frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fgene.2021.726596

Locally Adjust Networks Based on Connectivity and Semantic Similarities for Disease Module Detection

*Jia Liu[1], Huole Zhu[2,3] and Jianfeng Qiu[2,3]**

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China, [2]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, Hefei, China, [3]Information Materials and Intelligent Sensing Laboratory of Anhui Province, School of Artificial Intelligence, Anhui University, Hefei, China

For studying the pathogenesis of complex diseases, it is important to identify the disease modules in the system level. Since the protein-protein interaction (PPI) networks contain a number of incomplete and incorrect interactome, most existing methods often lead to many disease proteins isolating from disease modules. In this paper, we propose an effective disease module identification method IDMCSS, where the used human PPI networks are obtained by adding some potential missing interactions from existing PPI networks, as well as removing some potential incorrect interactions. In IDMCSS, a network adjustment strategy is developed to add or remove links around disease proteins based on both topological and semantic information. Next, neighboring proteins of disease proteins are prioritized according to a suggested similarity between each of them and disease proteins, and the protein with the largest similarity with disease proteins is added into a candidate disease protein set one by one. The stopping criterion is set to the boundary of the disease proteins. Finally, the connected subnetwork having the largest number of disease proteins is selected as a disease module. Experimental results on asthma demonstrate the effectiveness of the method in comparison to existing algorithms for disease module identification. It is also shown that the proposed IDMCSS can obtain the disease modules having crucial biological processes of asthma and 12 targets for drug intervention can be predicted.

Keywords: complex disease, module identification, protein-protein interaction network, locally adjust networks, connectivity and semantic similarities

OPEN ACCESS

Edited by:
Xiangxiang Zeng,
Hunan University, China

Reviewed by:
Jinting Guan,
Xiamen University, China
Huaming Chen,
University of Adelaide, Australia

****Correspondence:***
Jianfeng Qiu
qiujianf@ahu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *17 June 2021*
Accepted: *22 September 2021*
Published: *25 October 2021*

Citation:
Liu J, Zhu H and Qiu J (2021) Locally Adjust Networks Based on Connectivity and Semantic Similarities for Disease Module Detection. Front. Genet. 12:726596. doi: 10.3389/fgene.2021.726596

1 INTRODUCTION

There exist a number of complex diseases, which are not caused by the malfunction of an individual gene product, but the dysfunction of biological systems formed by several disease-related genes (Zheng et al., 2006; Zheng et al., 2008; Schadt, 2009; Zanzoni et al., 2009; Albert-László et al., 2011; Su et al. 2019). These disease-related genes and their products (e.g., proteins) are not randomly distributed on a molecular network, but they prefer to work together as a group for similar biological functions Sol et al., 2010. The above evidence suggests the existence of disease modules, which were firstly defined by Barabasi et al. as the connected subgraphs formed by proteins associated with a disease (Menche et al., 2015). The disease modules can be considered as the characteristic of a particular disease phenotype (Susan Dina et al., 2015). It becomes quite important to identify the disease modules, which is helpful for understanding the molecular mechanisms of disease origin and progression, and thus aiding the identification of synergistic drug combinations (Cheng et al., 2019).

With the rapid accumulation of protein-protein interactions, the investigation of interactions between proteins in the human protein-protein interaction (PPI) networks has become one of the primary approaches for detecting disease modules of complex diseases (Igor et al., 2008; Sebastian et al., 2008; Wang et al., 2011). These approaches usually are performed by using the connectivity information in the PPI network, and can be roughly classified into four categories, i.e., neighborhood scoring methods (Krauthammer et al., 2004; Jonsson and Bates, 2006; Tu et al., 2006; Xu and Li, 2006), seed expanding-based methods (Sharma et al., 2012; Susan Dina et al., 2015; Zhang et al., 2017b), diffusion-based methods Sebastian et al. (2008) and representation learning methods (Härtner et al., 2018). However, the disease modules achieved by these connectivity-based approaches usually show insufficient reliability to illustrate a specific disease phenotype, since nearly 80% of actual associations between proteins are not included in the existing PPI network and these missing associations leave many disease proteins isolated from their disease modules (Menche et al., 2015). Besides, high throughput experiments often produce a large number of interactions with noise, which makes several irrelevant proteins included in the disease module (Cho and Montanez, 2013).

To obtain better detection results, several studies have been performed by combining the protein-protein interaction data with other types of biological data, such as sequence-based features, epigenomic data, gene ontology (GO) annotation and expression patterns (Csaba and Mauno, 2009; Franke et al., 2006b; Liu et al., 2015). Among these biological data, GO annotation has shown to be an effective semantic resource which usually serves as a complement to protein-protein interactions to reflect functional information, where the semantic information of a gene is defined as the molecular function of genes and the biological processes in which the genes are involved (Franke et al., 2006b; Liu et al., 2015). Disease modules achieved by existing approaches have shown the ability to combine the connectivity information with the semantic information for the prioritizing of candidate disease genes (Franke et al., 2006b; Liu et al., 2015). For example, in Franke et al. (2006b), a gene network is developed by the intergation of the GO annotation information, interactions between proteins and microarray coexpressions, and genes are ranked based on the network. In Liu et al. (2015), Liu et al. proposed a method combining the topological similarity in the PPI network with the semantic similarity to select the candidate disease genes. However, the detection results of existing methods need to be further improved, since several unreliable interactions will hinder the detection effectiveness.

Recent studies on complex networks show that an ambiguous community structure can be converted into a structure much clearer than the original one by adding and reducing several links in the network (Su et al., 2021). It is known that about 80% of the disease proteins are disconnected from disease modules because of the incomplete biological network, where these proteins tend to be localized in the neighborhood of the disease modules (Menche et al., 2015). This means that the implementing of removing associations from the PPI network and adding into associations around the known disease proteins can compensate for the incomplete and incorrect interactions between the proteins in the PPI network, which will facilitate the detection of disease modules. For this reason, we proposed a connectivity and semantic similarities based method (termed as IDMCSS) to identify disease modules by locally adjusting a given PPI network in the detection process in a conference paper (Su et al., 2020). The connectivity similarity reflects the closeness of proteins based on protein-protein interactions and the semantic similarity represents functional similarities of proteins based on GO annotation information. In Su et al. (2020), due to the page limitation, the IDMCSS was only briefly presented and some simple experiments demonstrated the effectiveness of the algorithm for disease module identification. In this paper, we give an extended version of the paper in Su et al. (2020) by adding more analysis and discussions on the algorithm. Specifically, we present a detailed description of the strategies used in the IDMCSS and a series of experimental results are reported with detailed discussions to illustrate the competitiveness of the IDMCSS. We also add the related work section to highlight the difference between the IDMCSS and existing algorithms, as well as the complexity analysis of the IDMCSS. To sum up, the IDMCSS algorithm contains the following two main contributions:

