Front. Comput. Sci., 2022, 16(1): 161901
https://doi.org/10.1007/s11704-021-0476-8

RESEARCH ARTICLE

A framework combines supervised learning and dense subgraphs discovery to predict protein complexes

Suyu MEI (✉)

Software College, Shenyang Normal University, Shenyang 110034, China

Abstract Rapidly identifying protein complexes is significant to elucidate the mechanisms of macromolecular interactions and to further investigate the overlapping clinical manifestations of diseases. To date, existing computational methods majorly focus on developing unsupervised graph clustering algorithms, sometimes in combination with prior biological insights, to detect protein complexes from protein-protein interaction (PPI) networks. However, the outputs of these methods are potentially structural or functional modules within PPI networks. These modules do not necessarily correspond to the actual protein complexes that are formed via spatiotemporal aggregation of subunits. In this study, we propose a computational framework that combines supervised learning and dense subgraphs discovery to predict protein complexes. The proposed framework consists of two steps. The first step reconstructs genome-scale protein co-complex networks via training a supervised learning model of l_2-regularized logistic regression on experimentally derived co-complexed protein pairs; and the second step infers hierarchical and balanced clusters as complexes from the co-complex networks via effective but computationally intensive k-clique graph clustering method or efficient maximum modularity clustering (MMC) algorithm. Empirical studies of cross validation and independent test show that both steps achieve encouraging performance. The proposed framework is fundamentally novel and excels over existing methods in that the complexes inferred from protein co-complex networks are more biologically relevant than those inferred from PPI networks, providing a new avenue for identifying novel protein complexes.

Keywords protein complexes, protein co-complex networks, machine learning, L_2-regularized logistic regression, graph clustering

1 Introduction

Protein complexes, as a fundamental macromolecular organization of multiple subunit proteins, provide insights into understanding how individual gene products form the structures required for advanced biological activities [1]. Once some subunits within multi-protein complexes malfunction, patients would develop a single symptom or overlapping clinical manifestations of diseases [2]. Systematically investigating protein complexes helps to elucidate the cellular mechanisms underlying various disorders. Tandem Affinity Purification with mass spectrometry (TAP-MS), as a well-established experimental technique, has been successfully employed to detect Yeast [3] and human [4,5] protein complexes. Besides experimental methods, recent years have witnessed intensive investigations of computational modeling for complexes identification as reviewed in the comprehensive surveys [6–8]. From data point of view, existing methods could be classified into protein-protein interaction (PPI) networks-based methods and TAP-MS based methods. From methodological point of view, the PPI-based methods largely belong to unsupervised graph clustering, while the TAP-MS based methods often combine supervised learning with graph clustering. In essence, the TAP-MS based methods still identify complexes from PPI networks, the difference is that these methods use supervised learning to learn the edge weights of PPI networks from TAP-MS data. The former category far outnumbers the latter one [6–8], partly because development of novel graph clustering methods on PPI networks seems to be more attractive to bioinformatics researchers. Graph clustering especially dense subgraphs discovery could be mathematically viewed as discovery of cliques in a graph and finding all maximal cliques has been proven to be a challenging NP-complete problem [9]. To achieve optimal structural modularity, sophisticated techniques are highly needed to achieve good balances between cluster size, the number of clusters, cluster hierarchy, cluster overlap and algorithmic complexity.

The PPI networks-based methods detect functionally or structurally cohesive substructures in the form of cliques in PPI networks and treat the inferred clusters as protein complexes. A portion of densely connected regions in PPI networks discovered via graph clustering have been validated to surely correspond to the experimentally verified protein complexes [6–8], indicating the effectiveness of identifying complexes from PPI networks. These graph clustering methods focus on topologically partitioning PPI networks into clusters, e.g., molecular complex detection (MCODE) [10], markov clustering (MCL) [11], ClusterONE [12], COREPEEL [13], DAPG [14], etc. MCODE [10] prioritizes the vertex with the highest clustering

Received September 24, 2020; accepted March 9, 2021

E-mail: meisygle@gmail.com

density as a seed and recursively moves outward to generate a cluster. MCL [11] simulates random walks via expansion and inflation operators to extract dense regions from a graph. In each iteration, the two operators make the flow thicker in dense regions and thinner in sparse regions, so that the probability of intra-cluster walks is increased and the probability of inter-cluster walks is decreased. ClusterONE [12] defines a metric of cohesiveness score to guide a greedy growth process, so that the vertexes with high cohesiveness scores in PPI networks are clustered together. COREPEEL [13] uses a tight upper bound to guide the core decomposition for discovery of quasi cliques and then peels out the vertexes with minimum degree. DAPG [14] uses node ordering algorithms to turn PPI networks into a directed acyclic prefix graph (DAPG), based on which to detect the maximal dense subgraphs via objective function optimization.

The above-mentioned graph clustering methods usually treat PPI networks as unweighted graphs. To measure the confidence level of each interaction in PPI networks, several methods have been proposed to take advantage of biological insights to weigh the edges in PPI networks, e.g., gene ontology, gene expression profiles and TAP-MS data [15]; socio-affinity scores measuring the frequencies that two proteins are observed to co-occur in CoIP purifications [16], etc. Besides weighted PPI networks, bipartite graph has also been used to represent the co-immunoprecipitation data to identify protein complexes [17]. These PPI network-based methods, though appealing and intuitive in themselves, suffer from two major aspects of drawbacks. The first major drawback is that the modules identified from PPI networks are potentially less biological relevant, and the structural or functional modules do not necessarily correspond to spatiotemporally formed protein complexes. The second major drawback lies in the complexity of graph clustering on PPI networks. First, the dense and sparse regions in PPI networks are prone to yield highly unbalanced clusters in terms of cluster size; Second, the small complexes in sparse regions are prone to be ignored with a high probability; Third, it is challenging for graph clustering methods to recover the hierarchical organizations between super-complexes and sub-complexes; Fourth, the complexes yielded via most graph clustering algorithms are usually not overlapped; and lastly, the experimentally verified complexes and subunits are not easy to be explored as prior to guide the generation of biologically interpretable clusters. TAP-MS is a well-established experimental technique that naturally captures the co-complex relationships between proteins, but to our knowledge the available TAP-MS data are only used as auxiliary data to assign confidence scores to the edges of PPI networks to date. For instances, Krogan et al. [18] train a supervised learning model on TAP-MS data to assign weights to the edges of PPI networks, from which to further infer complexes via Markov Clustering (MCL) [11]. Wu et al. [15] quantify protein affinities from TAP-MS data via so-called C2S scores, which are defined as the log-likelihood ratio of probability that a protein pair is co-complexed relative to the probability that the protein pair is drawn randomly; and then train a linear ranking SVM to integrate heterogeneous data to assign scores to the edges of PPI networks, from which to identify protein complexes via hierarchical clustering. Qi et al. [19] extract local graph features from the complexes captured via TAP-MS technique to train a supervised Bayesian network and then use the learned features or patterns to search for novel complexes in PPI networks. In this study, we propose a framework that combines supervised learning and dense subgraphs discovery to predict human protein complexes. The critical difference between this framework and existing methods is that protein complexes are identified from protein co-complex networks rather than PPI networks. Protein co-complex networks are defined as a graph , where each node in the vertex set V denotes a protein and each edge in the edge set E denotes the relationship that two proteins are co-complexed. In this framework, the process flow consists of two steps. The first step reconstructs genome-scale protein co-complex networks via exploring the experimentally verified co-complexed protein pairs to train a predictive model of l_2-regularized logistic regression. To validate the effectiveness of predicting protein co-complex relationships, we compare this framework with existing methods on the experimental co-complexed proteins pairs of Saccharomyces cerevisiae via cross validation and independent test. The second step identifies protein complexes via hierarchical graph clustering on the predicted protein co-complex networks. The clusters detected from protein co-complex networks are more biological relevant and are easier to interpret as complexes than those detected from PPI networks. To cope with the complexity of graph clustering, we evaluate a variety of available graph clustering methods on the experimental complexes from CORUM [4] and HPRD [5] and choose the optimum method that achieves a good balance between performance and efficiency.

2 Materials and methods

2.1 Data and materials

To our knowledge, the database CORUM (version 2.0) [4] and HPRD (as of November 2008) [5] have curated a number of human protein complexes, and Reactome (version 54) [20] has collected a large number of co-complexed protein pairs from the study [21]. CORUM [4] and HPRD [5] provide a full list of subunits for each protein complexes, while Reactome [20,21] only contains pair-wise protein co-complex interactions without the information of subunit-complexes membership. The first step of this framework predicts protein co-complex interactions to reconstruct genome-scale protein co-complex networks, and thus uses the co-complexed protein pairs from Reactome [20,21] as training data. The second step of this framework splits protein co-complex networks inferred by the first step into complexes, and thus uses the subunit-complexes memberships from CORUM [4] and HPRD [5] as independent test data. Of course, the complexes from CORUM [4] and HPRD [5] could be binarized into co-complexed protein pairs as training data, but Reactome [20,21] does not provide subunit-complexes memberships and thus cannot be used as independent test data to evaluate the second-step clustering performance of this framework.

