predict disease-related genes. This framework can provide useful insights for the study of the multiscale module structure and its application in such issues as a disease-gene prediction.

In this study, multiscale module identification is critical to the HyMM framework, but we confirmed the effectiveness of multiscale modules in enhancing the ability of disease-gene prediction by using only MO and two other multiscale algorithms. There is a great possibility that HyMM can be further improved by using more advanced module identification algorithms, sampling methods and parameter estimation methods. Motifs, i.e. small patterns recurring in a network, widely exist in many biological networks (e.g. metabolic networks and PPI networks), which are generally considered as building blocks of biological networks [141], while we do not specifically consider network motifs in module identification. The study of motifs in networks is an important topic, and there has been some research on network clustering using motifs. For example, HiSCF (Higher-order Structural Clustering Framework) [142] is able to perform the clustering analysis by exploiting a variety of network motifs, which demonstrates that the consideration of higher-order network motifs gains new insight into the analysis of biological networks. Moreover, the extracted multiscale modules are used by HyMM in the way of integrating independent rankings of genes, but they can also form a (feature) matrix reflecting the module affiliations of genes at different scales, which may be used to infer disease-associated genes by kernel method or machine learning algorithm [143–146]. These interesting ideas are worth further trying in the future.

Biomedical data is an important basis for the research of complex diseases and their related genes [7, 8]. Individual status (disease or health) can be reflected through gene expression, which is affected by multiple factors such as gene mutation, methylation and transcription factors, so the analysis of multi-omics data is very important for disease research, which may promote the discovery of unknown biological knowledge. With the devel-

opment of high-throughput sequencing technologies, a large amount of omics data (e.g. from genomics and transcriptomics to proteomics and metagenomics) are continuously being generated [9, 147–149], and disease-related research will benefit from the increase in the amount and type of the data as well as the improvement in data quality [14]. The integrative use of omics data is expected to improve the ability of disease-related association prediction, and may promote the innovation of relevant technologies and methods (e.g. dimension reduction techniques, network embedding, structured sparsity regularization and multilayer network methods), accelerating the development of systems biology [14, 16, 146, 150–155]. However, it is still difficult to manage, analyze and use these data, though many studies for integrative bioinformatics and omics data source interoperability are actively promoting the solution of the related problems [9, 149, 156–158]. For example, HE repositories with multiple formats and different quality levels hinder the integration of genomic data [156]. The disease-related datasets also have similar problems: diversity of disease terminology systems, disease term redundancy, lack/incompleteness of mapping between terminologies in different datasets, data reliability, etc.

Moreover, precision medicine is to realize the personalized diagnosis and treatment of diseases, while the research on disease-gene prediction in literature basically focuses on the disease class or its subclass. Patient-level datasets with genotypic data and phenotypic data provide the possibility to study individual pathogenic genes [10], [154]. Especially with the development of single-cell sequencing technologies, a large amount of cell-level multi-omics data is growing explosively [147, 158], which provides new opportunities for the research of tissue heterogeneity, and cell function, as well as the personalized study of diseases and pathological genes. The integrative analysis of single-cell data at different molecular levels is expected to reveal the overall complexity of biological systems. We believe that these are worthy of further exploration in the future, though the integration of the single-cell data is still a challenge due to their intrinsic heterogeneity.

Key Points

- Developing computational methods for predicting disease-related genes is important to the study of human diseases, due to the high cost and time consumption of biological experiments.
- We proposed a hybrid framework for disease-gene-prediction by integrating multiscale module structure (HyMM), which can utilize multiscale information from local to global structure to more effectively predict disease-related genes.
- HyMM extracts module partitions from local to global scales by multiscale modularity optimization with exponential sampling, and estimates the disease relatedness of genes in partitions by the abundance of disease-related genes within modules. A probabilistic model for aggregation of gene rankings is designed in order to integrate multiple predictions derived from multiscale module partitions and network propagation, and a parameter estimation strategy based on functional information is proposed to further enhance HyMM's predictive power.
- By a series of experiments, we reveal the importance of module partitions at different scales, and verify the good performance of HyMM and its further performance improvement derived from the parameter estimation.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

National Key Research and Development Program of China (Grant No. 2019YFA0706202); the Training Program for Excellent Young Innovators of Changsha (Grant No. kq2106075), the Fundamental Research Funds for the Central Universities of Central South University (Grant no. 2019zzts279), the Project funded by China Postdoctoral Science Foundation (Grant No.2021M703633). National Natural Science Foundation of China (Grant No. 61702054).

