Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

# Overlap maximum matching ratio (OMMR): a new measure to evaluate overlaps of essential modules*

Xiao-xia ZHANG†1, Qiang-hua XIAO2, Bin LI1, Sai HU1, Hui-jun XIONG1, Bi-hai ZHAO†‡1
([1]Department of Mathematics and Computer Science, Changsha University, Changsha 410003, China)
([2]School of Information Science and Engineering, Central South University, Changsha 410083, China)
†E-mail: zhangxx111@yeah.net; bihaizhao@163.com

Received Aug. 6, 2014; Revision accepted Dec. 16, 2014; Crosschecked Mar. 9, 2015

**Abstract:** Protein complexes are the basic units of macro-molecular organizations and help us to understand the cell's mechanism. The development of the yeast two-hybrid, tandem affinity purification, and mass spectrometry high-throughput proteomic techniques supplies a large amount of protein-protein interaction data, which make it possible to predict overlapping complexes through computational methods. Research shows that overlapping complexes can contribute to identifying essential proteins, which are necessary for the organism to survive and reproduce, and for life's activities. Scholars pay more attention to the evaluation of protein complexes. However, few of them focus on predicted overlaps. In this paper, an evaluation criterion called overlap maximum matching ratio (OMMR) is proposed to analyze the similarity between the identified overlaps and the benchmark overlap modules. Comparison of essential proteins and gene ontology (GO) analysis are also used to assess the quality of overlaps. We perform a comprehensive comparison of serveral overlapping complexes prediction approaches, using three yeast protein-protein interaction (PPI) networks. We focus on the analysis of overlaps identified by these algorithms. Experimental results indicate the important of overlaps and reveal the relationship between overlaps and identification of essential proteins.

**Key words:** Protein-protein interaction network, Essential protein modules, Overlap, Overlap maximum matching ratio
**doi:**10.1631/FITEE.1400282 **Document code:** A **CLC number:** TP311; R857.3

## 1 Introduction

Functional modules or protein complexes are the basic units of macro-molecular organizations and perform all kinds of fundamental biological functions in cells. Recently, the focus of bioinformatics has transferred from sequence to protein interactions. The identification of functional modules or protein complexes is a hot topic in the post-genomic era.

Methods for protein complexes identification can be classified into two types: experimental method and computational method. A lot of computational approaches for identifying functional modules or complexes from protein-protein interaction (PPI) networks have been developed. The development of the yeast two-hybrid, tandem affinity purification, and mass spectrometry has supplied a large amount of protein-protein interaction data and provided fundamental and abundant data for computional approaches to the inference of protein complexes.

Generally, a PPI network can be modeled as a graph. Thus, identifying protein complexes is translated into detecting a dense subgraph containing many connections in a graph. Several approaches based on graph theory have been developed to identify protein complexes from PPI networks, such as molecular complex detection (MCODE) (Bader and Hogue, 2003), the Markov cluster algorithm (MCL)

‡ Corresponding author

* Project supported by the National Scientific Research Foundation of Hunan Province, China (Nos. 14C0096, 10C0408, and 10B010), the Natural Science Foundation of Hunan Province, China (Nos. 13JJ4106 and 14JJ3138), and the Science and Technology Plan Project of Hunan Province, China (No. 2010FJ3044)
ORCID: Bi-hai ZHAO, http://orcid.org/0000-0003-0870-7468

(Enright *et al.*, 2002), and the clique percolation method (CPM) (Palla *et al.*, 2005). Adamcsek *et al.* (2006) developed a software component for uncovering overlapping complexes from PPI networks. Clustering based on the maximal clique algorithm (CMC) has constructed a weighted PPI network and predicted complexes from the constructed network (Liu *et al.*, 2009). Speed and performance in the clustering algorithm (SPICi) is a variant of CMC, but it can achieve a faster running speed (Jiang and Singh, 2010). Based on a repeated random walk (RRW), Macropol *et al.* (2009) proposed an efficient algorithm for discovering protein complexes. This new algorithm is biologically sensitive and uses weights of given edges to find overlapping complexes. Recently, clustering with overlapping neighborhood expansion (ClusterONE) has been developed to detect potentially overlapping protein complexes from PPI networks (Nepusz *et al.*, 2012). A new protein complexes identification method was proposed based on graph entropy and clique seeds (Chen *et al.*, 2013). To address the limitation of MCL, Shih and Parthasarathy (2012) proposed the regularized MCL (R-MCL), allowing for highly overlapped clusters. Inspired by the bacteria foraging optimization mechanism (BFO) and an intuitionistic fuzzy set, Lei *et al.* (2013) proposed an improved clustering method to detect overlapping modules in PPI networks.