The quality of training data is of pivotal importance to machine learning modeling of biological problems. Choosing training data heavily depends on particular problems or applications. In this study, we follow three inclusion criteria to choose data. First, the proteins that lack gene symbols are removed, because the proteins may be hypothetical and are not helpful for biological interpretation of the results; Second, the genes or gene products that are obsolete or uncurated in UniprotKB are removed, because the GOA database annotates genes/proteins based on UniprotKB. Obsolete genes/proteins would result in null feature vectors (see the next section *Multi-instance GO feature construction*); and lastly, the less-studied genes/proteins that have not been annotated in GOA database are also removed for similar reasons. As such, we obtain 61,755 co-complexed protein pairs from Reactome [20,21] as the positive training data (see Online Resource S1). However, there are no available protein pairs that are observed not to be co-complexed and we need negative data to train a two-class l_2-regularized logistic regression model. For the reason, we have to randomly sample the negative training data from the huge space of protein pairs. To reduce the risk of false negative sampling, we resort to the available human physical PPI networks to increase the probability of sampling protein pairs that are actually not co-complexed. We restrict the sampling to the three cases (i) protein pairs with no path between them in human physical PPI networks; (ii) protein pairs whose shortest path is more than one; (iii) protein pairs whose shortest path is just one. As the chances that two proteins are co-complexed in the three cases decrease, we empirically determine the sampling ratio of these three cases to be 6:3:1. The first case could most probably reduce the risk of coincidence of sampled protein pairs with the known PPIs. Nevertheless, only consideration of this case would yield bias, because two physically-interacting (path length equals to one) and indirectly-interacting (path length equals to two or more) proteins still are potentially not co-complexed. The last two cases are considered to cover these protein pairs. The ratio could be treated as a hyperparameter and is empirically determined here for simplicity. The human physical PPI networks guiding the negative data sampling are constructed from the physical PPIs from HPRD [5], BioGrid (as of September 2014) [22], and IntAct (as of November 2013) [23]. The sampled negative training data are provided in Online Resource S2.

As regards the independent test data, we obtain 2157 and 1502 non-redundant protein complexes from CORUM [4] and HPRD [5], respectively. After filtering out the complexes whose co-complexed protein pairs already occur in the training data, we finally obtain 1757 and 1375 non-redundant complexes from CORUM [4] and HPRD [5], respectively. These complexes are used as the positive independent test data to evaluate both steps of the proposed framework. The first step uses the co-complexed protein pairs from CORUM [4] and HPRD [5] for co-complex networks evaluation; and the second step uses the complexes and their member subunits from CORUM [4] and HPRD [5] for complexes recovery evaluation. For co-complex networks evaluation, the complexes of independent test from CORUM [4] and HPRD [5] are binarized into co-complexed protein pairs. For a given complex that consist of N subunits, $C_N^2 = N * (N-1)/2$ co-complexed protein pairs are obtained. In such a way, we obtain 37, 228 and 23,973 co-complexed protein pairs as positive independent test data from CORUM [4] and HPRD [5], respectively (see Online Resource S3 and S4). To control the risk of model bias, we sample the equivalent size of negative independent test set in the same way that the negative training data are sampled. This step is to evaluate the performance of predicting novel co-complexed protein pairs. For the complexes recovery evaluation of the second step, the complexes and their member subunits are used evaluate the performance of recovering actual complexes from the predicted co-complex networks via a variety of graph clustering methods.

Existing methods are generally developed for *Saccharomyces cerevisiae*. For methodological comparisons, we also evaluate the proposed framework on the complexes of *Saccharomyces cerevisiae*. We use the 9070 co-complexed protein pairs of high quality provided by Collins et al. [24] as positive training data and resort to the physical PPI networks of *Saccharomyces cerevisiae* to guide the sampling of negative training data. The physical PPI networks of *Saccharomyces cerevisiae* are taken from the studies [25–27] The 408 complexes provided by Pu et al. [28] (also called CYC2008 Complexes) and the 482 complexes provided by Gavin et al. [16] are used as the positive independent test data.

2.2 Supervised learning for genome-scale reconstruction of protein co-complex networks

Multi-instance GO feature construction. The first step of this proposed framework is to train a supervised learning model to predict whether two proteins are co-complexed. Recent studies [29–33] have shown that gene ontology (GO) terms are the most discriminative features to represent protein pairs and predict protein-protein interactions. Our previous work [31–33] has proposed a multi-instance feature representation of proteins to cope with the sparsity and potential unavailability of GO terms, especially for the less-studied genes/proteins. Each gene or gene product is depicted with two instances. The target instance depicts the GO knowledge of the gene/protein itself, and the homolog instance depicts the GO knowledge of the homologs. The homolog instance serves the purpose of enriching or substituting the target instance of less-studied genes/proteins. Such the representation method has been successfully applied to predict pathogen-host protein interactions [31–33]. In this study, we apply the representation method to the problem of co-complex prediction. The homologs are searched via PSI-BLast [34] against SwissProt (downloaded as of April 2012) [35] and the GO terms are retrieved from GOA (downloaded as of December 2017) [36]. For each protein i in the training set U, i.e., $i \in U$, we obtain a homolog set of GO terms (S_H^i) and a target set of GO terms (S_T^i). The whole set of GO terms of the training set is defined as follows.

$$S = \bigcup_{i \in U} (S_T^i \cup S_H^i). \tag{1}$$

For each protein pair (i_1, i_2), the target instance and the homolog instance are formally defined as follows.

$$Vec_T^{(i_1,i_2)}[g] = \begin{cases} 0, & g \notin S_T^{i_1} \wedge g \notin S_T^{i_2}, \\ 2, & g \in S_T^{i_1} \wedge g \in S_T^{i_2}, \\ 1, & \text{otherwise}, \end{cases}$$

$$Vec_H^{(i_1,i_2)}[g] = \begin{cases} 0, & g \notin S_H^{i_1} \wedge g \notin S_H^{i_2}, \\ 2, & g \in S_H^{i_1} \wedge g \in S_H^{i_2}, \\ 1, & \text{otherwise}, \end{cases} \tag{2}$$

For GO term $g \in S$, $Vec_T^{(i_1,i_2)}[g]$ and $Vec_H^{(i_1,i_2)}[g]$ denote the component g of the target instance and the homolog instance, respectively. The GO terms that are not contained in the whole set of GO terms ($g \notin S$) are discarded. The two-instance feature representation as described in Eq. (2) is symmetric and could intuitively depict the distribution of GO terms between two proteins (i_1, i_2).

L_2-regularized logistic regression. L_2-regularized logistic regression [37] is a well-established machine learning method that could efficiently deal with large data with linear time complexity and penalize noise to prevent model overfitting via regularization technique. We have applied its toolbox LIBLINEAR [38] to counteract potential noise from homolog instances in our previous work [31–33]. In this study, we also choose l_2-regularized logistic regression as the base learner to predict protein co-complex relationships.

Given the training data (x_i, y_i), $i = 1, 2, \ldots, l; x_i \in R^n; y_i \in \{-1, +1\}$ where x_i denotes the ith instance and y_i denotes its label, l_2-regularized logistic regression transforms the prime logistic regression hypothesis $h(x) = 1/(1 + \exp(-y\omega^T x))$ to the dual optimization problem as follows.

$$\min_{\omega} \frac{1}{2}\omega^T \omega + \zeta \sum_{i=1}^{l} \log(1 + e^{-y,\omega^T x_i}), \tag{3}$$

where ω denotes weight vector, ζ denotes regularizer or slack variable. The second term adopts regularization technique to penalize noise and prevent overfitting. The prime problem as defined in Eq. (3) is solved via its dual form as follows.

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}} O\alpha + \sum_{i:\alpha_i>0}^{l} \alpha_i \log \alpha_i + \\ &\sum_{i:\alpha_i<C} \zeta - \alpha_i) \log(\xi - \alpha_i) - \sum_{i}^{l} \zeta \log \zeta, \\ &\text{subject to } 0 \leqslant \alpha_i \leqslant \zeta, i = 1, \ldots, l, \end{aligned} \tag{4}$$

where α_i denotes Lagrangian operator and $O_{ij} = y_i y_j x_i^{\mathrm{T}} x_j$.

For each test protein pair (i_1, i_2), we combine the target-instance and homolog-instance outputs ($h(Vec_T^{(i_1,i_2)})$, $h(Vec_H^{(i_1,i_2)})$) into one final decision value as defined below.

$$Decision\_value(i_1, i_2) = \begin{cases} h(Vec_T^{(i_1,i_2)}), & \text{if } |h(Vec_T^{(i_1,i_2)})| > |h(Vec_H^{(i_1,i_2)})|, \\ h(Vec_H^{(i_1,i_2)}), & \text{otherwise,} \end{cases} \tag{5}$$

Then the predicted label for the test protein pair (i_1, i_2) is determined as follows.

$$L(i_1, i_2) = \begin{cases} 1, & Decision\_value(i_1, i_2) > 0 \wedge decision\_value(i_1, i_2) - 0.5 > \delta, \\ -1, & Decision\_value(i_1, i_2) < 0 \wedge decision\_value(i_1, i_2) + 0.5 > \delta, \\ \infty & \text{otherwise,} \end{cases} \tag{6}$$

where threshold δ is introduced to increase the confidence level of predictions, generally assuming 0 if weak positive predictions still provide valuable information. The choice of thresholdcould affect the predicted protein co-complex networks and yield different dense subgraphs or protein complexes. Symbol ∞ means that the labels cannot be determined.