References

1. Kann MG. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinform* 2010;**11**:96–110.
2. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. *Brief Funct Genomics* 2011;**10**:280–93.
3. Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 2012;**13**:523–36.
4. Barabasi A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.
5. Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. *N Engl J Med* 2009;**360**:1699–701.
6. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;**33**:228–37.
7. Zhang W, Zhang H, Yang H, *et al.* Computational resources associating diseases with genotypes, phenotypes and exposures. *Brief Bioinform* 2019;**20**:2098–115.
8. Zeeshan S, Xiong R, Liang BT, *et al.* 100 years of evolving gene–disease complexities and scientific debutants. *Brief Bioinform* 2019;**21**:885–905.
9. Rombo SE, Ursino D. Integrative bioinformatics and omics data source interoperability in the next-generation sequencing era—editorial. *Brief Bioinform* 2021;**22**:1–2. https://doi.org/10.1093/bib/bbaa398

10. Gutiérrez-Sacristán A, De Niz C, Kothari C, *et al.* GenoPheno: cataloging large-scale phenotypic and next-generation sequencing data within human datasets. *Brief Bioinform* 2020;**22**:55–65.
11. Luo P, Xiao Q, Wei P-J, *et al.* Identifying disease-gene associations with graph-regularized manifold learning. *Front Genet* 2019;**10**:270.
12. Li YI, Wong G, Humphrey J, *et al.* Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat Commun* 2019;**10**:994.
13. del Sol A, Balling R, Hood L, *et al.* Diseases as network perturbations. *Curr Opin Biotechnol* 2010;**21**:566–71.
14. Yan J, Risacher SL, Shen L, *et al.* Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform* 2017;**19**:1370–81.
15. Ata SK, Wu M, Fang Y, *et al.* Recent advances in network-based methods for disease gene prediction. *Brief Bioinform* 2021;**22**:bbaa303. https://doi.org/10.1093/bib/bbaa303.
16. Oulas A, Minadakis G, Zachariou M, *et al.* Systems bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches. *Brief Bioinform* 2017;**20**:806–24.
17. Bebek G, Koyutürk M, Price ND, *et al.* Network biology methods integrating biological data for translational science. *Brief Bioinform* 2012;**13**:446–59.
18. Leung EL, Cao Z-W, Jiang Z-H, *et al.* Network-based drug discovery by integrating systems biology and computational technologies. *Brief Bioinform* 2012;**14**:491–505.
19. Zhang H, Ferguson A, Robertson G, *et al.* Benchmarking network-based gene prioritization methods for cerebral small vessel disease. *Brief Bioinform* 2021;**22**:bbab006. https://doi.org/10.1093/bib/bbab006
20. Hu K, Hu J-B, Tang L, *et al.* Predicting disease-related genes by path structure and community structure in protein–protein networks. *J Stat Mech Theory Exp* 2018;**2018**:100001.
21. Zeng X, Liao Y, Liu Y, *et al.* Prediction and validation of disease genes using HeteSim scores. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:687–95.
22. Luo J, Liang S. Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. *J Biomed Inform* 2015;**53**:229–36.
23. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods* 2015;**12**:841–3.
24. Peng J, Bai K, Shang X, *et al.* Predicting disease-related genes using integrated biomedical networks. *BMC Genomics* 2017;**18**:1043.
25. Lei X, Zhang Y. Predicting disease-genes based on network information loss and protein complexes in heterogeneous network. *Inform Sci* 2019;**479**:386–400.