These existing approaches fail to take into account inherent organization. Research (Dezső *et al.*, 2003; Gavin *et al.*, 2006) has shown that a complex consists of a core component and some attachment proteins. Based on the core-attachment concept, some algorithms have been proposed by Wu *et al.* (2009) and Leung *et al.* (2009). Taking into account the uncertainty of interactions in the PPI network, Ni *et al.* (2013) constructed a probabilistic protein-protein interaction (Pro-PPI) network, in which the reliability of each interaction is represented as a probability using the topology of the PPI network. After constructing the Pro-PPI network, a new method named WN-PC (weighted network based method for predicting protein complexes) was proposed to identify overlapping protein complexes from the Pro-PPI network.

Generally, approaches based on graph clustering are not ideal due to the predicted non-overlapping modules, while proteins may have a variety of functions. So, a protein may appear in more than one protein complex. In other words, there are some overlaps between the identified protein complexes. Protein complexes and their overlaps are helpful for the detection of essential proteins (Hart *et al.*, 2007) and others. Hart *et al.* (2007) have pointed out that essentiality is a product of the protein complex rather than the individual protein. Han *et al.* (2004) suggested that there are two sorts of essential proteins (hubs) in the yeast PPIN, such as party hubs and date hubs. Date hubs are just like mediators and adaptors, while party hubs appear within distinct modules. Inspired by these studies, we have succeeded in predicting essential proteins based on overlapping essential modules in previous research (Zhao *et al.*, 2014). Results indicate that proteins in these overlaps tend to be date hubs, which play more important roles than party hubs. So, it is necessary to identify overlapping essential protein modules and evaluate the quantity of overlaps. Essential modules are made up of some densely connected and function-shared proteins, which contain a large number of essential proteins. Effective evaluation of overlapping complexes and their overlaps could improve the precision of identification of essential proteins. Overlaps supporting these algorithms (ClusterONE excluded) ignore the assessment of the identified overlapping module pairs. Even for ClusterONE, only overlap size distribution between generated complexes pairs is shown. Further study of evaluating the quality of overlaps between cluster pairs is lacking.

As mentioned above, overlapping protein complexes and their overlaps play an important role in identifying essential proteins. To make a comprehensive comparison about module overlaps identified by various overlapping algorithms, a measure called the overlap maximum matching ratio (OMMR) is introduced. OMMR is used to match the benchmark overlaps and the identified overlaps. We also use the predicted protein complexes and overlaps identified by various algorithms for the identification of essential proteins. We apply compared methods in yeast PPI networks downloaded from several different high throughput datasets and perform a comprehensive comparison of the state-of-the-art approaches, such as WN-PC (Ni *et al.*, 2013), ClusterONE (Nepusz *et al.*, 2012), COACH (Wu *et al.*, 2009), RRW (Macropol *et al.*, 2009), and CMC (Liu *et al.*, 2009). Experimental results indicate the importance of overlaps and reveal the relationship between overlaps and essential protein identification.

## 2 Results and discussion

### 2.1 Experimental data

The PPI networks used for computational analysis come from bakers' yeast. We have applied five overlapping protein complex detection algorithms, namely WN-PC, ClusterONE, COACH, RRW, and CMC on two yeast PPI networks, including Gavin data (Gavin *et al*., 2006) and Krogan data (Krogan *et al*., 2006). BioGRID data (Stark *et al.*, 2006), a highly clustered network, was also used to test the effectiveness of OMMR. For all these selected algorithms, the optimal parameters were set as recommended by respective authors. The aggregation coefficient threshold and sample degree threshold of WN-PC were 0.3 and 0.1, respectively. For ClusterONE, the minimum size was set as 3 and the overlap threshold 0.8. For COACH, the threshold to filter redundant complex cores was 0.225. As for CMC, the minimum size of the clusters generated was 4 and the threshold used to remove or merge highly overlapped clusters 0.5. For RRW, the maximum cluster size was 11, the minimum cluster size 5, and the overlap threshold 0.2.