1) A strategy of network structure adjustment is proposed to locally change the structure of the existing PPI network by adding several missing links which are likely to be related to disease proteins and removing some existing links which have an extremely weak correlation to disease proteins. To this end, the strong-linked or weak-linked proteins are firstly selected from the neighbors of disease proteins, where the strong-linked proteins and the weak-linked proteins have large and small connective similarities with disease proteins, respectively. Then, two key operators, i.e., adding link operator and removing link operator, are designed to add several links between strong-linked proteins and disease proteins, and remove some links between strong-linked proteins and disease proteins.
2) A disease module detection method IDMCSS is proposed by using the strategy of network structure adjustment based on both connective and semantic similarity. In the proposed method, a strategy to expand the set of disease proteins is tailored for the disease module identification. The proposed IDMCSS is verified to be superior over some representative disease module identification approaches.

The rest of the paper is organized as follows. **Section 2** presents the disease module detection problem and reviews the related methods for disease module identification. Then, we describe the details of the proposed algorithm in **Section 3**.

Section 4 shows the experimental results and **Section 5** concludes the paper and gives the future work.

2 RELATED WORK

Recently, the PPI network has become a popular resource for disease module identification (Cagney et al., 2000; Navlakha and Kingsford, 2010). Several disease protein prioritization strategies have been developed to detect disease modules by taking advantage of the existing PPI networks (Agrawal et al., 2017; Cui et al., 2018; Tian et al., 2020). Due to the unreliability of the connective information, there exist some disease modules that are not observable in the PPI networks (Wu et al., 2013). There are also some approaches which are performed by combining connective information and other information such as GO annotation information and expression patterns, to change the structure of the PPI networks (Liu et al., 2015; Franke et al., 2006a; Luo and Liang, 2015; Zhang et al., 2017a). In what follows, we only recall several approaches based on changing network structure, which can be roughly divided into two groups.

The first group changes the network structure by adding several potential missing links to make the network more reliable or adding extra nodes to connect disassociated disease proteins. In order to achieve a reliable network, Franke et al. (2006a) collected a set of validated protein-protein interactions and made use of GO annotation, coexpression data to predict interactions of the remaining protein pairs by a Bayesian classifier. The achieved network was applied to detect candidate disease proteins. To avoid spurious interactions in the PPI networks, a network was reconstructed by connecting pairs of disconnected proteins in the PPI network whose higher-order topological similarities were larger than a certain threshold, where the higher-order topological similarity between two proteins was measured by a link prediction algorithm. Then, candidate inherited disease proteins were prioritized by a random walk-based algorithm on the reconstructed network (Luo and Liang, 2015). Based on a similar idea, Liu et al. developed an algorithm (CTSS) to detect disease proteins by adding the weak interactions between genes which were not connected in the existing network based on the semantic similarity between them (Liu et al., 2015). Experimental results indicated that the PPI network became more perfect by involving reliable associations. In order to connect known disease proteins to be a coherent network module, a seed connector algorithm was developed to detect disease modules by adding as few extra hidden proteins to the set of known proteins as possible (Wang and Loscalzo, 2018). The newly added proteins have been demonstrated useful, since they show significant biological relevance in terms of their functional similarity to known disease proteins and their enrichment of drug targets.

The second group focuses on eliminating potential incorrect associations in the existing networks to achieve a more reliable network or removing several links which are not related to a particular disease phenotype to obtain a disease-specific network. For instance, in order to eliminate potential incorrect associations, the structure of the human PPI network is adjusted by measuring the correlation coefficient between a pair of connected proteins and removing those with a low correlation coefficient (<0.75) in gene expression data (Liu et al., 2011). In Zhang et al., 2017a), a gene co-expression network was constructed according to the expression patterns of genes, and the links which were not included in the gene co-expression network were removed from the existing PPI network to improve the prediction accuracy of disease proteins. As for a disease-specific network, only the interactions between the immunome proteins in the PPI network were taken into account for the construction of primary immunodeficiencies network, where no new nodes were added, and proteins without interactions were removed (Ortutay and Vihinen, 2008. Similarly, in Bragina et al. (2016), an associative network, which represents molecular interactions between proteins and genes associated with Tuberculosis, was reconstructed and analyzed, and new candidate genes for TB susceptibility were discovered.

Although various network structure based techniques have been developed for the identification of disease modules, traditional approaches are still far from satisfactory, since little approaches focus on dealing with the missing and incorrect links simultaneously. In this paper, we propose a disease module identification method, which is achieved by both adding several potential missing interactions and removing several potential incorrect interactions from the existing PPI networks, based on two types of data, i.e., connective information and semantic information of proteins.

3 THE IDMCSS METHOD

In this section, we give the details of the proposed IDMCSS algorithm. Firstly, the general framework of IDMCSS is presented, and then the network adjustment strategy as well as the way to identify disease proteins which are the main components of IDMCSS are elaborated.

3.1 Framework of IDMCSS

The proposed IDMCSS is a network-based disease module detection method, where the keypoint is to expand a seed module based on an adjusted PPI network. To be specific, let a biological network be G and let the set of known disease proteins be S_0, the IDMCSS performs seven main steps to detect a disease module. First, we initialize the disease protein set S to be the set of known disease proteins S_0, and let the candidate disease protein set C be empty. Then, we select all the neighbors of known disease proteins, i.e., $NS = (b_1, \ldots, b_\alpha)$, based on the current network G, where b_i ($i = 1, \ldots, \alpha$) is a neighbor of a certain node in S. Third, the structure of the current network is locally changed into a new network, G_{new}, by the suggested network adjustment strategy, which focuses on removing the potential incorrect links and adding the potential missing links around the nodes in S. Fourth, the neighbors of the nodes in S, i.e., NS, are updated according to the adjusted network G_{new}. Fifth, we select the protein b from NS which is most likely to be a disease protein by the suggested similarity, and add the node b into the set S and the candidate disease protein set *C*. The above the second to the fifth steps are repeated until a certain disease-related information (gene ontology, differential expression genes, pathways) is not significantly enriched in the set *C*, where the significance estimation used in Wen et al. (2013) is adopted here for enrichment analysis. Sixth, the subnetwork G_s is extracted from the adjusted network G_{new}, where the node set of the subnetwork is *S*. Note that, G_s may be disconnected. Finally, the connected network with the largest number of nodes in G_s is selected as a disease module, denoted as G_{cs}. **Algorithm 1** presents the pseudo code of the framework of IDMCSS.

