Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 720960, 9 pages
http://dx.doi.org/10.1155/2014/720960

Research Article

A Least Square Method Based Model for Identifying Protein Complexes in Protein-Protein Interaction Network

Qiguo Dai,[1] Maozu Guo,[1] Yingjie Guo,[1] Xiaoyan Liu,[1] Yang Liu,[1] and Zhixia Teng[1,2]

[1] *School of Computer Science and Technology, Harbin Institute of Technology, P.O. Box 319, 92 Xidazhi Street, Harbin 150001, China*
[2] *School of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China*

Correspondence should be addressed to Maozu Guo; maozuguo@hit.edu.cn

Received 22 July 2014; Accepted 27 August 2014; Published 23 October 2014

Academic Editor: Yudong Cai

Protein complex formed by a group of physical interacting proteins plays a crucial role in cell activities. Great effort has been made to computationally identify protein complexes from protein-protein interaction (PPI) network. However, the accuracy of the prediction is still far from being satisfactory, because the topological structures of protein complexes in the PPI network are too complicated. This paper proposes a novel optimization framework to detect complexes from PPI network, named PLSMC. The method is on the basis of the fact that if two proteins are in a common complex, they are likely to be interacting. PLSMC employs this relation to determine complexes by a penalized least squares method. PLSMC is applied to several public yeast PPI networks, and compared with several state-of-the-art methods. The results indicate that PLSMC outperforms other methods. In particular, complexes predicted by PLSMC can match known complexes with a higher accuracy than other methods. Furthermore, the predicted complexes have high functional homogeneity.

1. Introduction

Proteins do not function in isolation but interact together to form complexes. Protein complex plays an important role in cellular activities, such as signal transduction, cell cycle, DNA transcription, and DNA repair [1–3]. Identifying protein complexes is crucial for understanding molecular mechanism in cellular activities. It is important to develop computational methods for identifying complexes [1]. Recent developments in high-throughput technologies have produced large amount of high-quality protein-protein interaction (PPI) data that can be represented as a PPI network, an undirected graph, in which nodes denote that proteins and edges are interactions between pairs of proteins. Graph clustering techniques are used to identify protein complexes by finding dense regions in a PPI network [4]. Since proteins may belong to several complexes, most of previous methods detect overlapping clusters [1, 4–6].

Many methods [7–9] detect complexes from PPI network by finding cliques, in which all nodes connect to each other. CFinder is one of the most popular clique-based methods, which searches adjacent cliques in the network [8, 10, 11]. OCG [12] takes the cliques as initial classes for hierarchy fusion to detect overlapping clusters in PPI networks. Another kind of methods detects complexes by expanding a set of seed proteins or clusters. MCODE [13] chooses the proteins with high weights as seeds and expands these seeds by including their neighboring proteins with weights higher than a threshold. ClusterONE [14], the latest and powerful seed-expansion method, starts from a set of seed complexes and expands them by maximizing the cohesiveness function. The expanding method depends on the density-based definition of the complexes. Random walking techniques have been also used to detect complexes. Markov clustering (MCL) algorithm [15] iteratively applies "expansion" and "inflation" steps to the transition matrix that denote the Markov chain of random walk. Reference [16] proposes a new spectral method based on the two-hop transition matrix of Markov random walk (SLCP2). In general, although much progress has been made, identifying protein complexes from PPI network still remains a challenge. The complexes derived by existing methods match few known complexes. The reason is that the topological structures of complexes are too complicated. It is difficult to define the topology by a specific type of pattern. It

is necessary to develop a new method to avoid the problem of topological dependence.

In this paper, we present an optimization framework that uses a penalized least squares method to identify complexes from PPI network, named PLSMC. Intuitively, our method is on the basis of the fact that if two proteins are in a common complex, they are likely to be interacting [1, 4, 5]. PLSMC employs this relation to detect complexes using a penalized least squares method. By optimization, the propensities of proteins to complexes can be determined. The PLSMC is tested and compared with other methods on several public PPI networks of yeast. The results show that PLSMC has higher accuracy on matching with known complexes than other state-of-the-art methods. Moreover, the analysis of functional homogeneity indicates that complexes identified by PLSMC are biological relevance.

2. Materials and Methods

2.1. Penalized Least Squares Method for Complex Detection. In order to introduce our method, we first introduce several notations. A PPI network is denoted by a matrix of $G_{N\times N}$, where N is the number of proteins and G_{ij} is equal to 1 if proteins i and j are interacting, 0 otherwise. Since an interaction may be a false positive one when the corresponding proteins share less common interacting partners, we compute the weight matrix S for a PPI network as in [17],

$$S_{ij} = \begin{cases} \dfrac{|N(i) \cap N(j)|}{\sqrt{|N(i)| \times |N(j)|}}, & \text{if } G_{ij} = 1, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $N(i)$ is a set consisting of protein i and all of its neighbors.