2.3 Graph clustering for complexes identification

In protein co-complex networks, any two subunits within the same complex potentially have a connective link or edge, while the edges between complexes are much sparser. Identification of complexes from protein co-complex networks is a problem of dense subgraphs discovery. In this study, we use the classical k-clique finding method CFinder [39,40] to identify dense subgraphs in protein co-complex networks. CFinder [39] uses a clustering algorithm with exponential computational complexity to find all k-cliques (i.e., maximal complete subgraphs), from which to further analyse the clique-clique overlap matrix. However, finding all maximal cliques is a NP-complete problem [9], so that CFinder could not be applicable anywhere. For this reason, we adopt the maximum modularity clustering method (MMC) [41,42] as an alternative solution in the case that CFinder [39] cannot yield desirable complexes. The MMC method [41,42] includes a coarsening step and a refining step. The coarsening step iteratively merges clusters and the refining step refines the resulting clusters by iteratively moving individual vertices between clusters according to a criterion called modularity increase. In order for readers to grasp the core idea of this method, we choose to briefly reformulate the MMC graph clustering method below.

Given a graph (V,f) that consists of vertex set V and a function $f : V \times V \rightarrow N$, the function f assigns a weight to each edge connecting two vertexes within V and the degree of vertex v is defined as $\deg(v) = \sum_{u \in V} f(u, v)$. The degree of a set of vertices is generalized as $\deg(V) = f(V, V) = \sum_{u \in V, v \in V} f(u, v)$. Assuming that the vertex set V is partitioned into k non-empty subsets C_i by a clustering $C = \{C_1, \ldots, C_k\}$, the weight of edge $(u, v) \in V^2$, i.e., $f(u, v)$, is binomially distributed with the expected value $\deg(u)\deg(v)/\deg(V)^2$ in the null model that the end-vertices of edges are chosen at random, and then the expected value of the whole edge set within cluster C_i is generalized as $\deg(C_i)^2/\deg(V)^2$ [43]. Then the modularity of clustering $C = \{C_1, \ldots, C_k\}$ is defined as follows.

$$Q(C) := \sum_{C_i \in C} \left(\frac{f(C_i, C_i)}{f(V, V)} - \frac{\deg(C_i)^2}{\deg(V)^2} \right), \tag{7}$$

where the first term is the actual fraction of intra-cluster edge weight and the second term is its expected value in the null model. Then the modularity increase of coarsening by merging cluster C_i with C_j is defined as follows.

$$\Delta Q_{C_i, C_j} := \frac{2f(C_i, C_j)}{f(V, V)} - \frac{2\deg(C_i)\deg(C_j)}{\deg(V)^2}. \tag{8}$$

Accordingly, the modularity increase of refining by moving a vertex v from its cluster C_i to cluster C_j is defined as follows.

$$\Delta Q_{v \to C_j} := \frac{2(f(v, C_j) - f(v, C_i - v))}{f(V, V)} - \frac{2(\deg(v)\deg(C_j) - \deg(v)\deg(C_i - v))}{\deg(V)^2}. \quad (9)$$

The coarsening as defined in Eq. (8) and the refining as defined in Eq. (9) iterate greedily until the maximum modularity is achieved without modularity increase (i.e., $\Delta Q_{v \to C_j} < 0$).

2.4 Experimental settings and model evaluation

Experimental settings

For the first step of this framework, three experimental settings (namely combine-instance, homolog-instance and target-instance) are provided to check how well the homolog knowledge transfer via homolog instance solves the problem of GO term sparsity. For model evaluation, the combined decision value as defined in Eq. (5), the decision value of the homolog instance alone (i.e., $h(Vec_H^{i_1,i_2})$ and the decision value of the target instance alone (i.e., $h(Vec_T^{i_1,i_2})$ are used in the combined-instance setting, the homolog-instance setting and the target-instance setting, respectively.

For the first step of this framework, the experimentally verified protein complexes from CORUM [4] and HPRD [5] are binarized into co-complexed protein pairs to check whether CFinder [39,40] and MMC [41,42] could correctly recover these complexes from the predicted protein co-complex networks.

Model evaluation

For the first step of protein co-complex networks reconstruction via supervised l_2-regularized logistic regression, the performance is estimated using the five frequently-used metrics including PR (precision), SE (sensitivity), MCC (Matthews correlation coefficient), ROC-AUC (Area Under ROC Curve) and F1 score. Except that ROC-AUC is calculated based on the decision values as defined in Eq. (5), all the other metrics are calculated based on a confusion matrix M, whose element $M_{i,j}$ records the counts that class i ($i = 1, 2, \ldots, L$) are classified to class j ($j = 1, 2, \ldots, L$). In this framework, L assumes 2 for the binary classification of protein co-complex associations. To calculate PR, SE and MCC, several intermediate variables as defined in Eq. (10) are first calculated based on M. Based on these variables, we further calculate PRl, SEl and MCCl for each label according to Eq. (11) and the overall MCC according to Eq. (12).

$$\begin{aligned} p_l &= M_{l,l}, q_l = \sum_{i=1, i \neq l}^{L} \sum_{j=1, j \neq l}^{L} M_{i,j}, \\ r_l &= \sum_{i=1, i \neq l}^{L} M_{i,l}, s_l = \sum_{j=1, j \neq l}^{L} M_{l,j}, \\ p &= \sum_{l=1}^{L} p_l, q = \sum_{l=1}^{L} q_l, r = \sum_{l=1}^{L} r_l, s = \sum_{l=1}^{L} s_l, \end{aligned} \quad (10)$$

$$\begin{aligned} PR_l &= \frac{p_l}{p_l + r_l}, l = 1, 2, \ldots, L, \\ SE_l &= \frac{p_l}{p_l + s_l}, l = 1, 2, \ldots, L, \\ MCC_l &= \frac{(p_l q_l - r_l s_l)}{(p_l + r_l)(p_l + s_l)(q_l + r_l)(q_l + s_l)}, l = 1, 2, \ldots, L, \end{aligned} \quad (11)$$

$$Acc = \frac{\sum_{l=1}^{L} M_{l,l}}{\sum_{i=1}^{L} \sum_{j=1}^{L} M_{i,j}}, \quad MCC = \frac{(pq - rs)}{(p + r)(p + s)(q + r)(q + s)}, \quad (12)$$

F1 score is further calculated as follows.

$$F1\ score = \frac{2 \times PR_l \times SE_l}{PR_l + SE_l}, l = 1 \text{ denotes the positive class.} \quad (13)$$

For the second step of protein complexes detection via unsupervised maximum modularity graph clustering, Jaccard Index as defined in Eq. (14) is used to measure how well the predicted complex P matches the reference complex R.

$$Jaccard(P, R) = |P \cap R| / |P \cup R|. \quad (14)$$

Given a threshold ξ, P is deemed to match R if $Jaccard(P, R) \geqslant \xi$ is satisfied (ξ generally assumes 0.5). Given a set of reference complexes $R = \{R_1, \ldots, R_l\}$ and a set of predicted clusters $P = \{P_1, \ldots, P_m\}$, the metrics of Precision, Recall and F-score are defined as follows.

$$\begin{aligned} \text{Precision} &= \frac{|\{P_i \in P | \exists R_j \in R, \text{Jaccard}(P_i, R_j) \geqslant \xi\}|}{|P|}, \\ \text{Recall} &= \frac{|\{R_i \in R | \exists P_j \in P, \text{Jaccard}(P_i, R_j) \geqslant \xi\}|}{|R|}, \\ \text{F-score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \quad (15)$$

3 Results

3.1 Performance of cross validation and independent set on predicting protein co-complex relationships

Cross validation

As described in the section *Data and materials*, the co-complexed proteins pairs from Reactome [20,21] are used as the positive training data and the negative training data are randomly sampled under the guidance of prior knowledge of human physical PPI networks to train a two-class l_2-regularized logistic regression model. The ROC curves of 5-fold cross validation are illustrated in Fig. 1(a). It is evident that the ROC curves of the three experimental settings nearly coincide with each other with the AUC scores all above 0.91, indicating that the solution of homolog knowledge transfer via homolog instances is effective to address the problem of GO terms sparsity. The other performance metrics are provided in Table 1. As demonstrated by these metrics, the l_2-regularized logistic regression model achieves encouraging performance of predicting protein co-complexed relationships with a good balance as a whole without showing a high risk of model bias. Nevertheless, the metrics of PR, SE and MMC on the two classes still