Algorithm 1. Framework of the IDMCS.

```
Input: A network $G$, the set of known disease proteins $S_0$.
Output: A disease module $G_{cs}$.
$S \leftarrow S_0$, $C \leftarrow \emptyset$;
while $max(Sig_{go}(C), Sig_{de}(C), Sig_{pa}(C)) < 0.01$ do
    $NS \leftarrow disease\_protein\_neighbor(G, S)$;
    $G_{new} \leftarrow network\_adjustment(G, S, NS)$;
    $NS \leftarrow disease\_protein\_neighbor(G_{new}, S)$;
    $(S, C) \leftarrow disease\_protein\_identification(NS, S, C)$;
end while
$G_s \leftarrow subnetwork\_selection(G_{new}, S)$;
$G_{cs} \leftarrow disease\_module\_selection(G_s)$.
```

3.2 Network Adjustment Strategy

For the network *G* = (*V*, *E*) and the disease protein set *S*, the IDMCSS starts to locally change the network structure of the original network *G* around the nodes in *S*, in order to discard several potential incorrect links and retrieve several missing links in *G*. To this end, a network adjustment strategy is developed to focus on removing several potential incorrect links associated to the nodes in *S* and adding potential missing links between a node *S* and its neighbors. **Algorithm 2** details the procedure of network adjustment strategy, which is performed as follows.

Algorithm 2. Network-adjustment (*G, S, NS*).

```
Input: A network $G = (V, E)$, a disease protein set $S$ and a set of disease proteins' neighbors $NS$.
Output: A network $G_{new} = (V, E')$.
$(CS, SS) \leftarrow cal\_connective\_semantic(NS)$;
$(SN, WN) \leftarrow select\_strong\_weak\_nodes(CS)$;
for each $p' \in SN$ do
    $S'_1 \leftarrow select\_connected\_nodes(S, p')$;
    $S'_2 \leftarrow select\_unconnected\_nodes(S, p')$;
    $\varphi_1 \leftarrow mean(SS, S'_1)$;
    for each $p_{i_e} \in S'_2$ do
        if $ss(p', p_{i_e}) > \varphi_1$ then
            $E' \leftarrow E \cup \{e_{p' p_{i_e}}\}$;
        end if
    end for
end for
for each $p'' \in WN$ do
    $S''_1 \leftarrow select\_unconnected\_nodes(S, p'')$;
    $S''_2 \leftarrow select\_connected\_nodes(S, p'')$;
    $\varphi_2 \leftarrow mean(SS, S''_1)$;
    for each $p_{j_e} \in S''_2$ do
        if $ss(p'', p_{j_e}) < \varphi_2$ then
            $E' \leftarrow E / \{e_{p'' p_{j_e}}\}$;
        end if
    end for
end for
```

First, we calculate both the connective similarity and the semantic similarity between each protein in *NS* and the diseases proteins in *S* = (p_1, . . ., p_n). For a node *b* ∈ *NS*, it is supposed that the node *b* has the degree *k* and connects to k_s nodes in *S*. The connective similarity between node *b* and the nodes in *S* is calculated by a hypergeometric test as **Eq. 1**., which represents how closely protein *b* connects to disease proteins in *S* (Susan Dina et al., 2015).

$$cs(b, S) = 1 - \sum_{t=k_s}^{k} \frac{C_n^t C_{N-n}^{k-t}}{C_N^k}, \tag{1}$$

where *n* is the number of nodes in *S*, and *N* is the number of nodes in *G*.

Then, we can calculate the semantic similarity between protein *b* and disease proteins *S*. Assume that the set $\mathbf{T} = \{t_i | i = 1, \ldots, \mathcal{M}\}$ consists of all of the terms annotating *N* proteins in network *G*.

$$ss(b, S) = \sum_{i=1}^{n} \frac{\sum_{t_i \in (A_b \cap A_{pi})} I(t_i)}{I_{max}(S)}, \tag{2}$$

where $A_b = \{t_{x_k} | k = 1, \ldots, m\}$ and $A_{p_i} = \{t_{y_j} | j = 1, \ldots, m'\}$ are the sets of terms used to annotate the proteins *b* and p_i, and t_* represents a term in **T**. $I(t_i) = -log\,[pro\,(t_i)]$ is the information of the term t_i, where $pro\,(t_i)$ denotes the probability of the presence of the term t_i and its descendants in the term set **T**. The information of protein *p* is $I(p) = \sum_{k=1}^{m} I(t_{x_k})$. $I_{max}(S) = max[I\,(p_1), \ldots, I\,(p_n)]$ denotes the largest value of the information of proteins in *S*.

Second, the strong-linked nodes (SN) and the weak-linked nodes (WN) are selected from *NS*, denoting proteins in *NS* closely and weakly related with disease proteins in *S*, where a strong-linked node is defined as the protein having a connective similarity with *S* larger than 0.99, and a weak-linked node is defined as the protein when it has a connective similarity with *S* smaller than the average value in *NS*. Note that, the connective similarity ranges from 0 to 1, and the average value of connective similarity is always smaller than 0.99. Thus, there is no intersection between the strong-linked nodes (SN) and the weak-linked nodes (WN). Third, the network *G* is changed to G_{new} by adding or removing several links associated with the strong-linked or weak-linked nodes, according to the suggested network adjustment strategy. The network adjustment strategy includes two key operators, i.e., adding and removing links, which are designed as follows.

1) Adding link operator: For a strong-linked node $p' \in SN$, we check whether a link needs to be added between p' and the node in *S* which is not connected with p' in the current network. Let $S'_1 = \{p_{i_1}, \ldots, p_{i_r}\} \subseteq S$ and $S'_2 = S/S'_1$ be the two sets of nodes which are connected and not connected to node p'. For each node $p_{i_e} \in S'_2$, a link between node p' and node p_{i_e} is added into the current network when $ss(p', p_{i_e}) > \varphi_1$. This means that a link is added if the semantic similarity $ss(p', p_{i_e})$ between p' and node p_{i_e} is larger than φ_1, where φ_1 is the mean semantic similarity between p' and each node in S'_1.