Let θ_{iz} ($\theta_{iz} > 0$) be the propensity denoting how likely protein i belongs to complex z, which is an unknown variable needing to be estimated. The cocomplex coefficient C_{ij} of proteins i and j denotes the likelihood that they participate in the same complexes. Given that there are at most K complexes existing in the PPI network, C_{ij} is calculated as

$$C_{ij} = \sum_{z=1}^{K} \theta_{iz}\theta_{jz}. \tag{2}$$

Hence, the sum of distances between interaction weights and cocomplex coefficients over all pairs of proteins can be written as follows:

$$L = \sum_{i,j}^{N} \frac{1}{2}\left(C_{ij} - S_{ij}\right)^2 = \sum_{i,j}^{N} \frac{1}{2}\left(\sum_{z=1}^{K} \theta_{iz}\theta_{jz} - S_{ij}\right)^2. \tag{3}$$

Minimizing L with respect to $\Theta = [\theta_{iz}]$ is to make the cocomplex coefficient close to interaction weight for each pair of proteins. If two proteins are not interacting, the cocomplex coefficient of them is supposed to be minimized to 0. However, only considering the cocomplex coefficient is not sufficient for complex detection, since a protein may have large number of propensities with high values. It will assign a protein to too many complexes and thus produce pervasive overlapping complexes. Therefore, to control overlapping rate, we augment (3) with a penalty term to shrink the propensities as in (4). Consider

$$L = \sum_{i,j}^{N} \frac{1}{2}\left(\sum_{z}^{K} \theta_{iz}\theta_{jz} - S_{ij}\right)^2 + \lambda \sum_{i}^{N}\sum_{z}^{K} \theta_{iz}^2, \tag{4}$$

where λ ($\lambda > 0$) is the parameter of the penalization. Finally, the optimization in PLSMC is written as

$$\begin{aligned} \min_{\Theta} \quad & L(\Theta) = \sum_{i,j}^{N} \frac{1}{2}\left(\sum_{z}^{K} \theta_{iz}\theta_{jz} - S_{ij}\right)^2 + \lambda \sum_{i}^{N}\sum_{z}^{K} \theta_{iz}^2, \\ \text{s.t.} \quad & \Theta \geq 0. \end{aligned} \tag{5}$$

2.2. Estimating Protein Propensities. Estimating the propensities $\Theta = [\theta_{iz}]$ in (5) is a nonnegative constrained optimization problem. Let $\Phi = [\phi_{iz}]$ be the Lagrange multiplier for the constraint $\Theta \geq 0$. The Lagrange function L is as

$$\begin{aligned} L(\Theta, \Phi) = & \sum_{i,j}^{N} \frac{1}{2}\left(\sum_{z}^{K} \theta_{iz}\theta_{jz} - S_{ij}\right)^2 + \lambda \sum_{i}^{N}\sum_{z}^{K} \theta_{iz}^2 \\ & + \sum_{i}^{N}\sum_{z}^{K} \phi_{iz}\theta_{iz}. \end{aligned} \tag{6}$$

Taking the derivation of (6) with respect to θ_{ik} and setting it to zero give

$$2\sum_{j}^{N} \theta_{jk} \sum_{z}^{K} \theta_{iz}\theta_{jz} - 2\sum_{j}^{N} \theta_{jk}S_{ij} + 2\lambda\theta_{ik} + \phi_{ik} = 0. \tag{7}$$

It is difficult to estimate θ_{ik} in above equation using an analytical method, as it depends on θ_{jz}, where $j \neq i$ and $z \neq k$. Therefore, we use an iterative method, to find the optimal θ_{ik}. Because $\theta_{ik}\phi_{ik} = 0$ for the Karush-Kuhn-Tucker condition, we multiply both sides of the equation by θ_{ik} and get

$$\theta_{ik}\left(\sum_{j}^{N} \theta_{jk} \sum_{z}^{K} \theta_{iz}\theta_{jz} + \lambda\theta_{ik}\right) = \theta_{ik}\sum_{j}^{N} \theta_{jk}S_{ij}. \tag{8}$$

Then, we can write the multiplicative updating rule as

$$\theta_{ik}^{\text{new}} \longleftarrow \theta_{ik} \frac{\sum_{j}^{N} \theta_{jk}S_{ij}}{\sum_{j}^{N} \theta_{jk} \sum_{Z}^{K} \theta_{iz}\theta_{jz} + \lambda\theta_{ik}}. \tag{9}$$

As suggested in the literature [18], we use the updating rule as

$$\theta_{ik}^{\text{new}} \longleftarrow \frac{\theta_{ik}}{2} + \frac{\theta_{ik}}{2} \frac{\sum_{j}^{N} \theta_{jk}S_{ij}}{\sum_{j}^{N} \theta_{jk} \sum_{Z}^{K} \theta_{iz}\theta_{jz} + \lambda\theta_{ik}}. \tag{10}$$

With the updating rule, we could estimate the propensities θ_{ik}. The reason why we use the multiplicative updating rule is that it is a gradient descent method with an adaptive step length and is guaranteed to converge to an optimum [19–21].

Input: G: PPI network;
λ: penalty parameter;
τ: propensity threshold;
N_s: max size of sub networks;
Output: C: predicted complexes
Algorithm:
(1) $C \leftarrow \emptyset$;
(2) get the sub-networks $\{G'\}$ in G with max size of N_s;
(3) for each sub-network G_p in $\{G'\}$
(4) compute the weight matrix S_p of G_p;
(5) Initialize the propensity matrix Θ_p of G_p;
(6) for $i = 1$ to N_p
(7) for $k = 1$ to K_p
(8) update θ_{ik} in Θ_p using the rule in (10);
(9) end for
(10) end for
(11) return (6) until convergent;
(12) get complexes C_p from Θ_p and $C \leftarrow C \cup C_p$;
(13) end for
(14) return C.

Algorithm 1: PLSMC (G, λ, τ, N_s).