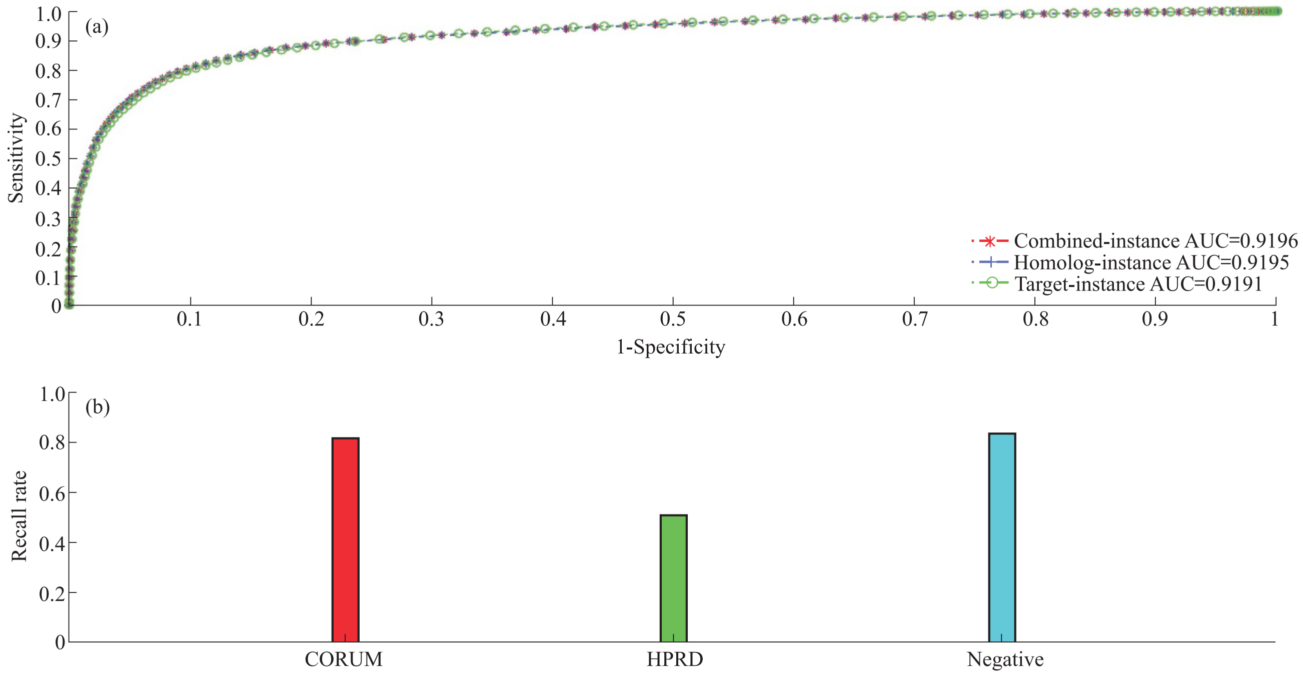

Fig. 1 Performance of cross validation and independent test on identifying human protein co-complex relationships. (a) Homo sapiens (cross validation); (b) homo sapiens (independent test)

Table 1 Results of 5-fold cross validation and independent set on identifying human protein co-complex relationships

Cross validation	Combined-instance			Homolog-instance			Target-instance		
	PR	SE	MCC	PR	SE	MCC	PR	SE	MCC
Co-complexed	0.7948	0.8967	0.7161	0.7952	0.8966	0.7161	0.8092	0.8974	0.7192
Not co-complexed	0.8813	0.7682	0.7068	0.8810	0.7683	0.7067	0.8695	0.7636	0.7030
Accuracy		83.25%			83.25%			83.42%	
MCC		0.7070			0.7071			0.7097	
ROC-AUC		0.9196			0.9195			0.9191	
F1 Score		0.8427			0.8429			0.8510	
Independent test		CORUM (positive)			HPRD (positive)			Negative independent data	
		81.47%			50.15%			83.45%	

Note: the performance metric of independent test denotes recognition rate.

show a little bias towards the positive class, for instance, PR = 0.7948, SE = 0.8967, and MCC = 0.7161 on the positive class versus PR = 0.8813, SE = 0.7682, and MCC=0.7068 on the negative class in the combined-instance setting. The bias partly results from the noise of negative data sampling that potentially comes from two sources: (1) the randomly sampled protein pairs connected by at least one path in PPI networks probably yields false negatives; and (2) the protein pairs without paths connecting them are also probably co-complexed because the available physical PPI networks are not complete. In spite of the lower risk of model bias, the performance is quite encouraging as compared to existing methods (see the section *Comparisons with the existing methods*).

Independent test

As described in the section *Data and materials*, 1757 and 1375 complexes are obtained from CORUM [4] and HPRD [5] as the positive independent test data, respectively. These complexes do not contain any co-complexed protein pair that already occur in the training data. These complexes from CORUM [4] and HPRD [5] are further binarized into 37,228 and 23,973 co-complexed protein pairs, respectively. To estimate the potential risk of model bias, we randomly sample 37,228 protein pairs as the negative independent test in the same way that the negative training data are sampled. The independent test performance is illustrated in Fig. 1(b) and provided in Table 1. 81.47% of the protein co-complex relationships in CORUM [4] are correctly recognized. Meanwhile, 83.45% of the protein pairs in the negative independent test data are correctly recognized. These results show a low risk of bias between the positive and the negative class. However, only 50.15% of the protein co-complex relationships from HPRD [5] are correctly recognized, which means that there are a large fraction of missing links in the predicted protein co-complex networks. If the missing links do not affect the recovery of protein complexes via MMC graph clustering, this proposed framework could be deemed to be robust against predictive errors of protein co-complex relationships. This problem will be discussed in the next section.

3.2 Performance of complexes identification via graph clustering

Validating the feasibility of MMC graph clustering in complexes identification

Before embedding the MMC graph clustering method [41,42] into our proposed framework, we first validate its feasibility in identifying complexes from protein co-complex networks. The 2157 complexes from CORUM [4] and the 1502 complexes from HPRD [5] are individually binarized into protein co-complex networks, in which any two proteins within the same complexes are assigned a link to indicate that they are co-complexed. For the singleton complexes that contain only one subunit, a link is assigned to the orphan subunit connecting itself. As such, all the complexes are naturally the dense subgraphs in protein co-complex networks, in which the link density within complexes is high and the link density between complexes is low. We need to verify that the MMC graph clustering method [41,42] could successfully recover most complexes from the protein co-complex networks.

The performance of MMC graph clustering method on the protein co-complex networks from CORUM [4] and HPRD [5] is illustrated in Figs. 2(a) and 2(b), respectively. As perfect match between predicted complex and reference complex is hard to achieve, the threshold of Jaccard index is relaxed to vary from 0.2 to 1. With the decrease of Jaccard index threshold, we could find more predicted complexes to match the reference complexes. We focus on the perfect match case $\xi = 1$ and the case $\xi = 0.5$ that more than a half of subunits overlap between the predicted complexes and the reference complex. The details of performance at $\xi = 1$ and $\xi = 0.5$ are provided in Table 2. In the case of perfect match ($\xi = 1$), 11.71% of CORUM and 11.78% of HPRD reference complexes are exactly hit respectively as shown by the recall metrics; and 32.57% of CORUM and 16.67% of HPRD predicted clusters exactly match the reference complexes respectively as shown by the precision metrics. These results show that the proposed framework achieves fairly encouraging performance of perfect match between predicted complexes and reference complexes, though the task is very challenging. If the threshold of Jaccard indexis relaxed to 0.5, 52.34% of CORUM and 54.26% of HPRD reference complexes are hit, respectively; and 80.99% and 77.68% of the predicted clusters match CORUM and HPRD reference complexes, respectively. These results demonstrate the feasibility of MMC graph clustering [39,40] to recover the complexes from protein co-complex networks.

Further comparison of performance on CORUM with that on HPRD shows that the MMC graph clustering method [41,42] achieves equivalent recall rates on both datasets (Table 2), which indicates that an equivalent number of CORUM and HPRD reference complexes are matched by the predicted clusters. Nevertheless, MMC achieves higher precision on CORUM than on HPRD, e.g., 0.3257 versus 0.1667 ($\xi = 1$), 0.8099 versus 0.7768 ($\xi = 0.5$), indicating that more predicted complexes match the reference complexes on CORUM than on HPRD. These results are potentially attributed to the quality and the number of reference complexes from CORUM and HPRD, and are also potentially associated with the ratio of reference complexes to predicted complexes, e.g., 1502 reference complexes to 1062 predicted complexes on HPRD (ratio 1.41), 2157 reference complexes to 1173 predicted complexes on CORUM (ratio 1.84). Under these ratios, a potentially larger fraction of predicted complexes match the reference complexes on CORUM than on HPRD.

Performance of identifying CORUM and HPRD independent test complexes

In the first step that predicts protein co-complex relationships, the proposed framework correctly recognizes 81.47% of the co-complexed protein pairs on CORUM independent test

Table 2 MMC clustering performance on CORUM and HPRD co-complex networks

	Exact match ($\xi = 1$)			Match ($\xi = 0.5$)		
	Precision	Recall	F-score	Precision	Recall	F-score
CORUM	0.3257	0.1171	0.2294	0.8099	0.5234	0.6359
HPRD	0.1667	0.1178	0.1381	0.7768	0.5426	0.6389

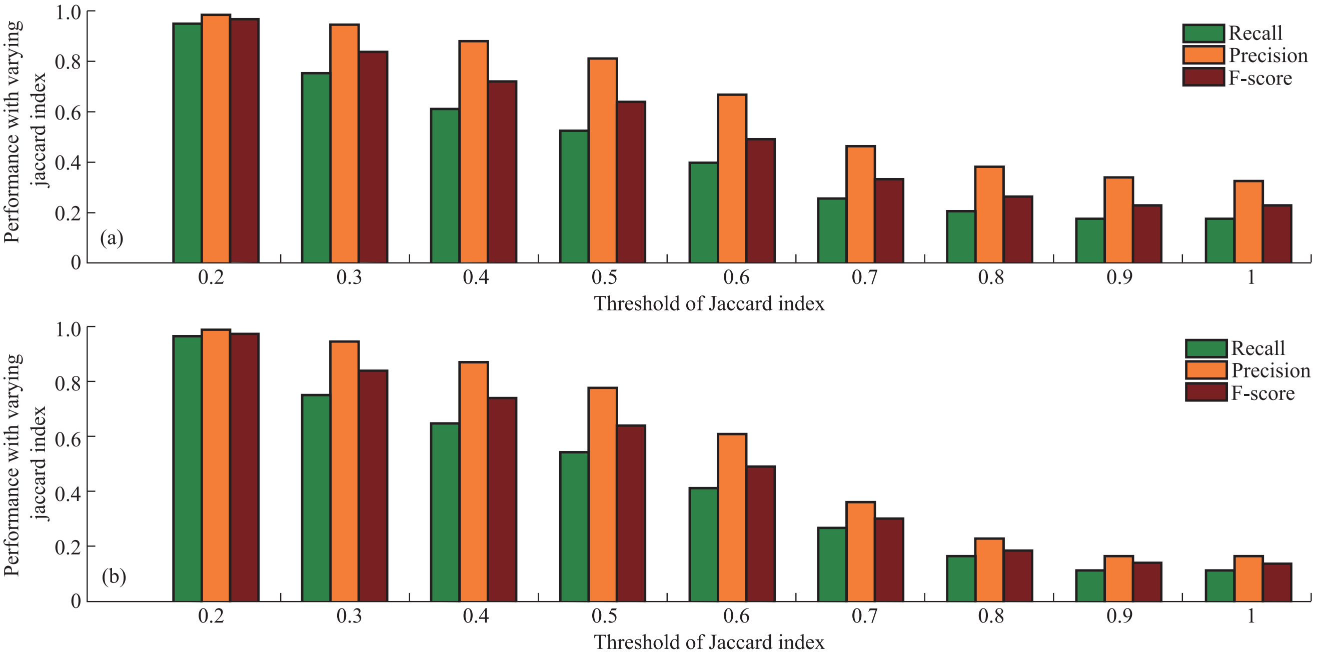

Fig. 2 Graph clustering performance that validates the feasibility of MMC method for complexes identification. (a) MMC clustering performance on CORUM co-complex networks; (b) MMC clustering performance on HPRD co-complex networks