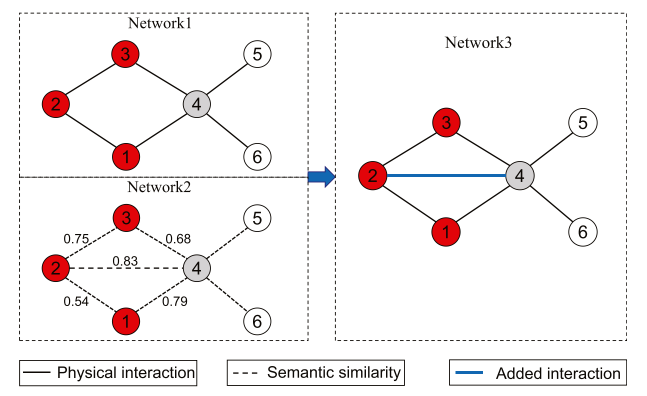

FIGURE 1 | An illustrative example of the suggested adding link operator. Network 1 is the original network, where the red nodes denote disease proteins **1**, **2** and **3** in *S*, and the gray node **4** represents a neighbor of nodes in *S*, and node **4** is a strong-linked node; Network 2 represents the semantic similarity network, where the marked edge weights are the semantic similarity. Network 3 represents the adjusted network, where a link is added between nodes **2** and **4**.

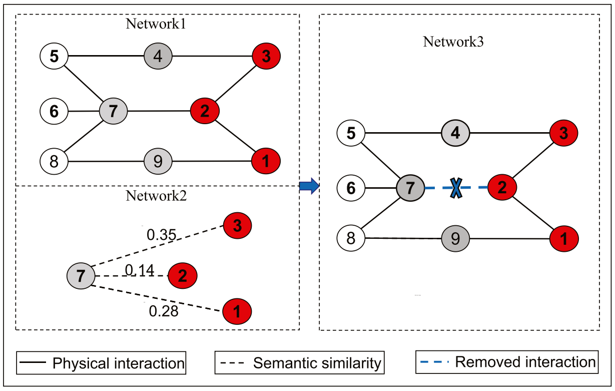

FIGURE 2 | An illustrative example of the suggested removing link operator. Network 1 is the original network, where the red nodes denote disease proteins **1**, **2** and **3** in *S*, and the gray nodes **4**, **7**, and **9** represent the neighbors of nodes in *S*, i.e., *NS* = (4, 7, 9); Node **7** is a weak-linked node in *NS*. Network 2 represents the values of the semantic similarity between **7** and each node in *S*. Network 3 represents the adjusted network, where the link between nodes **7** and **2** is removed.

Figure 1 presents an example to show how the suggested adding link operator works. As shown in this figure, the set of disease proteins *S* contains three nodes **1**, **2** and **3**, and *NS* = (**4**). For node **4**, $S_1' = \{\mathbf{1}, \mathbf{3}\}$ includes the nodes in *S* which are connected with **4**, and $S_2' = \{\mathbf{2}\}$ contains node **2** which is not connected with node **4**. Node **4** is a strong-linked node in *NS*, since the connective similarity between node **4** and *S* = (1, 2, 3) is 0.9964 according to **Eq. 1**, which is larger than the threshold 0.99. Further, the link between node **2** and node **4** is added, since the semantic similarity between them is 0.83 which is larger than the threshold $\varphi_1 = \frac{0.68+0.79}{2}$.

2) Removing link operator: For a weak-linked node $p'' \in WN$, the network adjustment strategy checks whether some links deserve to be removed to ensure that the weak-linked node p'' is not connected to any node in *S*. Let $S_1'' = \{p_{j_1}, \ldots, p_{j_s}\} \subseteq S$ and $S_2'' = S/S_1''$ be the two sets of nodes which are not connected and connected to node p''. For each node $p_{j_e} \in S_2''$, a link between p'' and p_{j_e} is removed when the semantic similarity between p'' and p_{j_e} is smaller than φ_2, where φ_2 denotes the mean semantic similarity between node p'' and each node in S_1''.

Figure 2 presents an illustrative example of the removing link operator. In this example, *S* = (1, 2, 3) represents the set of disease proteins and *NS* = (4, 7, 9) consists of all neighbors of nodes in *S*. For node **7**, there are two nodes 1 and 3 which are not connected with it ($S_1'' = \{\mathbf{1}, \mathbf{3}\}$), and one node 2 which is connected with it ($S_2'' = \{\mathbf{2}\}$). By simple calculation, we can obtain that the connective similarity between node **7** and set *S* is 0.9964 and the average connective similarity of the nodes in *NS* is 0.9984. Since the connective similarity is smaller than the average value, the node **7** is weak-linked. Hence, we need to remove the link between nodes **7** and **2** from the network, due to the fact that the threshold $\frac{0.35+0.28}{2}$ is larger than the semantic similarity between nodes **7** and **2** in S_2'' (i.e., 0.14).

3.3 The Similarity Between a Protein and Disease Proteins

In the IDMCSS, the protein having the largest similarity with the nodes in *S* is selected as a disease protein, where the similarity is measured based on both connective similarity and semantic similarity. Specifically, considering a protein *p* and a set of disease proteins *S* = (p_1, . . ., p_t), the similarity between the protein *p* and the set of disease proteins *S*, denoted as *sv* (*p*, *S*), is the normalization of the sum of the connective similarity and the semantic similarity, which is defined as **Eq. 3**.

$$sv(p, S) = \frac{cs(p, S) + ss(p, S)}{2}, \tag{3}$$

where *cs* (*p*, *S*) represents the connective similarity between *p* and *S*, and *ss* (*p*, *S*) represents the semantic similarity between *p* and *S*.

3.4 Complexity Analysis

Here, an upper bound of the time complexity of the IDMCSS is presented. As described above, the main complexity of IDMCSS lies in the following five steps: 1) the identification of *NS*, 2) the network adjustment, 3) the selection of disease protein, 4) extracting the subnetwork G_s from the adjusted network, 5) selecting a disease module G_{cs}. Note that, the first three steps are in a while loop.