26. Cáceres JJ, Paccanaro A. Disease gene prediction for molecularly uncharacterized diseases. *PLoS Comput Biol* 2019;**15**:e1007078.
27. Valdeolivas A, Tichit L, Navarro C, *et al.* Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 2018;**35**:497–505.
28. Lin C-H, Konecki DM, Liu M, *et al.* Multimodal network diffusion predicts future disease–gene–chemical associations. *Bioinformatics* 2018;**35**:1536–43.
29. Dwivedi SK, Tjärnberg A, Tegnér J, *et al.* Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat Commun* 2020;**11**:856.
30. Menche J, Sharma A, Kitsak M, *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015;**347**:841.
31. Kovács IA, Luck K, Spirohn K, *et al.* Network-based prediction of protein interactions. *Nat Commun* 2019;**10**:1240.
32. Yang K, Lu K, Wu Y, *et al.* A network-based machine-learning framework to identify both functional modules and disease genes. *Hum Genet* 2021;**140**:897–913.
33. Liu Y, Guo Y, Liu X, *et al.* Pathogenic gene prediction based on network embedding. *Briefings in Bioinformatics* 2021;**22**:bbaa353. https://doi.org/10.1093/bib/bbaa353.
34. Xiang J, Zhang J, Zheng R, *et al.* NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Brief Bioinform* 2021;**22**:bbab080.
35. Oti M, Brunner H. The modular nature of genetic diseases. *Clin Genet* 2007;**71**:1–11.
36. Gustafsson M, Nestor CE, Zhang H, *et al.* Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med* 2014;**6**:82.
37. Sharma A, Menche J, Huang CC, *et al.* A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet* 2015;**24**:3005–20.
38. Lage K, Karlberg EO, Størling ZM, *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;**25**:309–16.
39. Fortunato S, Barthélemy M. Resolution limit in community detection. *Proc Natl Acad Sci U S A* 2007;**104**:36–41.
40. Xiang J, Tang Y-N, Gao Y-Y, *et al.* Phase transition of surprise optimization in community detection. *Phys A: Stat Mech Appl* 2018;**491**:693–707.
41. Xiang J, Wang Z-Z, Li H-J, *et al.* Community detection based on significance optimization in complex networks. *J Stat Mech Theory Exp* 2017;**2017**:053213.
42. Choobdar S, Ahsen ME, Crawford J, *et al.* Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**:843–52.
43. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. *Nat Rev Genet* 2016;**17**:615.
44. Lee LY-H, Loscalzo J. Network medicine in pathobiology. *Am J Pathol* 2019;**189**:1311–26. https://doi.org/10.1016/j.ajpath.2019.03.009
45. Wang W, Han R, Zhang M, *et al.* A network-based method for brain disease gene prediction by integrating brain connectome and molecular network. *Brief Bioinform* 2022;**23**:bbab459. https://doi.org/10.1093/bib/bbab459
46. Fortunato S, Hric D. Community detection in networks: a user guide. *Phys Rep* 2016;**659**:1–44.
47. Jin D, Yu Z, Jiao P, *et al.* A survey of community detection approaches: from statistical modeling to deep learning. *IEEE Trans Knowl Data Eng* 2021;1–1.
48. Singhal A, Cao S, Churas C, *et al.* Multiscale community detection in Cytoscape. *PLoS Comput Biol* 2020;**16**:e1008239.
49. Ruan P, Wang S. DiSNEP: a disease-specific gene network enhancement to improve prioritizing candidate disease genes. *Brief Bioinform* 2021;**22**:bbaa241. https://doi.org/10.1093/bib/bbaa241
50. Ding P, Ouyang W, Luo J, *et al.* Heterogeneous information network and its application to human health and disease. *Brief Bioinform* 2019;**21**:1327–46.
51. Dotolo S, Marabotti A, Rachiglio AM, *et al.* A multiple network-based bioinformatics pipeline for the study of molecular mech-