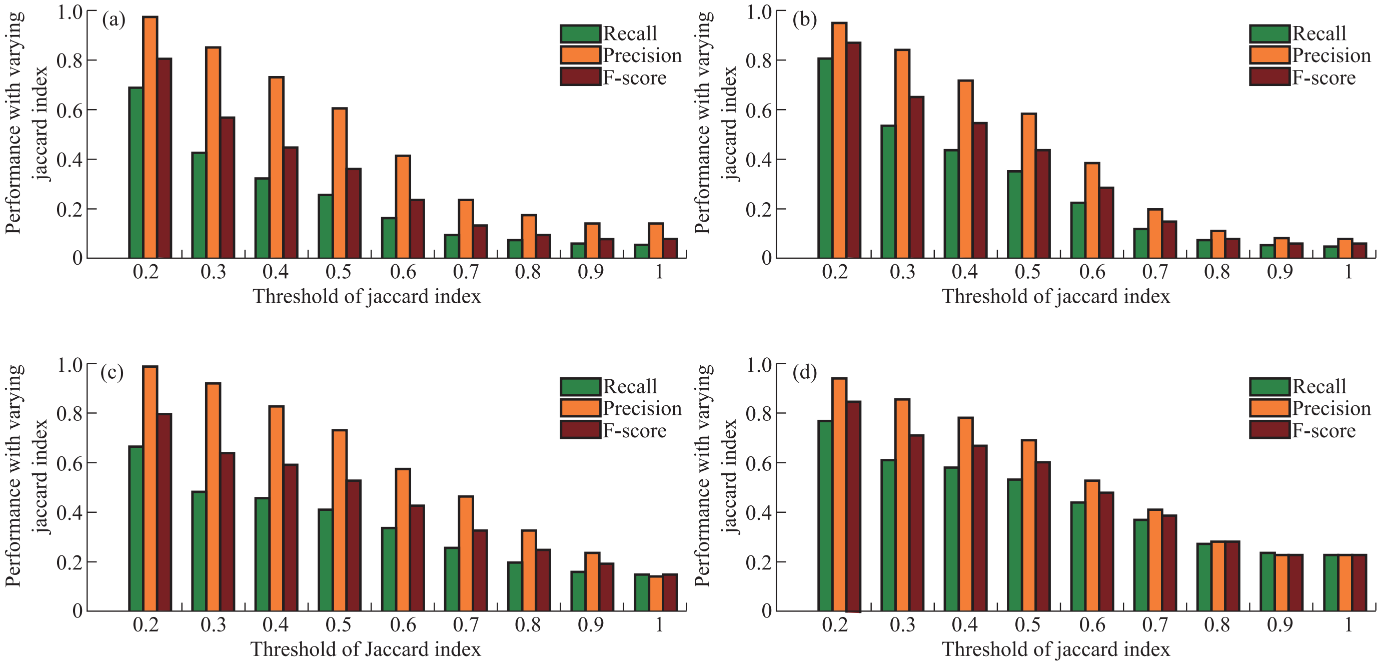

Fig. 3 Performance of identifying CORUM and HPRD independent test complexes with threshold of Jaccard indexvarying from 0.2 to 1. (a) MMC performance of identifying CORUM independent test complexes; (b) MMC performance of identifying HPRD independent test complexes; (c) CFinder performance of identifying CORUM independent test complexes; (d) CFinder performance of identifying HPRD independent test complexes

complexes and 50.15% of the co-complexed protein pairs on HPRD independent test complexes, respectively. In the second step of the proposed framework, we further estimate how well CFinder [39,40] and MMC [41,42] recovers the CORUM and HPRD independent test complexes from the predicted protein co-complex networks. The performance of independent test with the threshold of Jaccard indexvarying from 0.2 to 1 is illustrated in Figs. 3(a)–(d). Comparing Figs. 3(a) and (b) with Figs. 3(c) and (d), we find that CFinder performs much better than MMC in terms of identifying actual complexes from predicted protein co-complex networks. However, CFinder could not yield outputs on CORUM because of its inherent NP-complete characteristic, though it yields results on HPRD in a reasonable time. After setting the time limit per node at 0.1 second according to the instructions at the website of CFinder, CFinder adopts the policy of approximate clique finding and its performance on CORUM independent test complexes is illustrated in Fig. 3(c). The policy of approximate clique finding performs a little worse (see Figs. 3(c) and 3(d)), though it could reduce the computational complexity. Furthermore, the approximate policy of CFinder is still much slower than MMC in dense subgraphs discovery.

As shown in Table 3, MMC perfectly predicts 5.98% of the CORUM and 4.87% of the HPRD reference complexes, respectively (ξ = 1, Recall); and 14.46% of the CORUM and 7.94% of the HPRD predicted clusters by MMC exactly match the reference complexes, respectively (ξ = 1, Precision). If the threshold of Jaccard index is relaxed to ξ = 0.5, 26.18% of the CORUM and 34.55% of the HPRD reference complexes match the predicted complexes by MMC, respectively; and 60.33% of the CORUM and 58.41% of the HPRD predicted clusters match the reference complexes, respectively.

Table 3 Performance of identifying CORUM and HPRD independent test on protein complexes

MMC	Exact match (ξ = 1)			Match (ξ = 0.5)		
	Precision	Recall	F-score	Precision	Recall	F-score
CORUM	0.1446	0.0598	0.0846	0.6033	0.2618	0.3652
HPRD	0.0794	0.0487	0.0604	0.5841	0.3455	0.4341
CFinder	Exact match (ξ = 1)			Match (ξ = 0.5)		
	Precision	Recall	F-score	Precision	Recall	F-score
CORUM	0.1512	0.1554	0.1533	0.7324	0.4172	0.5316
HPRD	0.2340	0.2371	0.2355	0.6927	0.5360	0.6044

As shown in Table 3, CFinder comparatively performs much better than MMC. 41.72% of the CORUM and 53.60% of the HPRD reference complexes are matched by the clusters predicted by CFinder, respectively (ξ = 0.5); and 73.24% of the CORUM and 69.27% of the HPRD predicted clusters by CFinder are matched by the reference complexes, respectively (ξ = 0.5). These results encourage us to choose CFinder as the optimal solution to dense subgraphs discovery in protein co-complex networks. However, CFinder, due to its NP-complete complexity, would be highly restricted in some particular applications, in which it cannot efficiently yield outputs even when some approximate strategy of clique finding is adopted. In such cases, MMC could be used as alternative solution. In the last decades, many more novel and sophisticated graph clustering methods have been developed to find dense subgraphs, e.g., Molecular COmplex Detection (MCODE) [10], Markov Clustering (MCL) [11], ClusterONE [12], COREPEEL [13], DAPG [14], etc. In the next section, we also provide performance comparisons with these methods. Interestingly, the computational results show that CFinder and MMC are strongly robust to a large fraction of missing and false links in the predicted protein co-complex networks. In the case that only 50.15% of HPRD co-complexed protein pairs are correctly recognized, CFinder and MMC still achieve 0.5360 and 0.3455 recall performance (ξ = 0.5), respectively.