The complexity for the identification of *NS* is O ($d_{max} \times n$), where $|S| = n$, the largest degree of nodes in *S* is d_{max}. Suppose the number of nodes in *NS* is n', a complexity of O ($4 \times n' + n'^2$) is needed for the network adjustment, since the complexity for calculating connective and semantic similarity as well as selecting strong and weak nodes is O ($4 \times n'$), and the maximum complexity for adding and removing links is O (n'^2). The maximum complexity for the selection of disease protein is O (n'). The first three steps holds a time complexity of O ($d_{max} \times n + n'^2$), since O ($d_{max} \times n + n'^2$) ≈ O ($d_{max} \times n + 4 \times n' + n'^2 + n'$). After the iteration of *maxgen* times, it needs a complexity of O

$((d_{max} \times n + n'^2) \times maxgen)$ for identifying the disease proteins. The fourth step needs a time complexity of $O(M)$ to extract the subnetwork G_s from the adjusted network G_{new}, where M is the number of links in G_{new}. Finally, it holds a time complexity of $O(M')$ to select a disease module, where M' is the number of links in G_s. Therefore, the IDMCSS holds a computational complexity of $O(d_{max} \times n \times maxgen + n'^2 \times maxgen + M)$, since $O((d_{max} \times n + n'^2) \times maxgen + M + M') \approx O(d_{max} \times n \times maxgen + n'^2 \times maxgen + M)$.

4 EXPERIMENTAL RESULTS

In this section, we first analyze the module of asthma obtained by the proposed IDMCSS, and then compare the performance of the IDMCSS with that of four existing algorithms for disease module detection.

4.1 Datasets

The IDMCSS performs the detection of asthma-related modules based on the protein-protein interaction network. The stopping criterion of the algorithm is set according to the information of gene ontology, differential expression genes and pathways which are related to the asthma. Specifically, the protein-protein interactions, microarray expression data, asthma-related genes and pathways are presented as follows.

First, the protein-protein interaction network is obtained by considering seven kinds of physical interactions simultaneously, which yields a network having 13, 460 proteins and 141, 296 physical interactions. The seven physical interactions considered here are regulatory interactions (Matys et al., 2003), biophysical interactions Aranda et al. (2009), Ceol et al. (2007), literature curated interactions Prasad et al. (2009), metabolic enzyme-coupled interactions Lee et al. (2008), protein complexes Ruepp and et al. (2010), kinase network Hornbeck and et al. (2012) and signaling interactions Vinayagam and et al. (2011) in human interactome. From the gene ontology annotation database (GOA) Huntley and et al. (2015), we extract 19, 707 genes annotated with GO terms and hence the obtained network consists of 12, 562 proteins and 130, 390 physical interactions.

Next, we adopt nine asthma-related microarray expression data sets consisting of the gene expression values for the differential expression analysis. The nine data sets are GSE470, GSE2125, GSE3004, GSE4302, GSE16032, GSE31773, GSE35571, GSE41649 and GSE43696, which can be available from the NCBI Gene Expression Omnibus database (GEO)[1]. It is worth noting that we use 107 known asthma-related genes in the protein-protein interaction network for experimental analysis in this paper, which are compiled from pervious literature Vercelli (2008) and several datasets[2]. In addition, 23 asthma-related pathways collected from the literature (Song and Lee, 2013; Sharma et al., 2012) are used in this paper (**Supplementary Appendix S1**).

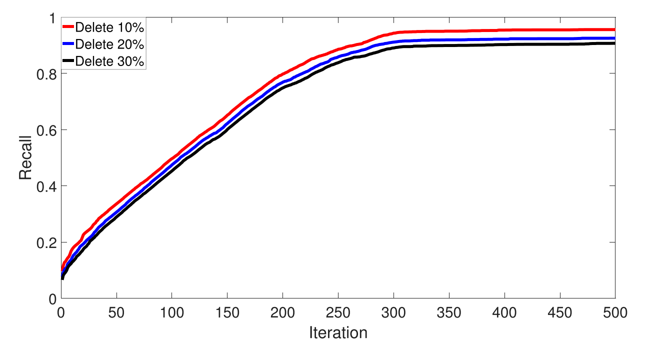

FIGURE 3 | The recall rate of disease module.

4.2 Identification of Disease Modules

We use the IDMCSS to identify disease modules based on an adjusted network, where the final disease module of asthma is achieved by running the proposed IDMCSS 217 iterations. The reason for the iterations for 217 times is that "differential expression genes" is not significantly enriched in current disease proteins earlier than "GO annotation information" and "pathway information", and the enrichment of the differential expression genes included in the disease proteins is smaller than 0.05 when the algorithm iterates 218 times.

For the disease module of asthma obtained by the suggested IDMCSS, it consists of 279 nodes and 2,819 links. Among the 279 nodes, 62 nodes are known asthma-related proteins and the other 217 nodes are newly discovered relating to asthma-related proteins. In the 2,819 links found in the disease module, 489 links are newly added and 19 links are removed from the original network by the proposed IDMCSS. It is worth noting that some known disease proteins associated with asthma are not included in the obtained disease module of asthma and hence they may be included in other connected subgraphs.

Finally, we take a close look at the closeness of the obtained disease module. We here use the ratio of the number of inner-links to that of external-links as the closeness of the disease module. The module has 2,819 inner-links and 47,657 external-links, and thus the closeness of the disease module is 0.0592. This confirms that the disease module is not a locally dense community as stated by Susan Dina et al. (2015). It can also be found that the obtained disease module has statistically larger closeness than the subnetworks randomly selected from the adjusted protein-protein interaction network according to the Student's t-test.

4.3 Asthma-Related Pathways and Genes in the Disease Module

In this subsection, we analyze the asthma-related pathways and genes in the disease module. To this end, from 304 human pathways in the Biocarta database given in **Supplementary Appendix S2**, we extract the 72 candidate pathways which has at least half of genes in the disease module obtained by the algorithm. It can be found that the 72 pathways are possible asthma-related pathways as shown in **Supplementary Appendix**

[1] http://www.ncbi.nlm.nih.gov/geo/

[2] http://gene2mesh.ncibi.org

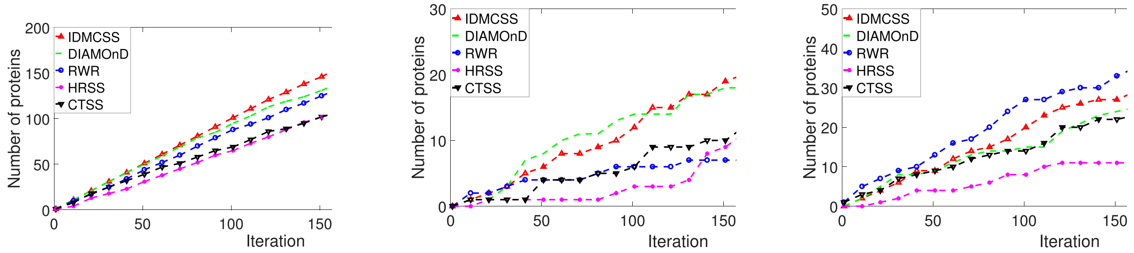

FIGURE 4 | Performance of IDMCSS, CTSS, DIAMOnD, RWR and HRSS on the asthma dataset. GO annotations: the number of asthma-related GO annotations enriched in the disease module; DifferExpre: the number of differential expressed genes in the disease module; Pathways: the number of asthma-related pathways enriched in the disease module.