2.3. Postprocessing. After estimating the propensities, we could obtain complexes using the estimated propensity matrix $\Theta = [\theta_{ik}]$. We introduce a propensity threshold τ to derive the complexes. If $\theta_{ik} \geq \tau$, the protein i is allocated to the complex k. Thus, a set of predicted complexes C in the network G is obtained, in which each element consists of a group of proteins. Moreover, as previous methods, the predicted complexes in set C that include less than 3 proteins are removed.

2.4. A Speeding-Up Strategy. The time-consuming is prohibitive when the optimizing process is directly conducted on a large-scale real world PPI network. Therefore, it is appropriate to execute the estimating process on a set of subnetworks that are of small scale but enough to identify complexes. To get the subnetworks, we recursively cluster the network into subnetworks containing proteins less than a specific size N_s. Then, apply the optimization procedure to each subnetwork to detect complexes. We use the tool of fastCommunity [22] to cluster the network. The reason is that it is a fast and robust algorithm in the field of network clustering.

In particular, we first use fastCommunity to cluster the input network and let each cluster be a subnetwork. Redo the process on each subnetwork larger than N_s, until there is no subnetwork larger than N_s.

2.5. PLSMC Algorithm. Three main steps in PLSMC are as follows: (1) get subnetworks from the input PPI network; (2) compute the weight matrix and initialize the propensity matrix with random values for each subnetwork; (3) estimate protein propensities in each subnetwork; (4) identify complexes of proteins using the postprocessing step. The pseudocode of PLSMC is in Algorithm 1.

3. Results and Discussion

We implemented a Java archive and a Web tool of the PLSMC algorithm, which is available at http://nclab.hit.edu.cn/PLSMC/. To examine its effectiveness, PLSMC is tested on several public PPI networks of yeast and compared with some state-of-the-art methods. The matching with known complexes and functional homogeneity of predicted complexes are both studied.

3.1. Dataset and Evaluation Metrics. We investigate the performance on several PPI networks of yeast (*Saccharomyces cerevisiae*), including Krogan [23], Collins [24], Gavin [2], and BioGRID [25] datasets. For Krogan, we use high confidence interactions with the probability higher than 0.273. For Gavin, only interactions with socioaffinity index larger than 5 are considered. For Collins network, we choose the top 9074 interactions with respect to purification enrichment score. The above cutoffs are suggested by original papers and [14]. In addition, all of physical interactions in BioGRID dataset (version 3.1.92) are downloaded. The general characteristics

of these networks are listed in Supplementary Table S1 available online at http://dx.doi.org/10.1155/2014/720960.

The matching between predicted complexes and known complexes is studied to evaluate the accuracy of the prediction. We use CYC2008 catalogue [26] as the gold standard of known complexes in this work, which is available at http://wodaklab.org/cyc2008/. The CYC2008 includes the complexes that are all validated by small-scale experiments and it is an up-to-date comprehensive dataset of known complexes of yeast. As in the literature [14], the known complexes in CYC2008 containing less than 3 proteins are removed.

Three metrics in the following are used to evaluate the accuracy of matching between a predicted complex set P and a gold standard B.

3.1.1. f-Measure. A predicted complex $p \in P$ and a known one $b \in B$ are considered to be matching, if the overlapping score $os(p, b)$ is greater than a matching threshold ov (ov is set to 0.25 as in [4]). The overlapping score is defined as

$$os(p, b) = \frac{|p \cap b|^2}{(|p| \times |b|)}. \tag{11}$$

Let N_{cp} be the number of predicted complexes that match at least one known complex and let N_{cb} be the number of known complexes that match at least one predicted complex. The precision and recall are defined as follows:

$$\text{precison} = \frac{N_{cp}}{|P|}, \qquad \text{recall} = \frac{N_{cb}}{|B|}. \tag{12}$$

The f-measure is the harmonic mean of precision and recall as

$$f\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}. \tag{13}$$

3.1.2. Acc Metric. Let T_{ij} be the number of common proteins between a known complex i and a predicted complex j. Then, the sensitivity (Sn) and positive predictive value (PPV) are as follows:

$$\begin{aligned} \text{Sn} &= \frac{\left(\sum_{i=1}^{|B|} \max_j \{T_{ij}\}\right)}{\left(\sum_{i=1}^{|B|} N_i\right)}, \\ \text{PPV} &= \frac{\left(\sum_{j=1}^{|P|} \max_i \{T_{ij}\}\right)}{\left(\sum_{i=1}^{|B|} \sum_{j=1}^{|P|} T_{ij}\right)}, \end{aligned} \tag{14}$$

where N_i is the number of proteins in a known complex i. Then, the accuracy metric [14] is defined as

$$\text{Acc} = \sqrt{\text{Sn} \times \text{PPV}}. \tag{15}$$

3.1.3. MMR Metric. Recently, [14] proposed a novel metric called maximum matching ratio (MMR) as follows:

$$\text{MMR} = \frac{\sum_{i=1}^{|B|} \max_{j=1}^{|P|} os(p_j, b_i)}{|B|}, \tag{16}$$

where b_i and p_j are ith known complex in B and jth predicted complex in P, respectively.

It is important to note that each of above evaluation metrics does not provide an adequate description of the matching between predicted complexes and known complexes. To make a comprehensive evaluation, we consider the composite score that is the sum of above three scores in this study. Similar composite score is also used in the literature [14].