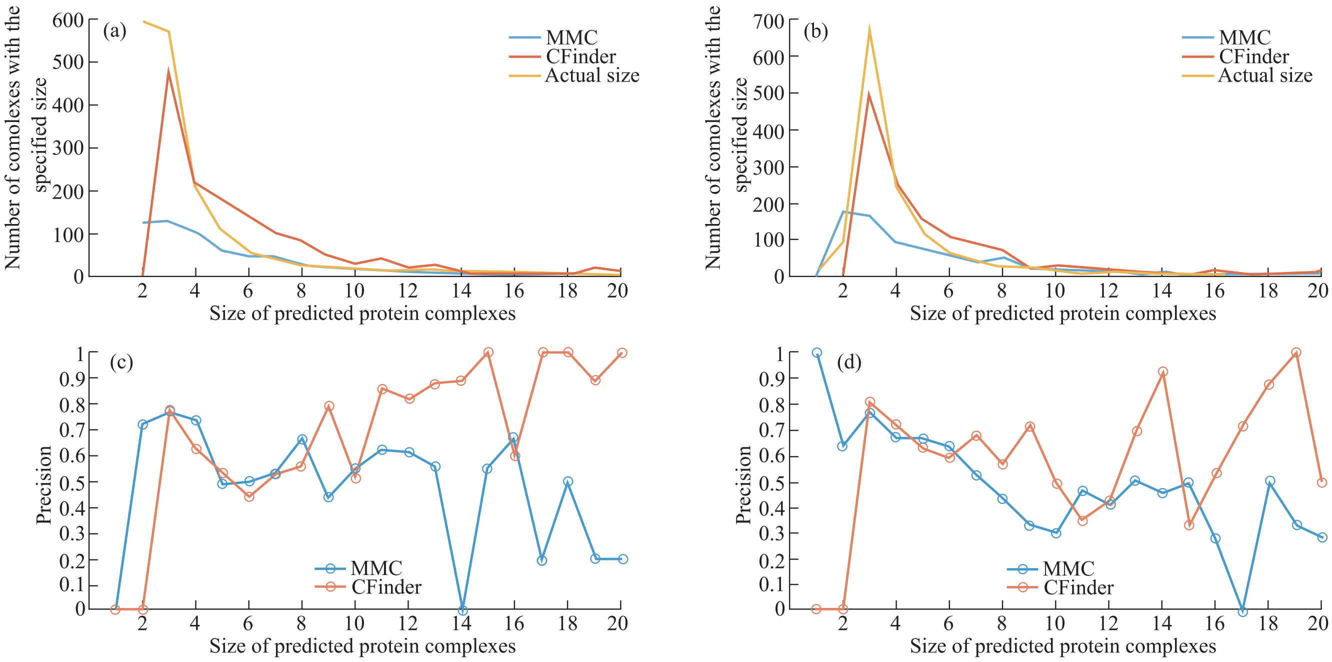

Fig. 4 Distribution of predicted complexes size on CORUM and HPRD independent test complexes and corresponding performance in terms of precision ($\xi = 0.5$). The horizontal axis denotes the size of complexes and the vertical axis denotes the number of complexes of the specified size. (a) Distribution of predicted complexes size on CORUM independent test complexes; (b) Distribution of predicted complexes size on HPRD independent test complexes; (c) Distribution varying with the size of CORUM predictod complexes; (d) Performance varying with the size of HPRD predicted compiexes

Cluster size is an important measure that evaluates a graph clustering algorithm as well as the Jaccard index-based performance metrics such as recall, precision and F-score. A good graph clustering algorithm naturally yield similar distributions between actual complexes size and predicted complexes size. The distributions of predicted complexes size that CFinder and MMC achieve on CORUM and HPRD independent test complexes are illustrated in Figs. 4(a) and (b). Most reference complexes and predicted complexes contain 2–5 subunits, and the complexes identified by CFinder are more approximate to the reference complexes in size than those identified by MMC. For the large complexes exceeding 5 subunits, MMC performs better with the size of predicted complexes more approximate to that of reference complexes, whereas CFinder tends to yield large clusters. Due to sparsity, the complexes exceeding 20 subunits are not considered. We further consider the performance in terms of precision that varies with the size of predicted protein complexes. As shown in Figs. 4(c) and (d), CFinder and MMC have a large fraction of small-sized predicted complexes (3–8 subunits) matched by the reference complexes with precision ranging from 0.4400 to 0.8056. Interestingly, CFinder achieves even higher precision on large predicted complexes, while MMC does not perform so satisfactorily. Except for the predicted complexes with 14 subunits on CORUM and 17 subunits on HPRD, the performance of MMC is acceptable.

3.3 Comparisons with existing methods

The proposed framework combines the first step of protein co-complex prediction via supervised learning and the second step of complexes identification via graph clustering into one integrated framework, whereas existing methods treat the prediction of protein co-complex relationships [44,45] and the identification of protein complexes via graph clustering [6–8] as two independent research topics. Besides this, the fundamental difference between the proposed framework and existing methods is that this framework identifies complexes from protein co-complex networks instead of protein-protein interaction (PPI) networks, which are used only to guide negative data sampling in this framework. To demonstrate the effectiveness of this proposed framework, we compare the two steps of the proposed framework with existing methods from methodological and performance points of view.

Prediction of protein co-complex networks via supervised learning

The first step of this framework is to predict protein co-complex networks via l_2-regularized logistic regression, in which each protein is represented with homolog instance and target instance to tackle the sparsity of GO terms. Existing methods exploit heterogeneous data as attributes of decision tree [44] (e.g., correlated mRNA expression, sequence homology, gene fusion, molecular function, etc.) or embed multiple features as individual kernels of kernel fusion [45] (e.g., GO terms, gene co-expression, gene co-regulation, interlogs, etc.) to predict protein co-complex relationships. These two existing methods both focus on *Saccharomyces cerevisiae*. To make methodological comparisons feasible, we rebuild the proposed framework on the available protein complexes of *Saccharomyces cerevisiae*, and we evaluate the proposed framework on the same independent test data as the two methods [44, 45] have used.

The training data and the negative independent test data of *Saccharomyces cerevisiae* are constructed as described in the section Data and materials. The ROC curves are illustrated in Fig. 5 and the other performance metrics are provided in Table 4. The first step of predicting protein co-complex networks via l_2-regularized logistic regression achieves satisfactory performance of cross validation with a low risk of bias on the two classes. The ROC curves of the three experimental settings nearly coincide with each other and the AUC scores are all above 0.92. These results again indicate that homolog knowledge transfer is effective to address the sparsity of GO terms. As illustrated in Fig. 5, the true positive rate is over 80% at 10% false discovery rate, which is equivalent to 83.9% achieved by the diffusion kernel method [45]. The method [45] does not report the performance on the positive class and the negative class individually, so that we have no knowledge about the risk of model bias.

The positive independent test data of *Saccharomyces cerevisiae* are taken from the studies [16,28]. The proposed framework achieves 78.18% recognition rate on the complexes from Gavin et al. [16], while diffusion kernel method [45] only achieves 48.5% recognition rate on the same complexes. These results demonstrate the superiority of the first step of this framework over existing methods. From methodological perspective, this framework uses much less feature information than the methods [44,45], but achieves much better performance of independent test. One reason is that the newly updated information of gene ontology potentially contributes to the performance increase; and the other reason is that the sampling of negative data guided by physical PPI networks is potentially more reliable than random sampling.

Identification of protein complexes via graph clustering

In recent years, several graph clustering methods, e.g., Markov clustering (MCL) [11], ClusterONE [12], COREPEEL [13] and DAPG [14], have been developed to identify protein complexes from protein-protein interaction networks. However, the clusters inferred by these methods are more structural or functional modules than protein complexes. Furthermore, the available PPI networks are far from complete and thus many actual complexes cannot be captured from PPI networks. This framework identifies complexes from the protein co-complex networks, which are constructed from experimental TAP-MS data and predicted co-complexed protein pairs. To demonstrate the superiority of CFinder [39,40] and MMC [41,42] over existing graph clustering methods including Markov Clustering (MCL) [11], ClusterONE [12], COREPEEL [13] and DAPG [14], we conduct performance comparisons on the protein co-complex

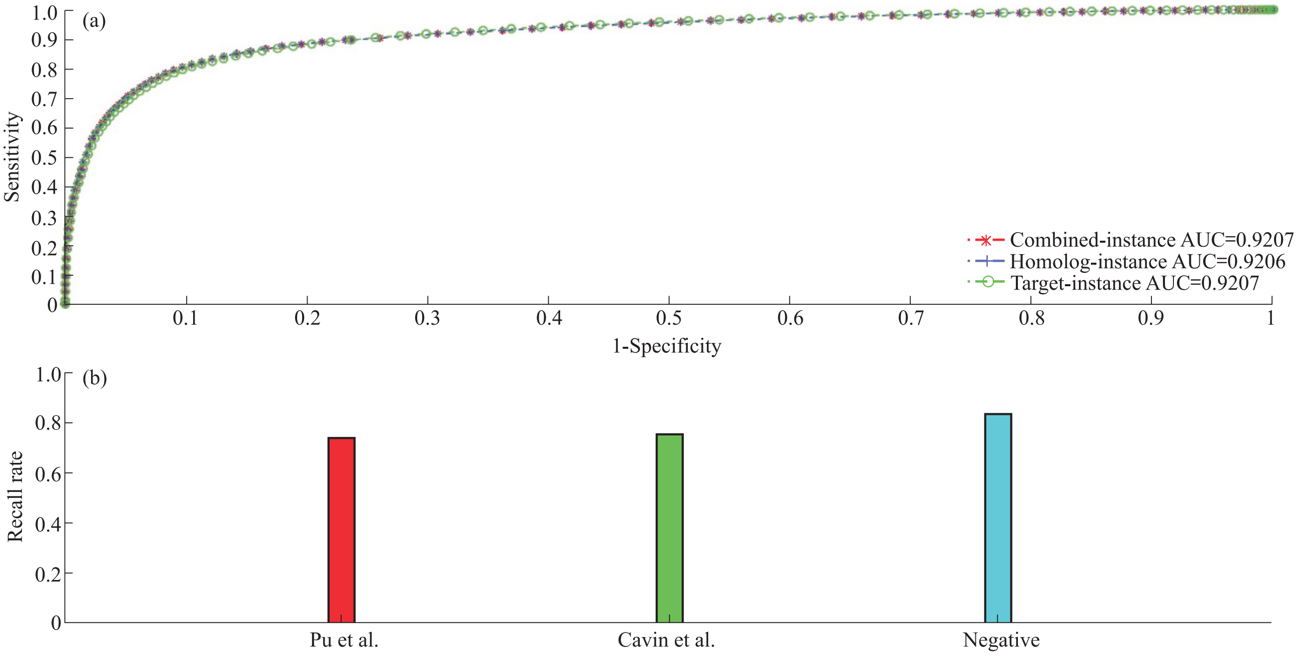

Fig. 5 Performance of cross validation and independent test on identifying *Saccharomyces cerevisiae* protein co-complex relationships. (a) Saccharomyces cerevisiae (cross validation); (b) saccharomyces cerevisiae (independent test)