anisms in oncological diseases for personalized medicine. *Brief Bioinform* 2021;**22**:bbab180. https://doi.org/110.1093/bib/bbab180.
52. van Dam S, Võsa U, van der Graaf A, *et al.* Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform* 2018;**19**:575–92.
53. Köhler S, Bauer S, Horn D, *et al.* Walking the Interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
54. Chen J, Bardes EE, Aronow BJ, *et al.* ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;**37**:W305–11.
55. Hsu C-L, Huang Y-H, Hsu C-T, *et al.* Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics* 2011;**12**:1–12.
56. Zhu C, Kushwaha A, Berman K, *et al.* A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Syst Biol* 2012;**6**:1–9.
57. Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 2010;**26**:1219–24.
58. Wu X, Jiang R, Zhang MQ, *et al.* Network-based global inference of human disease genes. *Mol Syst Biol* 2008;**4**:189–9.
59. Xie M, Xu Y, Zhang Y, *et al.* Network-based phenome-genome association prediction by bi-random walk. *PLoS One* 2015;**10**:e0125138.
60. Singh-Blom UM, Natarajan N, Tewari A, *et al.* Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS One* 2013;**8**:e58977.
61. Xiang J, Zhang N-R, Zhang J-S, *et al.* PrGeFNE: predicting disease-related genes by fast network embedding. *Methods* 2021;**192**:3–12.
62. Liu X, Liu Z-P, Zhao X-M, *et al.* Identifying disease genes and module biomarkers by differential interactions. *J Am Med Inform Assoc* 2012;**19**:241–8.
63. Kitsak M, Sharma A, Menche J, *et al.* Tissue specificity of human disease module. *Sci Rep* 2016;**6**:35241.
64. Sun PG, Gao L, Han S. Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci* 2011;**7**:61–73.
65. Akram P, Liao L. Prediction of missing common genes for disease pairs using network based module separation on incomplete human interactome. *BMC Genomics* 2017;**18**:902.
66. Opap K, Mulder N. Recent advances in predicting gene–disease associations. *F1000Research* 2017;**6**:578.
67. Seyyedrazzagi E, Navimipour NJ. Disease genes prioritizing mechanisms: a comprehensive and systematic literature review. *Netw Model Anal Health Inform Bioinf* 2017;**6**:13.
68. Luo P, Chen B, Liao B, *et al.* Predicting disease-associated genes: computational methods, databases, and evaluations. *WIREs Data Mining and Knowledge Discovery* 2021;**11**:e1383.
69. Zolotareva O, Kleine M. A survey of gene prioritization tools for Mendelian and complex human diseases. *J Integr Bioinform* 2019;**16**:20180069. https://doi.org/10.1515/jib-2018-0069
70. Cowen L, Ideker T, Raphael BJ, *et al.* Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* 2017;**18**:551.
71. Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. *J Mol Cell Biol* 2015;**7**:214–30.
72. Dobay MP, Stertz S, Delorenzi M. Context-based retrieval of functional modules in protein–protein interaction networks. *Brief Bioinform* 2017;**19**:995–1007.
73. Lazareva O, Baumbach J, List M, *et al.* On the limits of active module identification. *Brief Bioinform* 2021;**22**:bbab066. https://doi.org/010.1093/bib/bbab1066.
74. Chen B, Fan W, Liu J, *et al.* Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Brief Bioinform* 2014;**15**:177–94.
75. Blondel VD, Guillaume J-L, Lambiotte R, *et al.* Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008;**2008**:P10008.
76. Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a Potts model. *Phys Rev Lett* 2004;**93**:218701.
77. Peixoto TP. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X* 2014;**4**:011047.
78. Ji J, Zhang A, Liu C, *et al.* Survey: functional module detection from protein-protein interaction networks. *IEEE Trans Knowl Data Eng* 2014;**26**:261–77.
79. Zhao B, Wang J, Li M, *et al.* Detecting protein complexes based on uncertain graph model. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**:486–497.
80. Meng X, Xiang J, Zheng R, *et al.* DPCMNE: detecting protein complexes from protein-protein interaction networks via multi-level network embedding. *IEEE/ACM Trans Comput Biol Bioinform* 2021. https://doi.org/10.1109/TCBB.2021.3050102.
81. Wu Z, Liao Q, Liu B. A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. *Brief Bioinform* 2020;**21**:1531–1548. https://doi.org/10.1093/bib/bbz085
82. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
83. Peng X, Wang J, Peng W, *et al.* Protein–protein interactions: detection, reliability assessment and applications. *Brief Bioinform* 2016;**18**:798–819.
84. Mucha PJ, Richardson T, Macon K, *et al.* Community structure in time-dependent, multiscale, and multiplex networks. *Science* 2010;**328**:876–8.
85. Ahn Y-Y, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature* 2010;**466**:761–4.
86. Arenas A, Fernández A, Gómez S. Analysis of the structure of complex networks at different resolution levels. *New J Phys* 2008;**10**:053039.
87. Xiang J, Tang Y-N, Gao Y-Y, *et al.* Multi-resolution community detection based on generalized self-loop rescaling strategy. *Phys A: Stat Mech Appl* 2015;**432**:127–39.
88. Xiang J, Zhang Y, Li J-M, *et al.* Identifying multi-scale communities in networks by asymptotic surprise. *J Stat Mech Theory Exp* 2019;**2019**:033403.
89. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* 2005;**6**:39.
90. Lewis A, Jones N, Porter M, *et al.* The function of communities in protein interaction networks at multiple scales. *BMC Syst Biol* 2010;**4**:100.
91. Wang J, Li M, Chen J, *et al.* A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:607–20.
92. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human Interactome. *PLoS Comput Biol* 2015;**11**:e1004120.