3.2. Investigation of PLSMC. The parameter N_s in PLSMC controls the size of subnetwork and is significantly related to the effect of the speed-up strategy. We test different values of $N_s = \{50, 100, 200, 300, 400, 500\}$. Because of the prohibitive cost of computation, N_s larger than 500 is not investigated. For each value of N_s, we try different values of penalty parameter λ ($\lambda \in \{2^{-5}, \ldots, 2^5\}$) and repeat executing the algorithm 100 times with random initialization. We choose the execution that the estimated propensity matrix gives the minimal value of L in (5). We choose the values of propensity threshold τ from 0.05 to 0.5 with increment 0.05 that gives the best composite score. Supplementary Table S2 shows the best parameter setting for each value of N_s.

We demonstrate the effect of N_s with different values on the four networks in Figures 1(a) and 1(b). As in Figure 1(a), on all networks, the composite score decreases with the parameter N_s when $N_s \leq 200$ and fluctuates when $N_s > 200$. Meanwhile, the execution time increases with the parameter dramatically as in Figure 1(b). It indicates that the speed-up procedure could make a good balance between the computation time and prediction performance when $N_s = 200$. Interestingly, this is also consistent with that in CYC2008 [26], in which there is no known complex including more than 200 proteins. Therefore, in the following of this study, N_s is set to 200.

To examine the effect of the penalty term introduced in (4), we compare the PLSMC using the term and the one without using it (denoted by LSMC) applied to the four networks. The parameter setting of LSMC is shown in Supplementary Table S3. Figure 1(c) illustrates the results of PLSMC and LSMC. As shown, the PLSMC outperforms LSMC applied to all four networks. This confirms that the penalty term in (4) is essential.

3.3. Comparison with Other Methods on Matching Known Complexes. We compare PLSMC with SLCP2 [16], ClusterONE [14], RSGNM [21], OCG [12], MCL [15], and CFinder [10]. The parameters of these algorithms are tuned as follows: ClusterONE: density (d) and merging threshold (mo) both from 0.1 to 1.0 with increment 0.1; RSGNM: rate parameter $\beta \in \{2^{-5}, \ldots, 2^5\}$ and the parameter $\lambda \in \{2^{-5}, \ldots, 2^5\}$; MCL: inflation from 1.2 to 5.0 with increment of 0.1; CFinder: the size (k) of k-clique is changed from 3 to 10; OCG: using centered cliques initialization and modularity maximization; SLCP2: no parameter needs to be tuned. We remove the predicted complexes of above methods with size smaller than 3 and choose the parameter setting that yields the best composite score. The general information including parameter settings of the algorithms applied to four networks is in Supplementary Table S4, where (Com.) is the number