Table 4 Results of 5-fold cross validation and independent set on identifying Saccharomyces cerevisiae protein co-complex relationships

Cross validation	Combined-instance			Homolog-instance			Target-instance		
	PR	SE	MCC	PR	SE	MCC	PR	SE	MCC
Co-complexed (9070 pairs)	0.8261	0.8799	0.7275	0.8293	0.8825	0.7291	0.8261	0.8799	0.7275
Not co-complexed (8500 pairs)	0.8623	0.8024	0.7187	0.8604	0.7994	0.7180	0.8623	0.8024	0.7187
Accuracy		84.24%			84.31%			84.24%	
MCC		0.7226			0.7236			0.7226	
ROC-AUC		0.9207			0.9206			0.9207	
F1 Score		0.8522			0.8551			0.8522	
Independent test		Pu et al. (positive)			Gavin et al. (positive)			Negative independent data	
		75.99%			78.18%			85.43%	

Note: the performance metric of the independent test denotes recognition rate.

networks that are predicted from 1757 CORUM and 1375 HPRD independent test complexes, respectively. MCODE [10] is not compared because it heavily depends on Yeast gene dictionary and only works for *Saccharomyces cerevisiae*.

As illustrated in Fig. 6((a)–(d)), CFinder and MMC achieve far better recall rates than the other methods on CORUM and HPRD, ranking first and second, respectively. These results show that CFinder and MMC hit more reference complexes than the other methods. However, CFinder and MMC are not so good as ClusterONE and COREPEEL from precision point of view, which indicates that ClusterONE and COREPEEL have more predicted complexes matched by the reference complexes. Nevertheless, the two methods have their advantages of precision sharply weakened by their worse recall values than CFinder and MMC. Furthermore, the better precision performance of ClusterONE and COREPEEL largely result from the small number of predicted complexes. As illustrated in Fig. 6(b) and Fig. 6(d), the number of complexes predicted by CFinder and MMC is much larger than that of ClusterONE, COREPEEL, and the other methods. Among these methods, the number of complexes predicted by CFinder and MMC is the closest to the actual number of reference complexes. ClusterONE and COREPEEL predict too small number of complexes and achieve high precision performance, but many actual complexes cannot not be captured. A graph clustering algorithm could be deemed good only if it could achieve a good trade-off between recall and precision. Contrary to ClusterONE and COREPEEL, MCL predicts the largest number of complexes on HPRD that is close to the actual number of reference complexes (see Fig. 6(d)), so that the actual complexes are easily captured to achieve good recall performance But MCL achieves the lowest precision performance on HPRD and 60.99% of the clusters predicted by MCL are unfortunately singleton clusters that contain only one protein (see Fig. 6(c)). Comparatively, MMC and CFinder seldom yield singleton clusters on CORUM and HPRD. F-score shows that MMC and CFinder perform the best among all the methods and yield a reasonable number of predicted complexes with a good trade-off between recall and precision (see Figs. 6(a) and 6(c)).

As two top-ranking methods, CFinder and MMC perform much better than the other methods, wherein CFinder performs the best. Furthermore, the number of complexes predicted by CFinder is the close to the actual number of reference complexes among these methods. However, the NP-complete complexity of CFinder restricts its practical applications. We cannot determine how much time it will take to yield outputs or whether or not it could yield output. As the second-best solution, MMC outperforms the other graph clustering methods, and could be used as an alternative solution in the case that CFinder fails to yield desirable outputs.

4 Discussion

Identifying protein complexes is significant to understand how individual proteins spatiotemporally form the structures required for biological activities. Investigation of subunits malfunction helps to understand the overlapping clinical manifestations of disease and to find potential drug targets. From computational point of view, existing methods focus on developing novel graph clustering methods to identify protein complexes from protein-protein interaction (PPI) networks. For instance, PCDq [46] comprehensively curates a large number of human protein complexes, but most complexes are predicted from human PPI networks. Short- or long-range biological signals are transmitted (possibly through complexes) to the receptors of protein or complexes via protein-protein interactions. Most PPIs only play the role of relaying signals in particular signaling pathways and do not form protein complexes, so that the clusters inferred via graph clustering on PPI networks are

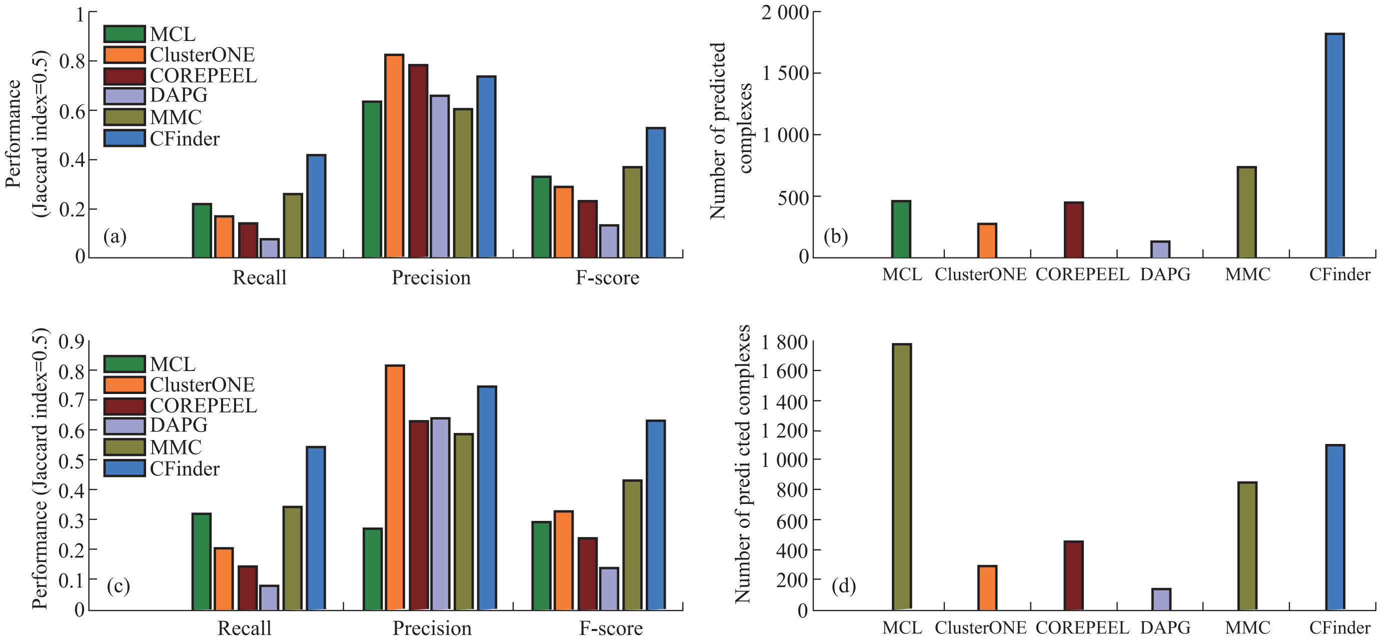

Fig. 6 Performance comparisons between CFinder, MMC, and the other graph clustering methods with $\xi = 0.5$. (a) Performance comparison on CORUM independent test complexes; (b) Number of predicted complexes on 1757 CORM reference complexes; (c) Performance comparison on HPRD independent test complexes; (d) Number of predicted complexes on 1 375 HPRD reference complexes

potentially structural or functional modules of PPI networks and the modules unnecessarily correspond to spatiotemporally formed protein complexes. These methods could to some extent capture a small number of complexes because the available physical PPI networks have already contained some complexes experimentally derived by TAP-MS technique. However, the physical PPI networks to date are by far not complete, let alone the embedded co-complex interactions, which restricts the applications of the purely PPI-based methods. For these reasons, we need to develop biologically interpretable computational methods to identify biological-relevant complexes.

In this study, we attempt to identify complexes from protein co-complex networks instead of protein-protein interaction networks. The available co-complexed protein pairs experimentally derived by TAP-MS technique are far from complete and need to be computationally augmented for complexes identification. For this purpose, we propose a framework that combines supervised learning and graph clustering to predict protein complexes. In this framework, the first step is to reconstructs genome-scale protein co-complex networks via l_2-regularized logistic regression; and the second step is to identify complexes from the co-complex networks via k-clique finding (CFinder) or maximum modularity clustering (MMC). According to the law of high link density within complexes and low link density between complexes, the genome-scale protein co-complex networks could be naturally split into a number of complexes. To critically estimate the model performance, we choose the experimentally derived co-complexed protein pairs from Reactome [20,21] as training data, and choose the experimental complexes from CORUM [4] and HPRD [5] as independent test data. The physical PPI networks herein are used to guide the sampling of negative only. It is worth noting that the first-step supervised learning of this framework aims at predicting protein-protein interactions that potentially form protein complexes (i.e., co-complex interaction) instead of general-purpose PPIs, e.g., PPI-Detect [47], PIPR [48], etc. The performance of cross validation and independent test on CORUM [4] and HPRD [5] demonstrates that the first-step supervised learning of this framework well recognizes the experimental co-complexed protein pairs with a low risk of model bias. Comparison on the previous Saccharomyces cerevisiae independent test data shows that the l_2-regularized logistic regression of this framework outperforms the diffusion kernel method [45] in terms of predicting protein co-complex relationships. Negative data sampling is critical to model performance and we take three measures to sample those protein pairs that are likely not to be co-complexed. As this study focuses on validating the assumption that protein complexes can be identified from protein co-complex networks, the sampling strategies could be chosen as an independent research topic and the performance averaging could be particularly conducted in practical application. Furthermore, the second-step CFinder or MMC graph clustering of this framework also outperforms the up-to-date graph clustering methods such as MCL, ClusterONE, COREPEEL and DAPG with a good trade-off between recall and precision (i.e., a high F-score). In addition, the number of complexes predicted by CFinder and MMC is much closer to the actual number of reference complexes. As regards the two top-ranking methods, CFinder performs much better than MMC and more approximate to the true complexes in terms of actual compositional subunits and complexes size. However, CFinder is seriously restricted in practical uses due to its NP-complete complexity.