S3, since they are statistically significantly enriched in the disease module. Among the 72 pathways, two are included in the 23 known asthma-related pathways and the rest 70 are the newly asthma-related pathways predicted by the algorithm. For the 70 pathways, five pathways, "h-il7Pathway", "h-pkcPathway", "h-melanocytepathway", "h-ngfPathway", and "h-trkaPathway", are considered to be associated with asthma in previous literature (Kelly and et al., 2009; Hou and et al., 2017; Raap et al., 2003; Abram, 2008).

Next, we will predict several targets of glucocorticoid based on the disease module of asthma, since they are an effective anti-inflammatory drug for asthma. The genes will be considered as the targets of glucocorticoid in asthma if they are differentially expressed between asthmatic fibroblasts untreated and asthmatic fibroblast cells treated with glucocorticoid, but not between normal untreated fibroblast cells and normal fibroblasts treated with glucocorticoid. For this reason, in this paper the 12 genes, *acvrl1*, *ar*, *cdk1*, *ctgf*, *ddit3*, *icam1*, *jak1*, *rora*, *smad1*, *snca*, *tgfb2*, and *tlr4*, are considered to be targets of glucocorticoid. To verify the effectiveness of the targets, we use the enrichment analysis of the differential expression genes before and after the treatment of glucocorticoid. For 217 expanded proteins, 23 and 17 expanded proteins are differentially expressed in normal and asthmatic samples, respectively. As for the 62 known asthma-related proteins, 10 and 8 known asthma-related proteins are differentially expressed in normal and asthmatic samples, respectively. Based on the Fisher's exact test, in normal and asthmatic samples the expanded proteins have the enrichment of differential expression genes 6.0324×10^{-4} and $2.70, \times, 10^{-3}$, and the known asthma-related proteins have the enrichment of differential expression genes 4.32×10^{-2} and $4.30, \times, 10^{-2}$. This means that the expanded proteins has significantly higher enrichment of differential expression genes than the known asthma-related proteins. Thus, we can conclude that the algorithm can provide effective targets for therapeutic intervention.

4.4 Robustness of IDMCSS

To show the robustness of IDMCSS, **Figure 3** gives the recall rate of the disease module when 10, 20, and 30% of the known asthma disease genes are randomly deleted, averaging over 30 times experiments (Warren et al., 2002). It can be found that the removal of the known disease genes has little influence on the performance of the suggested IDMCSS, and it always detect similar disease modules in the 217 iterations. Hence, we can conclude that the suggested IDMCSS shows a good robustness in detecting disease modules of asthma.

4.5 Performance Comparison

The IDMCSS is compared to four state-of-the-art disease module identification approaches, including a network structure change-based algorithm (CTSS) (Liu et al., 2015) and three traditional approaches without changing network structures (DIAMOnD Susan Dina et al. (2015), RWR Sebastian et al. (2008) and HRSS Wu et al. (2013)), where DIAMOnD and RWR are connective-based algorithms and HRSS is a semantic-based algorithm. Specifically, CTSS identifies disease genes by adding weak interactions between unconnected genes in the existing network based on the semantic similarity between them. The DIAMOnD algorithm is a seed-expanding method which identifies a disease module around a set of known disease proteins in the PPI network. RWR uses random walk analysis, which is a global network distance measure, to measure similarities among proteins in the PPI network. HRRS ranks all nodes by calculating the relative specificity similarity of each node in the network to known disease nodes, where the relative specificity similarity is calculated by taking the global position of relevant gene ontology terms into account. For the above comparison algorithms, the best parameters recommended in their original references are adopted.

Figure 4 presents performance (the number of proteins annotated by asthma-related GO terms, the number of differential expression genes, and the number of proteins in asthma-related pathways) obtained by five approaches on the asthma dataset. To be specific, the left one in **Figure 4** draws the number of proteins which are significantly annotated by 940 asthma-related GO terms for different iterations, where the 940 asthma-related GO terms are those enriched in the 107 known asthma proteins (**Supplementary Appendix S4**). From the figure, it can be found that IDMCSS achieves the largest number of proteins annotated by asthma-related GO terms.

The middle one in **Figure 4** plots the number of differential expression genes included in the disease module achieved by IDMCSS and those by four compared algorithms when the iteration ranges from 1 to 217. As can be seen from the figure, the algorithm IDMCSS gains the largest number of differential expression genes when the iteration is larger than 111. The main reason may be attributed to the fact that by enhancing the structure of PPI, it becomes relatively easy to detect the differential expression genes, thus the IDMCSS can achieve a competitive performance in detecting disease modules. The right one in **Figure 4** presents the number of proteins which belong to the 23 known asthma-related pathways. It is found that the IDMCSS is slightly worse than RWR, but it is better than other algorithms. The main reason for the phenomenon is that the proteins linked by physical interactions tend to collaborate with each other in the same pathway (Venkatesan et al., 2008. The proteins obtained by RWR are always the known disease proteins' neighbors which are connected to the known disease proteins by physical interactions in the PPI network, while those obtained by IDMCSS may be the nodes which are not linked with the known disease proteins. Therefore, we can conclude that the IDMCSS is a competitive disease module detection algorithm in terms of detection quality.