93. Hamosh A, Scott AF, Amberger JS, *et al.* Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**: D514–7.
94. Pletscher-Frankild S, Pallejà A, Tsafou K, *et al.* DISEASES: text mining and data integration of disease–gene associations. *Methods* 2015;**74**:83–9.
95. Piñero J, Bravo À, Queralt-Rosinach N, *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–9.
96. Guala D, Ogris C, Müller N, *et al.* Genome-wide functional association networks: background, data & state-of-the-art resources. *Brief Bioinform* 2019;**21**:1224–37.
97. Lee W-P, Tzou W-S. Computational methods for discovering gene networks from expression data. *Brief Bioinform* 2009;**10**: 408–23.
98. Przytycka TM, Singh M, Slonim DK. Toward the dynamic interactome: it's about time. *Brief Bioinform* 2010;**11**:15–29.
99. Perfetto L, Briganti L, Calderone A, *et al.* SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res* 2015;**44**:D548–54.
100. Pan A, Lahiri C, Rajendiran A, *et al.* Computational analysis of protein interaction networks for infectious diseases. *Brief Bioinform* 2016;**17**:517–26.
101. Van Steen K. Travelling the world of gene–gene interactions. *Brief Bioinform* 2011;**13**:1–19.
102. Zhou X, Menche J, Barabási A-L, *et al.* Human symptoms–disease network. *Nat Commun* 2014;**5**:4212.
103. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**: 15545–50.
104. Consortium TGO. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res* 2018;**47**:D330–8.
105. Kanehisa M, Sato Y, Kawashima M, *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2015;**44**:D457–62.
106. Matthews L, Gopinath G, Gillespie M, *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**:D619–D622. https://doi.org/10.1093/nar/gkn863
107. Wang J, Zhong J, Chen G, *et al.* ClusterViz: a Cytoscape APP for cluster analysis of biological network. *IEEE/ACM Trans Comput Biol Bioinform* 2015;**12**:815–22.
108. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E* 2004;**69**:026113. https://doi.org/10.1103/PhysRevE.69.026113
109. Xiang J, Hu X-G, Zhang X-Y, *et al.* Multi-resolution modularity methods and their limitations in community detection. *European Physical Journal B* 2012;**85**:1–10.
110. Xiang J, Hu K. Limitation of multi-resolution methods in community detection. *Phys A: Stat Mech Appl* 2012;**391**:4995–5003.
111. Badgeley MA, Chikina MD, Sealfon SC. Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation. *Bioinformatics* 2014;**31**:209–15.
112. Datta S, Datta S, Pihur V. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* 2007;**23**:1607–15.
113. Vilo J, Adler P, Kolde R, *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**: 573–80.
114. Li X, Wang X, Xiao G. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Brief Bioinform* 2017;**20**:178–89.
115. Yu G, Wang L-G, Yan G-R, *et al.* DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2015;**31**:608–9.
116. Takizawa C, Thompson PL, van Walsem A, *et al.* Epidemiological and economic burden of Alzheimer's disease: a systematic literature review of data across Europe and the United States of America. *J Alzheimers Dis* 2015;**43**:1271–84.