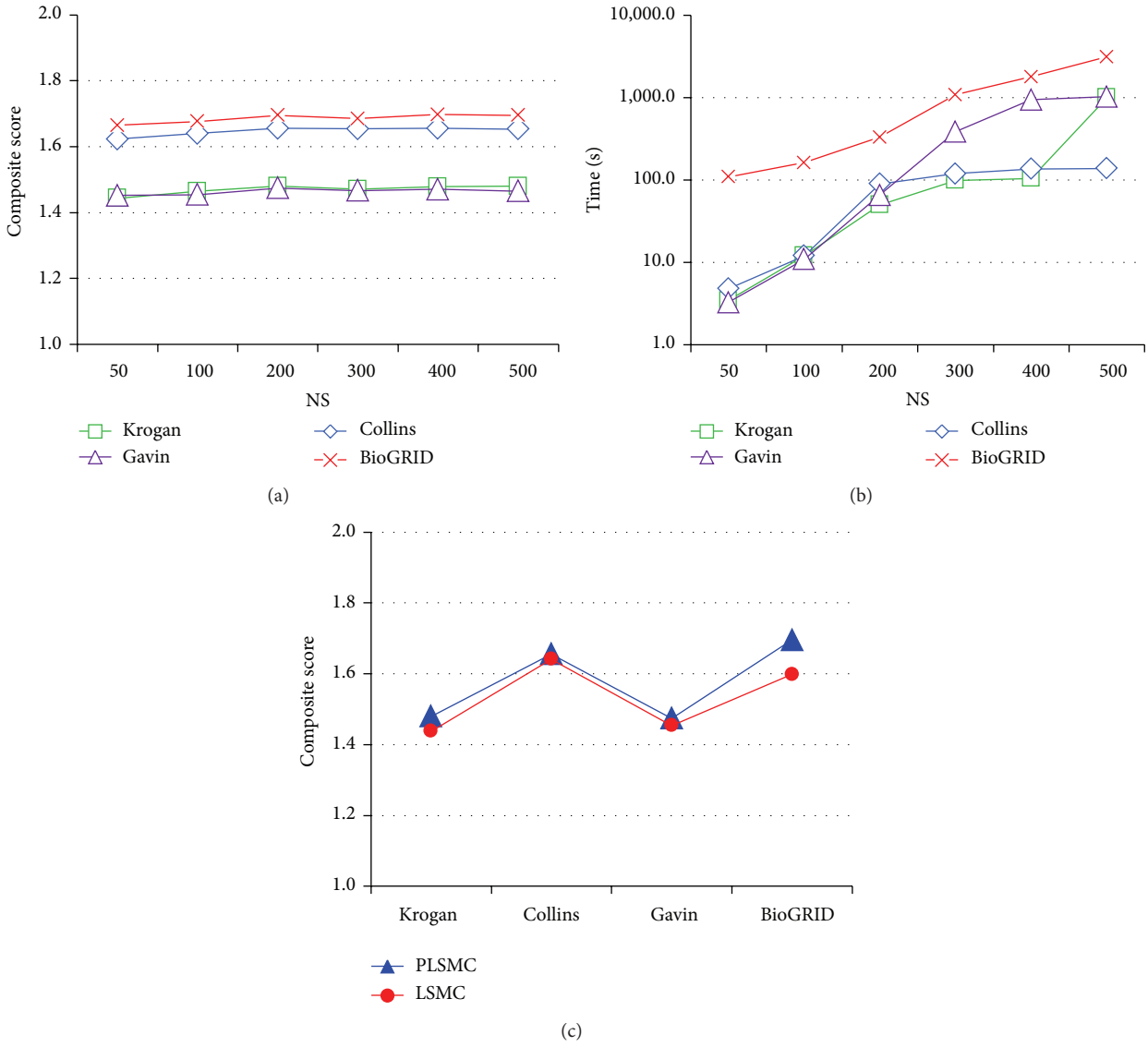


FIGURE 1: Comparison of PLSMC with different parameter setting. (a) and (b) are the comparison of composite score and execution time of PLSMC with different value of N_s (max size of subnetwork) applied to the four networks. (c) is the composite scores of PLSMC and PLSMC without the penalty term (denoted by LSMC).

of predicted complexes, (Prot.) is the number of covered proteins, and (Size) is the average size of predicted complexes. We cannot obtain the results of CFinder on BioGRID network, as the calculation requires more memory than a typical computer.

We present the comparison result of matching with gold standard in Figure 2. On all four networks, PLSMC could get better composite score than other methods. ClusterONE gets close results to PLSMC on all networks. SLCP2 and OCG provide good performance when applied to Collins network but make poor predictions about other networks. It indicates that these two methods are prone to be affected by different networks. MCL achieves poor performance when applied to all networks.