5 CONCLUSION AND FUTURE WORK

[3]In this paper, we have developed a disease module identification method IDMCSS by modifying the existing PPI networks. In the suggested IDMCSS, some potential interactions are added in the existing PPI network and some incorrect interactions are removed based on the connective and semantic similarities between the given disease proteins and their neighboring proteins. The basic idea of modifying the existing PPI network is that the incorrect links and the missing links are in the original PPI network, and we want to eliminate interference of the incorrect links and missing links for detecting disease module. However, due to the lack of the knowledge about the accurate protein-protein interactions, it is hard to analyze the validity of the modified PPI network, which may be verified in the future. The protein having the best connective and semantic similarities in the neighborhood of known disease proteins is extended into the set of disease proteins on the adjusted PPI network step by step until a stopping criterion is reached. Further, the connected subgraphs which include the disease proteins, as well as the interactions between them, are extracted from the adjusted network. Finally, the connected subgraph which contains the largest number of disease proteins is selected as a disease module.

We have performed a series of experiments on a particular disease, i.e., asthma to show the effectiveness of the IDMCSS. First, the disease module detected by the IDMCSS was not a dense community which is in accordance with traditional discovery, and it was also significantly different from the random subgraphs. Then, several pathways and genes discovered in the disease module have been verified to be related to asthma. Further, IDMCSS has little sensitivity to the number of known disease proteins. Finally, IDMCSS was superior to state-of-the-art approaches for disease module identification, since the disease module achieved by IDMCSS includes more proteins which are enriched in asthma-related GO terms, pathways and differential expression genes than those achieved by other approaches. From the above, the experiments have extensively demonstrated the superiority of IDMCSS in disease module identification.

In this work, we have locally adjusted the network structure by the suggested network adjustment strategy to deal with the PPI network which suffers from both high false positive and false negative rates. The IDMCSS performs based on the assumption that the detection results will become better if the PPI network becomes more perfect. Future attention can be given to combing connective information with other kinds of information, such as pathway information and phenotypic similarity information, to further improve the IDMCSS.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: In this paper, we adopt nine asthma-related microarray expression data sets consisting of the gene expression 252 values for the differential expression analysis. The nine data sets are GSE470, GSE2125, GSE3004, GSE4302, 253 GSE16032, GSE31773, GSE35571, GSE41649 and GSE43696, which can be available from the NCBI Gene Expression Omnibus database (GEO) http://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

JL: Software, Original draft preparation HZ: Data process, Experiments JQ: Methodology, Investigation, Reviewing.

FUNDING

The National Natural Science Foundation of China (Grant No. U1804262, 61822301, 61976001, and 61876184), the Ministry of Science and Technology of China Key Project of Science and Technology Innovation 2030 (Grant No. 2018AAA0101302 and 2018AAA0100105), and the Anhui Provincial Key Research and Development Plan(Grand No. 202004j07020005), the Fundamental Research Funds for the Central Universities (CUC210B001). Key Program of Natural Science Project of educational Commission of Anhui Province (KJ2019A0029), the Natural Science Foundation of Anhui Province(2008085QF294, 1908085MF218).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.726596/full#supplementary-material

[3] This paper is an extended version of a paper of our published in the 14th International Conference on Bio-inspired Computing: Theories and Applications (BIC-TA 2019).

REFERENCES

Abram, M. (2008). Ngf increases cell viability of isolated plasma cells from inflamed airways via trka signalling in a mouse model of allergic asthma. *J. Allergy Clin. Immunol.* 121, S200. doi:10.1016/j.jaci.2007.12.745

Agrawal, M., Zitnik, M., and Leskovec, J. (2017). Large-scale analysis of disease pathways in the human interactome. *Pac. Symp. Biocomputing* 23, 111–122. doi:10.1142/9789813235533_0011

Albert-László, B., Natali, G., and Joseph, L. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., et al. (2009). The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 38, D525–D531. doi:10.1093/nar/gkp878

Bragina, E. Y., Tiys, E. S., Rudko, A. A., Ivanisenko, V. A., and Freidin, M. B. (2016). Novel tuberculosis susceptibility candidate genes revealed by the reconstruction and analysis of associative networks. *Infect. Genet. Evol.* 46, 118–123. doi:10.1016/j.meegid.2016.10.030

Cagney, G., Uetz, P., and Fields, S. (2000). [1] High-throughput screening for protein-protein interactions using two-hybrid assay. *Methods Enzymol.* 328, 3–14. doi:10.1016/s0076-6879(00)28386-9

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., et al. (2007). Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res.* 35, 572–574. doi:10.1093/nar/gkl961

Cheng, F., Lu, W., Liu, C., Fang, J., Hou, Y., Handy, D. E., et al. (2019). A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat. Commun.* 10, 3476. doi:10.1038/s41467-019-10744-6

Cho, Y., and Montanez, G. (2013). Predicting false positives of protein-protein interaction data by semantic similarity measures. *Curr. Bioinformatics* 8, 339–346.

Csaba, O., and Mauno, V. (2009). Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.* 37, 622–628.

Cui, Y., Cai, M., and Stanley, H. E. (2018). Discovering disease-associated genes in weighted protein-protein interaction networks. *Physica A: Stat. Mech. its Appl.* 496, 53–61. doi:10.1016/j.physa.2017.12.080

del Sol, A., Balling, R., Hood, L., and Galas, D. (2010). Diseases as network perturbations. *Curr. Opin. Biotechnol.* 21, 566–571. doi:10.1016/j.copbio.2010.07.010

Franke, L., Bakel, H. V., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006a). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* 78, 1011–1025. doi:10.1086/504300

Franke, L., Bakel, H. v., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006b). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* 78, 1011–1025. doi:10.1086/504300

Härtner, F., Andrade-Navarro, M. A., and Alanis-Lobato, G. (2018). Geometric characterisation of disease modules. *Appl. Netw. Sci.* 3, 10. doi:10.1007/s41109-018-0066-3

Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., et al. (2012). Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 40, D261–D270. doi:10.1093/nar/gkr1122

Hou, L., Zhu, L., Zhang, M., Zhang, X., Zhang, G., Liu, Z., et al. (2017). Participation of antidiuretic hormone (adh) in asthma exacerbations induced by psychological stress via pka/pkc signal pathway in airway-related vagal preganglionic neurons (avpns). *Cell Physiol Biochem.* 41, 2230–2241. doi:10.1159/000475638

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., et al. (2015). The goa database: gene ontology annotation updates for 2015. *Nucleic Acids Res.* 43, D1057–D1063. doi:10.1093/nar/gku1113

Igor, F., Andrey, R., and Dennis, V. (2008). Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. United States America* 105, 4323–4328.