117. Crous-Bou M, Minguillón C, Gramunt N, *et al.* Alzheimer's disease prevention: from risk factors to early intervention. *Alzheimer's Res Ther* 2017;**9**:71.
118. Naj AC, Schellenberg GD, Consortium ftAsDG. Genomic variants, genes, and pathways of Alzheimer's disease: an overview. *Am J Med Genet B Neuropsychiatr Genet* 2017;**174**:5–26.
119. Park J, Choi H, Kim YD, *et al.* Aberrant role of ALK in tau proteinopathy through autophagosomal dysregulation. *Mol Psychiatry* 2021;**26**:5542–56. https://doi.org/10.1038/s41380-020-01003-y
120. Annunziata I, Patterson A, Helton D, *et al.* Lysosomal NEU1 deficiency affects amyloid precursor protein levels and amyloid-β secretion via deregulated lysosomal exocytosis. *Nat Commun* 2013;**4**:2734.
121. Lian H, Yang L, Cole A, *et al.* NFκB-activated astroglial release of complement C3 compromises neuronal morphology and function associated with Alzheimer's disease. *Neuron* 2015;**85**: 101–15.
122. Rasmussen KL, Nordestgaard BG, Frikke-Schmidt R, *et al.* An updated Alzheimer hypothesis: complement C3 and risk of Alzheimer's disease—a cohort study of 95,442 individuals. *Alzheimers Dement* 2018;**14**:1589–601.
123. Michele Z, Francesca D, Laura D, *et al.* Complement C4A and C4B gene copy number study in Alzheimer's disease patients. *Curr Alzheimer Res* 2017;**14**:303–8.
124. Pichiah PBT, Sankarganesh D, Arunachalam S, *et al.* Adipose-derived molecules–untouched horizons in Alzheimer's disease biology. *Front Aging Neurosci* 2020;**12**:17. doi: 10.3389/fnagi.2020.00017
125. Stoye NM, dos Santos GM, Endres K. Alzheimer's disease in the gut—major changes in the gut of 5xFAD model mice with ApoA1 as potential key player. *FASEB J* 2020;**34**: 11883–99.
126. Qi L-F-R, Liu S, Liu Y-C, *et al.* Ganoderic acid a promotes amyloid-β clearance (in vitro) and ameliorates cognitive deficiency in Alzheimer's disease (mouse model) through autophagy induced by activating Axl. *Int J Mol Sci* 2021;**22**: 5559.
127. Rai V. Methylenetetrahydrofolate reductase (MTHFR) C677T polymorphism and Alzheimer disease risk: a meta-analysis. *Mol Neurobiol* 2017;**54**:1173–86.
128. Feng Q, Luo Y, Zhang X-N, *et al.* MAPT/tau accumulation represses autophagy flux by disrupting IST1-regulated ESCRT-III complex formation: a vicious cycle in Alzheimer neurodegeneration. *Autophagy* 2020;**16**:641–58.
129. Saftig P, Haas A. Turn up the lysosome. *Nat Cell Biol* 2016;**18**: 1025–7.
130. Thal DR, Fändrich M. Protein aggregation in Alzheimer's disease: Aβ and τ and their potential roles in the pathogenesis of AD. *Acta Neuropathol* 2015;**129**:163–5.
131. Zare-shahabadi A, Masliah E, Johnson GVW, *et al.* Autophagy in Alzheimer's disease. *Rev Neurosci* 2015;**26**:385–95.
132. Mahajan UV, Varma VR, Griswold ME, *et al.* Dysregulation of multiple metabolic networks related to brain transmethylation and polyamine pathways in Alzheimer disease: a targeted metabolomic and transcriptomic study. *PLoS Med* 2020;**17**:e1003012.