In addition, we also investigate the number of known complexes that are matched by predicted complexes. The number of matched known complexes of various algorithms applied to Krogan, Collins, Gavin, and BioGRID networks is illustrated in Figures 3(a)–3(d), respectively. We show the results of the overlapping threshold ov from 0.5 to 1.0. It denotes a perfect matching when $ov = 1$. As shown, PLSMC can hit 15, 36, 16, and 23 known complexes with perfect matching on four networks, respectively. It can also be found that, on Krogan, Collins, and BioGRID networks, PLSMC can provide the greatest number using all thresholds. On Gavin network, PLSMC could get comparative results with ClusterONE with all thresholds and match more known complexes with perfect matching than others. Generally, the

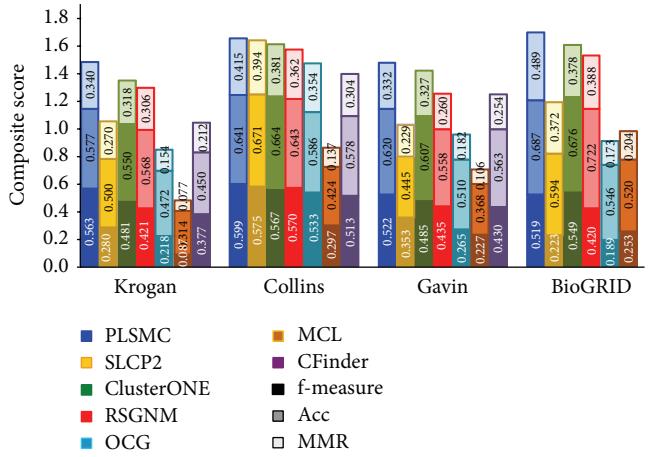
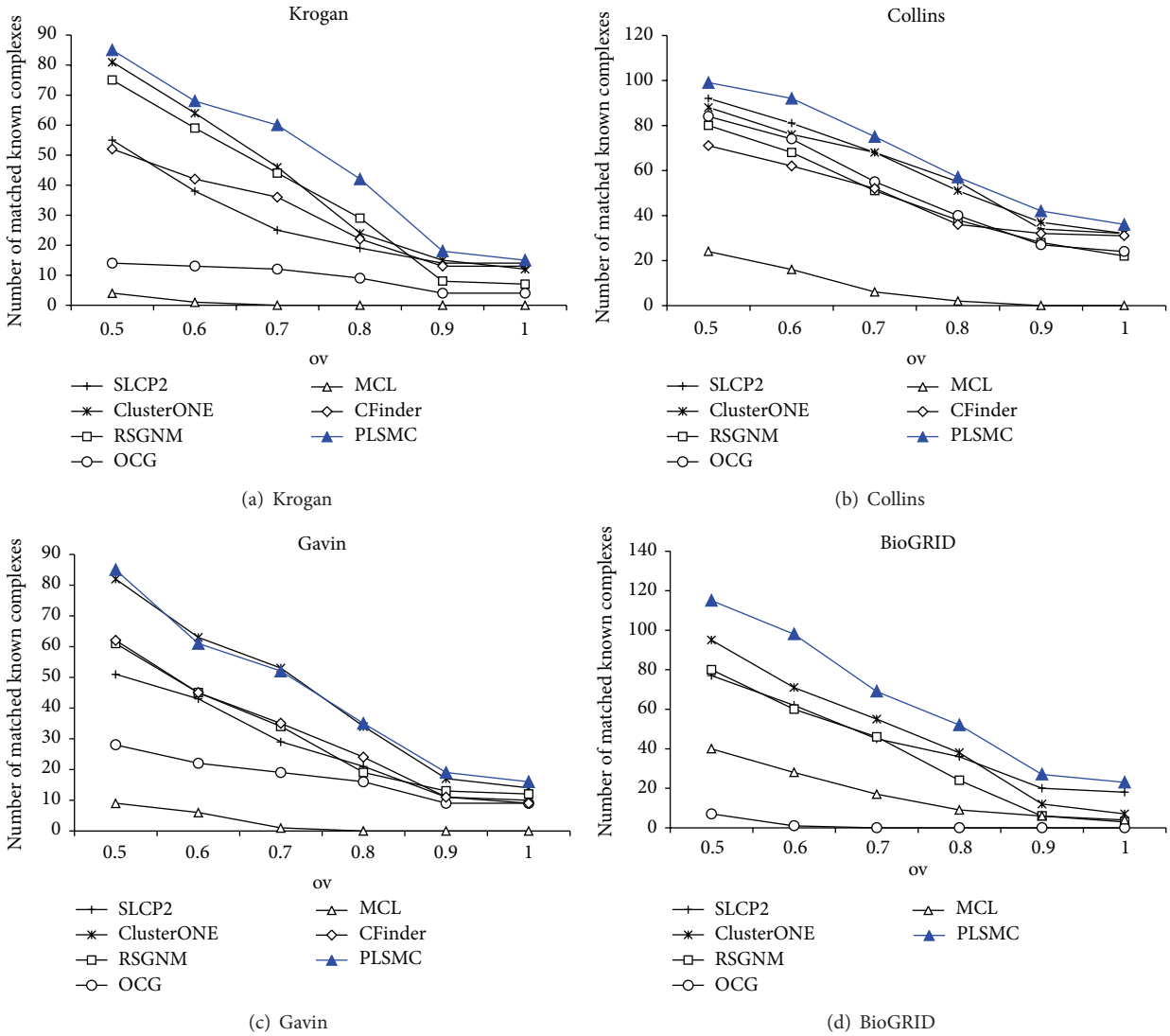