Jonsson, P. F., and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297. doi:10.1093/bioinformatics/btl390

Kelly, E. A. B., Koziol-White, C. J., Clay, K. J., Liu, L. Y., Bates, M. E., Bertics, P. J., et al. (2009). Potential contribution of il-7 to allergen-induced eosinophilic airway inflammation in asthma. *J. Immunol.* 182, 1404–1410. doi:10.4049/jimmunol.182.3.1404

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Res.* 37, D767–D772. doi:10.1093/nar/gkn892

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci.* 101, 15148–15153. doi:10.1073/pnas.0404315101

Lee, D. S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N., and Barabasi, A. L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci.* 105, 9880–9885. doi:10.1073/pnas.0802208105

Liu, B., Jin, M., and Zeng, P. (2015). Prioritization of candidate disease genes by combining topological similarity and semantic similarity. *J. Biomed. Inform.* 57, 1–5. doi:10.1016/j.jbi.2015.07.005

Liu, T.-Y., Liu, Z.-P., Zhao, X.-M., and Chen, L. (2011). Future Work. *J. Am. Med. Inform. Assoc.* 19, 241–248. doi:10.1007/978-3-642-14267-3_20

Luo, J., and Liang, S. (2015). Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data. *J. Biomed. Inform.* 53, 229–236. doi:10.1016/j.jbi.2014.11.004

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., et al. (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 374–378. doi:10.1093/nar/gkg108

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

Navlakha, S., and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26, 1057–1063. doi:10.1093/bioinformatics/btq076

Ortutay, C., and Vihinen, M. (2008). Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.* 37, 622–628. doi:10.1093/nar/gkn982

Raap, U., Brzoska, T., Sohl, S., Päth, G., Emmel, J., Herz, U., et al. (2003). α-Melanocyte-Stimulating Hormone Inhibits Allergic Airway Inflammation. *J. Immunol.* 171, 353–359. doi:10.4049/jimmunol.171.1.353

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., et al. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res.* 38, D497–D501. doi:10.1093/nar/gkp914

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi:10.1038/nature08454

Sebastian, K., Sebastian, B., Denise, H., and Peter N, R. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958.

Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., et al. (2012). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* 46, 957–961.

Song, G. G., and Lee, Y. H. (2013). Pathway analysis of genome-wide association study on asthma. *Hum. Immunol.* 74, 256–260. doi:10.1016/j.humimm.2012.11.003

Su, Y., Li, S., Zheng, C., and Zhang, X. (2019). A heuristic algorithm for identifying molecular signatures in cancer. *IEEE Trans. Nanobioscience* 19, 132–141. doi:10.1109/TNB.2019.2930647

Su, Y., Liu, C., Niu, Y., Cheng, F., and Zhang, X. (2021). A community structure enhancement-based community detection algorithm for complex networks. *IEEE Trans. Syst. Man. Cybern, Syst.* 51, 2833–2846. doi:10.1109/tsmc.2019.2917215

Su, Y., Zhu, H., Zhang, L., and Zhang, X. (2020). "Identifying disease modules based on connectivity and semantic similarities," in *Proceedings of 14th International Conference on Bio-inspired Computing: Theories and Applications*, 1–8. doi:10.1007/978-981-15-3415-7_3

Susan Dina, G., Jorg, M., and Albert-Laszlo, B. (2015). A disease module detection algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *Plos Comput. Biol.* 11, e1004120.

Tian, Y., Su, X., Su, Y., and Zhang, X. (2020). EMODMI: A multi-objective optimization based method to identify disease modules. *IEEE Trans. Emerging Top. Comput. Intelligence* 1, 13. doi:10.1209/TETCI.2020.3325117

Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., and Sun, F. (2006). Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* 7, 31–13. doi:10.1186/1471-2164-7-31

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., et al. (2008). An empirical framework for binary interactome mapping. *Nat. Methods* 6, 83–90. doi:10.1038/nmeth.1280

Vercelli, D. (2008). Discovering susceptibility genes for asthma and allergy. *Nat. Rev. Immunol.* 8, 169–182. doi:10.1038/nri2257

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signaling* 4, rs8. doi:10.1126/scisignal.2001699

Wang, R.-S., and Loscalzo, J. (2018). Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *J. Mol. Biol.* 430, 2939–2950. doi:10.1016/j.jmb.2018.05.016

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* 10, 280–293. doi:10.1093/bfgp/elr024

Warren, R. M. L., Pointon, L., Caines, R., Hayes, C., Thompson, D., and Leach, M. O. (2002). What is the recall rate of breast mri when used for screening asymptomatic women at high risk? *Magn. Reson. Imaging* 20, 557–565. doi:10.1016/s0730-725x(02)00535-0

Wen, Z., Liu, Z. P., Liu, Z., Zhang, Y., and Chen, L. (2013). An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. *J. Am. Med. Inform. Assoc.* 20, 659–667. doi:10.1136/amiajnl-2012-001168

Wu, X., Pang, E., Lin, K., and Pei, Z.-M. (2013). Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and ic-based hybrid method. *Plos One* 8, e66745. doi:10.1371/journal.pone.0066745

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 22, 2800–2805. doi:10.1093/bioinformatics/btl467

Zanzoni, A., Soler-López, M., and Aloy, P. (2009). A network medicine approach to human disease. *Febs Lett.* 583, 1759–1765. doi:10.1016/j.febslet.2009.03.001

Zhang, T., Wang, X., and Yue, Z. (2017a). Identification of candidate genes related to pancreatic cancer based on analysis of gene co-expression and protein-protein interaction network. *Oncotarget* 8, 71105–71116. doi:10.18632/oncotarget.20537

Zhang, X., Wang, C., Su, Y., Pan, L., and Zhang, H.-F. (2017b). A fast overlapping community detection algorithm based on weak cliques for large-scale networks. *IEEE Trans. Comput. Soc. Syst.* 4, 218–230. doi:10.1109/tcss.2017.2749282

Zheng, C.-H., Huang, D.-S., Kong, X.-Z., and Zhao, X.-M. (2008). Gene expression data classification using consensus independent component analysis. *Genomics, Proteomics & Bioinformatics* 6, 74–82. doi:10.1016/s1672-0229(08)60022-4

Zheng, C.-H., Huang, D.-S., and Shang, L. (2006). Feature selection in independent component subspace for microarray data classification. *Neurocomputing* 69, 2407–2410. doi:10.1016/j.neucom.2006.02.006

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.