133. Toledo JB, Arnold M, Kastenmüller G, *et al.* Metabolic network failures in Alzheimer's disease: a biochemical road map. *Alzheimers Dement* 2017;**13**:965–84.
134. Shoffner JM. Oxidative phosphorylation defects and Alzheimer's disease. *Neurogenetics* 1997;**1**:13–9.
135. Manczak M, Park BS, Jung Y, *et al.* Differential expression of oxidative phosphorylation genes in patients with Alzheimer's disease. *Neuromolecular Med* 2004;**5**:147–62.
136. Belaidi AA, Bush AI. Iron neurochemistry in Alzheimer's disease and Parkinson's disease: targets for therapeutics. *J Neurochem* 2016;**139**:179–97.
137. Goedert M. Neurodegeneration. Alzheimer's and Parkinson's diseases: the prion concept in relation to assembled Aβ, tau, and α-synuclein. *Science* 2015;**349**:1255555.
138. Finan C, Gaulton A, Kruger FA, *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* 2017;**9**:eaag1166.
139. Wang Y, Zhang S, Li F, *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res* 2019;**48**: D1031–41.
140. Freshour SL, Kiwala S, Cotto KC, *et al.* Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res* 2020;**49**:D1144–D1151. https://doi.org/10.1093/nar/gkaa1084
141. Milo R, Shen-Orr S, Itzkovitz S, *et al.* Network motifs: simple building blocks of complex networks. *Science* 2002;**298**:824–7.
142. Hu L, Zhang J, Pan X, Yan H, You Z-H. HiSCF: leveraging higher-order structures for clustering analysis in biological networks. *Bioinformatics*, 2021;**7**:542–50.
143. Chen H, Li F, Wang L, *et al.* Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. *Brief Bioinform* 2020;**22**:bbaa068. https://doi.org/010.1093/bib/bbaa068.
144. Karim MR, Beyan O, Zappa A, *et al.* Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform* 2021;**22**: 393–415.https://doi.org/10.1093/bib/bbz170
145. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2016;**19**:325–40.
146. Oh M, Park S, Kim S, *et al.* Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Brief Bioinform* 2020;**22**:66–76.
147. Li Y, Ma L, Wu D, *et al.* Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief Bioinform* 2021;**22**:bbab024. https://doi.org/10.1093/bib/bbab024
148. Comin M, Di Camillo B, Pizzi C, *et al.* Comparison of microbiome samples: methods and computational challenges. *Brief Bioinform* 2020;**22**:88–95.
149. Knyazev S, Hughes L, Skums P, *et al.* Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Brief Bioinform* 2020;**22**:96–108.
150. Meng C, Zeleznik OA, Thallinger GG, *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**:628–41.
151. Zhou T, Liu W, Li N, *et al.* Secure scheme for locating disease-causing genes based on multi-key homomorphic encryption. *Tsinghua Sci Technol* 2022;**27**:333–43.
152. Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform* 2020;**22**:77–87.
153. Tie J, Lei X, Pan Y. Metabolite-disease association prediction algorithm combining DeepWalk and random forest. *Tsinghua Sci Technol* 2022;**27**:58–67.
154. Galano-Frutos JJ, García-Cebollada H, Sancho J. Molecular dynamics simulations for genetic interpretation in protein coding regions: where we are, where to go and when. *Brief Bioinform* 2019;**22**:3–19.
155. Zhang Y, Lei X, Fang Z, *et al.* CircRNA-disease associations prediction based on metapath2vec++ and matrix factorization. *Big Data Mining and Analytics* 2020;**3**:280–91.
156. Pastor Ó, León AP, Reyes JFR, *et al.* Using conceptual modeling to improve genome data management. *Brief Bioinform* 2020;**22**: 45–54.
157. Bernasconi A, Canakoglu A, Masseroli M, *et al.* The road towards data integration in human genomics: players, steps and interactions. *Brief Bioinform* 2020;**22**:30–44.
158. Forcato M, Romano O, Bicciato S. Computational methods for the integrative analysis of single-cell data. *Brief Bioinform* 2020;**22**:bbaa042. https://doi.org/10.1093/bib/bbaa042