Figure 2: Comparison on composite score of the algorithms applied to four networks. Various shades of the same color denote f-measure, Acc, and MMR submetrics. The total height of each bar is the value of composite score.



(a) Krogan (b) Collins

(c) Gavin (d) BioGRID

Figure 3: The number of matched known complexes of the algorithms.

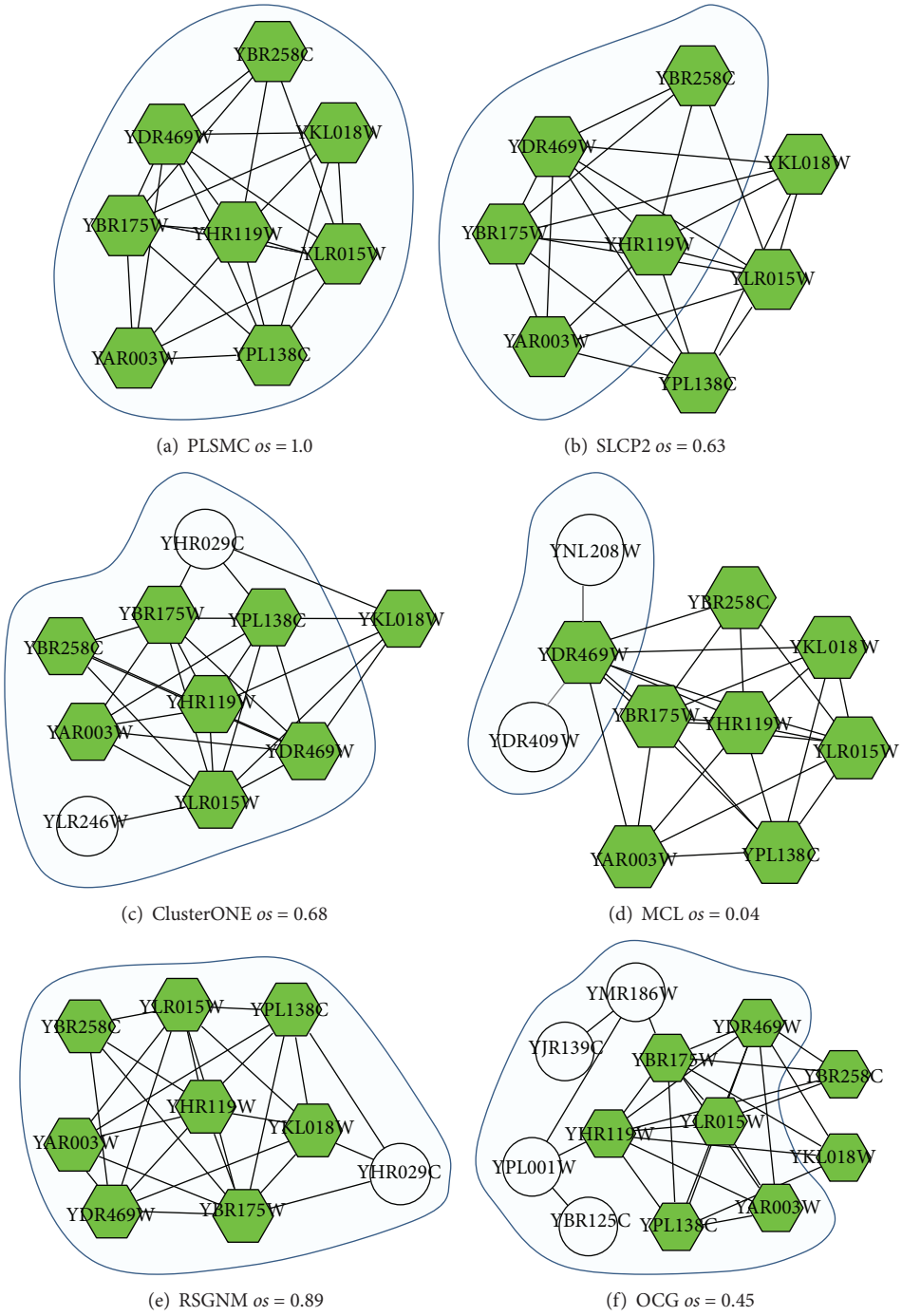


Figure 4: The COMPASS complex as detected by the six algorithms. Hexagon nodes represent the proteins involved in the COMPASS complex. Shaded areas are the clusters detected by the algorithms, which have the max overlapping scores (*os*) with COMPASS complex.

above comparisons confirm that the PLSMC outperforms other methods in terms of matching known complexes in gold standard.

We show how the studied algorithms identify the known COMPASS complex from the Krogan network in Figure 4. The COMPASS complex is an important conserved protein complex that catalyzes methylation of histone H3, which is collected in both CYC2008 and GO (GO: 0048188). The complex contains 8 proteins (YKL018W, YPL138C, YBR175W, YDR469W, YHR119W, YLR015W, YAR003W, and YBR258C), which are denoted by hexagon nodes in Figure 4. The clusters under the shaded areas are detected by the algorithms, which have the max overlapping scores (*os*) with COMPASS complex. As shown, PLSMC is the only algorithm that is able to detect this complex with perfect matching. All of the other algorithms make inaccurate prediction. SLCP2 detects

Table 1: Comparison on biological relevance of complexes predicted by the algorithms.

Network	Method	MF	BP	CC
Krogan	**PLSMC**	**0.479**	**0.457**	**0.592**
	SLCP2	0.394	0.114	0.094
	ClusterONE	0.311	0.291	0.357
	RSGNM	0.392	0.270	0.270
	OCG	0.199	0.185	0.331
	MCL	0.265	0.057	0.033
	CFinder	0.296	0.287	0.330
Collins	**PLSMC**	**0.536**	**0.460**	**0.620**
	SLCP2	0.405	0.353	0.410
	ClusterONE	0.401	0.377	0.419
	RSGNM	0.376	0.371	0.418
	OCG	0.519	0.439	0.612
	MCL	0.380	0.240	0.331
	CFinder	0.439	0.351	0.439
Gavin	**PLSMC**	**0.399**	**0.362**	**0.467**
	SLCP2	0.374	0.153	0.189
	ClusterONE	0.374	0.308	0.360
	RSGNM	0.382	0.333	0.389
	OCG	0.381	0.310	0.405
	MCL	0.308	0.112	0.210
	CFinder	0.387	0.350	0.401
BioGRID	**PLSMC**	**0.459**	**0.452**	**0.511**
	SLCP2	0.443	0.184	0.117
	ClusterONE	0.439	0.447	0.439
	RSGNM	0.363	0.277	0.267
	OCG	0.262	0.321	0.343
	MCL	0.400	0.176	0.140
	CFinder	—	—	—
CYC2008		0.458	0.424	0.525

MF, molecular function; BP, biological process; CC, cellular compartment.

a part of the complex and other algorithms include unrelated proteins into the complex. The result of CFinder is not shown, because the detected cluster that has the best matching with the complex is a huge cluster, which consists of 627 proteins.

3.4. Biological Relevance of Predicted Complexes. The known complex dataset is incomplete. For example, CYC2008 only covers 1627 proteins, while the number of proteins in yeast is more than 5000. Therefore, a predicted complex that does not match with any known complex is possibly not a false positive one and it is worth further in-depth analysis. To this end, we also examine the biological relevance of predicted complexes in terms of functional homogeneity. This is because the proteins within a complex tend to be located in the same cellular component (CC) or are involved in a common molecular function (MF) or biological process (BP) [4, 14]. We use the tool of GO::TermFinder (Version 0.83) [27] to compute the P value for each predicted complex. The GO corpus is downloaded from Saccharomyces Genome Database [28]. We investigate all three aspects of GO.

A predicted complex that has more than one annotation with the P value smaller than a threshold p is considered functional homogeneity. The threshold p is set to $1.0E-10$ [4]. The fraction of predicted complexes that are functional homogeneity is used to evaluate the performance of the prediction method.

Table 1 presents the comparison of functional homogeneity of complexes predicted by different methods. The result of known complexes in CYC2008 is also listed. It can be found that the complexes predicted by PLSMC are more functional homologous than those of other methods. Moreover, the results of PLSMC applied to Krogan, Collins, and Biological networks are all better than that of CYC2008. More interestingly, on all networks, the results of PLSMC in regard to CC aspect are better than MF and BP aspects. This tendency is consistent with that of CYC2008. On the whole, the comparison demonstrates that the complexes derived by PLSMC are more biologically relevant.