In the case that CFinder could not yield desirable outputs in a reasonable time, MMC is a good alternative because of its efficiency. Computational results show that actual complexes can be well recovered by CFinder and MMC from protein co-complex networks with a large fraction of missing and false links, that's, the second-step graph clustering is strongly tolerant to errors in the predicted protein co-complex networks to yield complexes of high quality. If all the co-complexed protein pairs from Reactome, CORUM and HPRD are merged into the training data, the derived protein co-complex networks would be greatly enhanced with a lower risk of missing and false links, which will further improve the quality and increase the coverage of complexes. Although CFinder and MMC achieve good trade-off between recall and precision with reasonable cluster size and number of clusters, graph clustering for complexes identification still is a challenging problem, e.g., detecting small complexes in sparse regions, inferring hierarchical and overlapping associations between complexes, etc. As an independent research topic, novel graph clustering algorithms have to consider many concerns to make the predicted clusters infinitely approximate to the actual complexes. For the pure PPI-based methods, prior knowledge of TAP-MS complexes and other biological implications should be embedded into the graph clustering methods, so that PPI networks contain more information of co-complex interactions. At present, the performance of pure PPI-based methods is not satisfactory and needs to be further improved. For instance, the supervised graph local clustering method [19] achieve much lower performance (0.489 recall, 0.312 precision and 0.381 F-score) on PPI networks than CFinder (0.6927 recall, 0.5360 precision and 0.6044 F-score, see Table 3) and MMC (0.5841 recall, 0.3455 precision and 0.4341 F-score, see Table 3) on protein-complex networks.

5 Conclusion

In this study, we propose an integrative framework that combines supervised learning and dense subgraphs discovery to predict protein complexes. The first-step supervised learning adopts l_2-regularized logistic regression as base learner to predict protein co-complex relationships. The second-step graph clustering adopts k-clique finding (CFinder) or maximum modularity clustering (MMC) to identify complexes from the protein co-complex networks inferred by the first-step supervised learning. Performance comparisons show both steps of this framework outperform existing methods. Identifying complexes from protein co-complex networks instead of protein-protein interaction networks provides a new avenue for computational modeling in complexes identification.

Supporting information The supporting information is available online at journal.hep.com.cn and link.springer.com.

References

1. Krogan N J, Peng W, Cagney G, Robinson M D, Haw R, Zhong G, et al. High-definition macromolecular composition of yeast RNA-processing

complexes. Molecular Cell, 2004, 13(2): 225–239

2. Lage K, Karlberg E O, Størling Z M, Olason P I, Pedersen A G, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology, 2007, 25(3): 309–316
3. Mewes H W, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Research, 2004, 32(suppl_1): D41-D44
4. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Research, 2010, 38(suppl_4): D497–D501
5. Keshava Prasad T S, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database—2009 update. Nucleic Acids Research, 2009, 37(suppl_1): D767–D772
6. Li X, Wu M, Kwoh C K, Ng S K. Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics, 2010, 11(1): 1–19
7. Srihari S, Yong C H, Patil A, Wong L. Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes. FEBS Letters, 2015, 589(19): 2590–2602
8. Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Sepasi H, et al. Protein complex prediction: a survey. Genomics, 2020, 112(1): 174–183
9. Bron C, Kerbosch J. Finding all cliques of an undirected graph. Communications of the ACM, 1973, 16(9): 575–580
10. Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 2003, 4(1): 1–27
11. Van Dongen S. Graph clustering by flow simulation. University of Utrecht, 2000
12. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 2012, 9(5): 471–472
13. Pellegrini M, Baglioni M, Geraci F. Protein complex prediction for large protein protein interaction networks with the Core&Peel method. BMC Bioinformatics, 2016, 17(12): 37–58
14. Hernandez C, Mella C, Navarro G, Olivera-Nappa A, Araya J. Protein complex prediction via dense subgraphs and false positive analysis. PLoS ONE, 2017, 12: e0183460
15. Wu M, Xie Z, Li X, Kwoh C K, Zheng J. Identifying protein complexes from heterogeneous biological data. Proteins, 2013, 81(11): 2023–2033
16. Gavin A C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006, 440(7084): 631–636
17. Geva G, Sharan R. Identification of protein complexes from co-immunoprecipitation data. Bioinformatics, 2011, 27(1): 111–117
18. Krogan N J, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature, 2006, 440(7084): 637–643
19. Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z. Protein complex identification by supervised graph local clustering. Bioinformatics, 2008, 24(13): i250–i268
20. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Research, 2016, 44(D1): D481–D487
21. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biology, 2010, 11(5): 1–23
22. Chatr-Aryamontri A, Breitkreutz B J, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Research, 2015, 43(D1): D470–D478
23. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research, 2014, 42(D1): D358–D363
24. Collins S R, Kemmeren P, Zhao X C, Greenblatt J F, Spencer F, Holstege F C, et al. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Molecular & Cellular Proteomics, 2007, 6(3): 439–450
25. Yu H, Braun P, Yildirim M A, Lemmens I, Venkatesan K, Sahalie J, et al. High-quality binary protein interaction map of the yeast interactome network. Science, 2008, 322(5898): 104–110
26. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of The United States of America, 2001, 98(8): 4569–4574
27. Uetz P, Giot L, Cagney G, Mansfield T A, Judson R S, Knight J R, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature, 2000, 403(6770): 623–627
28. Pu S, Wong J, Turner B, Cho E, Wodak S J. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research, 2009, 37(3): 825–831
29. Maetschke S, Simonsen M, Davis M, Ragan M A. Gene ontology-driven inference of protein-protein interactions using inducers. Bioinformatics, 2012, 28(1): 69–75
30. Qi Y, Tastan O, Carbonell J G, Klein-Seetharaman J, Weston J. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics, 2010, 26(18): i645–i652
31. Mei S, Zhu H. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. Scientific Reports, 2015, 5: 8034
32. Mei S. In silico enhancing *M. tuberculosis* protein interaction networks in STRING to predict drug-resistance pathways and pharmacological risks. Journal of Proteome Research, 2018, 17(5): 1749–1760
33. Mei S, Flemington E K, Zhang K. Transferring knowledge of bacterial protein interaction networks to predict pathogen targeted human genes and immune signaling pathways: a case study on M. tuberculosis. BMC Genomics, 2018, 19(1): 1–21
34. Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 1997, 25(17): 3389–3402
35. Boeckmann B, Bairoch A, Apweiler R, Blatter M C, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, 2003, 31(1): 365–370
36. Barrell D, Dimmer E, Huntley R P, Binns D, O'Donovan C, Apweiler R, et al. The GOA database in 2009–an integrated gene ontology annotation resource. Nucleic Acids Research, 2009, 37(D1): D396–D403
37. Yu F, Huang F, Lin C. Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning, 2011, 85: 41–75
38. Fan R, Chang K, Hsieh C, Wang X, Lin C. LIBLINEAR: a library for large linear classification. Machine Learning Research, 2008, 9(2): 1871–1874
39. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814–818
40. Adamcsek B, Palla G, Farkas I J, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics, 2006, 22(8): 1021–1023
41. Noack A, Rotta R. Multi-level algorithms for modularity clustering. In: Proceedings of the 8th International Symposium on Experimental Algorithms. 2009, 257–268
42. Rossi F, Villa-Vialaneix N. Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets. Journal de la Société

Française de Statistique, 2011, 152: 34–65

43. Newman M E. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006, 74: 036104
44. Zhang L V, Wong S L, King O D, Roth F P. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics, 2004, 5(1): 1–15
45. Qiu J, Noble W S. Predicting co-complexed protein pairs from heterogeneous data. PLoS Computational Biology, 2008, 4(4): e1000054
46. Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, et al. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset. BMC Systems Biology, 2012, 6(Suppl 2): S7
47. Romero-Molina S, Ruiz-Blanco Y B, Harms M, Münch J, Sanchez-Garcia E. PPI-Detect: a support vector machine model for sequence-based prediction of protein-protein interactions. Journal of Computational Chemistry, 2019, 40(11): 1233–1242
48. Chen M, Ju C J, Zhou G, Chen X, Zhang T, Chang K W, et al. Multi-faceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics, 2019, 35(14): i305–i314

Suyu Mei received his PhD in computer science from Fudan University, China. His research fields cover machine learning and bioinformatics. He further conducted postdoctoral research of computational biology in Southern Medical University, China. His research topics focused on studying pathogen-host signaling cross-talks and systems pharmacology. He has published more than 20 first-authored papers in international peer-review journals. His current research topics cover the studies of plant and soil microbiome, microbial ecology and human microbiome-associated diseases via microbiomics and machine learning approaches.