4. Conclusion

In this paper, we present PLSMC, a penalized least squares method, to detect complexes from PPI network. PLSMC identifies complexes by minimizing the distances between cocomplex coefficients and interaction weights of all pairs of proteins. We test it on several yeast PPI networks. The results show that PLSMC achieves higher accuracy in matching with known complexes than some state-of-the-art methods. Moreover, the predicted complexes also have good biological relevance to functional homogeneity. This study confirms that PLSMC, based on a least squares method, is an effective approach to identify complexes from the PPI network.

We note that integrating multiple biological data sources in addition to PPI network [29] can improve the identification of protein complexes. On the one hand, most of available protein-protein interaction networks are static. Combining dynamic information such as expression profiles can infer the dynamic properties of protein-protein interactions under different time points or various conditions [1, 30]. On the other hand, when two or more proteins form a complex, some interface information as physical folds [31], biochemical properties [32], and posttranslation modifications [33] is very important to the complex formation. In the future, based on PLSMC, we will study the identification of protein complexes from dynamic protein-protein interaction networks and interface datasets.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by Natural Science Foundation of China (61271346, 61172098, and 91335112), Specialized Research Fund for the Doctoral Program of Higher Education of China (20112302110040), and Fundamental Research Funds for the Central Universities (HIT.KISTP.201418). This

work was performed at the School of Computer Science and Technology, Harbin Institute of Technology, China.

References

[1] B. Chen, W. Fan, J. Liu, and F. X. Wu, "Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 177–179, 2014.

[2] A.-C. Gavin, P. Aloy, P. Grandi et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[3] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12123–12128, 2003.

[4] J. Z. Ji, A. D. Zhang, C. N. Liu, X. M. Quan, and Z. J. Liu, "Survey: functional module detection from protein-protein interaction networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 261–277, 2014.

[5] J. Wang, M. Li, Y. Deng, and Y. Pan, "Recent advances in clustering methods for protein interaction networks," *BMC Genomics*, vol. 11, no. 3, article S10, 2010.

[6] X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Computational approaches for detecting protein complexes from protein interaction networks: A survey," *BMC Genomics*, vol. 11, supplement 1, article S3, 2010.

[7] C. Zhang, S. Liu, and Y. Zhou, "Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast," *Journal of Proteome Research*, vol. 5, no. 4, pp. 801–807, 2006.

[8] J. Wang, B. Liu, M. Li, and Y. Pan, "Identifying protein complexes from interaction networks based on clique percolation and distance restriction," *BMC Genomics*, vol. 11, no. 2, article S10, 2010.

[9] B. Chen, J. Shi, S. Zhang, and F.-X. Wu, "Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy," *Proteomics*, vol. 13, no. 2, pp. 269–277, 2013.

[10] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: Locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, 2006.

[11] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[12] E. Becker, B. Robisson, C. E. Chapple, A. Guénoche, and C. Brun, "Multifunctional proteins revealed by overlapping clustering in protein interaction network," *Bioinformatics*, vol. 28, no. 1, Article ID btr621, pp. 84–90, 2012.

[13] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, article 2, 2003.

[14] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.

[15] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[16] Y. J. Wang and X. N. Qian, "Functional module identification in protein interaction networks by interaction patterns," *Bioinformatics*, vol. 30, no. 1, pp. 81–93, 2014.

[17] M. Mete, F. Tang, X. Xu, and N. Yuruk, "A structural approach for finding functional modules from large biological networks," *BMC Bioinformatics*, vol. 9, no. 9, article S19, 2008.

[18] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 5th SIAM International Conference on Data Mining (SDM '05)*, pp. 606–610, April 2005.

[19] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[20] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562, 2000.

[21] X.-F. Zhang, D.-Q. Dai, and X.-X. Li, "Protein complexes discovery based on protein-protein interaction data via a regularized sparse generative network model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 857–870, 2012.

[22] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, Article ID 066111, 2004.

[23] N. J. Krogan, G. Cagney, H. Yu et al., "Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[24] S. R. Collins, P. Kemmeren, X.-C. Zhao et al., "Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae," *Molecular and Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.

[25] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D535–D539, 2006.

[26] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, "Up-to-date catalogues of yeast protein complexes," *Nucleic Acids Research*, vol. 37, no. 3, pp. 825–831, 2009.

[27] E. I. Boyle, S. Weng, J. Gollub et al., "GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[28] E. L. Hong, R. Balakrishnan, Q. Dong et al., "Gene Ontology annotations at SGD: new data sources and annotation methods," *Nucleic Acids Research*, vol. 36, no. 1, pp. D577–D581, 2008.

[29] M. Wu, Z. P. Xie, X. L. Li, C. K. Kwoh, and J. Zheng, "Identifying protein complexes from heterogeneous biological data," *Proteins-Structure Function and Bioinformatics*, vol. 81, no. 11, pp. 2023–2033, 2013.

[30] J. Wang, X. Peng, W. Peng, and F. X. Wu, "Dynamic protein interaction network construction and applications," *Proteomics*, vol. 14, no. 4-5, pp. 338–352, 2014.

[31] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[32] H. Naveed and J. J. Han, "Structure-based protein-protein interaction networks and drug design," *Quantitative Biology*, vol. 1, no. 3, pp. 183–191, 2013.

[33] Y. Jiang, B. Q. Li, Y. Zhang et al., "Prediction and analysis of post-translational pyruvoyl residue modification sites from internal serines in proteins," *PLoS ONE*, vol. 8, no. 6, Article ID e66678, 2013.