We will first present in detail the results on Gavin data, and the results using Krogan will also be briefly presented to demonstrate the effectiveness of OMMR. The Gavin dataset is made up of 1855 proteins, and there are 7669 interactions among the proteins. The Korgan dataset consists of 3672 proteins and 14317 interactions. The BioGRID dataset consists of 5616 proteins and 52833 interactions. We have removed the self-interactions and repeated interactions from the PPI networks.

To evaluate and analyze the overlapping complexes predicted by these methods, CYC2008 (Pu *et al*., 2009) was used as a benchmark set, which consists of 408 protein complexes. A reference set of essential proteins used in our experiments was collected from the following databases: MIPS (Mewes *et al*., 2006), SGD (Cherry *et al*., 1998), DEG (Zhang and Lin, 2009), and SGDP, which consists of 1285 essential proteins.

### 2.2 Overlap maximum matching ratio

This study focuses on the evaluation of overlaps of predicted protein complexes, while it is firstly necessary to assess the quality of complexes produced by these methods mentioned above. Several evaluation measures, including precision, recall, F-measure, and the coverage rate, are adopted to analyze the overlapping complexes. To assess the quality of the produced overlapping complexes, we match the generated protein complexes with the benchmark sets. The overlap score (OS) of a predicted complex $A$ and a benchmark complex $B$ is defined as follows (Wu *et al*., 2009):

$$\mathrm{OS}(A, B)=\frac{|A \cap B|^2}{|A|\cdot|B|}, \tag{1}$$

where $|\cdot|$ denotes the number of elements in the set.

A predicted complex is considered to match the benchmark complex if their OS value is no less than a threshold, which typically was set at 0.2 (Bader and Hogue, 2003; Wu *et al*., 2009). Precision and recall are the general measures to evaluate the quality of protein complexes prediction methods. Precision is the proportion of detected modules that are matched, while recall is the proportion of benchmark sets that are matched. F-measure is the harmonic mean of precision and recall. Mathematically, the three measures are defined as follows:

$$\text{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\ \text{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{2}$$

$$\mathrm{F}_{\text{measure}}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{3}$$

where TP (short for true positive) is the number of detected complexes matched by benchmark sets, FP (short for false positive) is the number of detected complexes that are not matched by benchmark sets, and FN (short for false negative) is the number of benchmark sets that are not matched by detected complexes.

On the other hand, the coverage rate is introduced to measure how many proteins in the benchmark sets can be covered by the detected complexes. Given a benchmark set BM and a predicted complex set PM, $T_{ij}$ is the common number of proteins between the $i$th benchmark module and the $j$th predicted module. The coverage rate (CR) is then defined as

$$\mathrm{CR}=\sum_{i=1}^{|\mathrm{BM}|}\max_{1\le j\le|\mathrm{PM}|}T_{ij}\Bigg/\sum_{i=1}^{|\mathrm{BM}|}N_i, \tag{4}$$

where $N_i$ is the number of proteins within the $i$th benchmark complex.

Table 1 lists the basic information of prediction results for each method when the threshold of OS was set as 0.2. WN-PC detects 485 protein complexes which contain at least three proteins. The average and maximum sizes of those predicted modules are 11.2 and 39, respectively. Among the 485 protein complexes predicted by WN-PC, 284 modules match at least a benchmark set (MPF, matched predicted functional modules), and 157 benchmark sets are matched by at least a predicted one (MBS, matched benchmark set).

**Table 1 Information of prediction results**

| Algorithm | \|PM\| | Average size | Maximum size | MBS | MPF |
|---|---|---|---|---|---|
| WN-PC | 485 | 11.2 | 39 | 157 | 284 |
| ClusterONE | 294 | 6.9 | 40 | 154 | 130 |
| COACH | 361 | 8.3 | 36 | 159 | 185 |
| RRW | 152 | 6.5 | 11 | 92 | 110 |
| CMC | 165 | 7.9 | 37 | 130 | 104 |

Table 2 shows the overall comparison, including F-measure and CR. The F-measure of WN-PC is 0.56, which is 47.37%, 19.15%, 64.70%, and 47.37% higher than that of ClusterONE, COACH, RRW, and CMC, respectively.

**Table 2 Comparison of F-measure and coverage rate**

| Algorithm | Precision | Recall | F-measure | CR |
|---|---|---|---|---|
| WN-PC | 0.59 | 0.53 | 0.56 | 0.52 |
| ClusterONE | 0.44 | 0.34 | 0.38 | 0.42 |
| COACH | 0.51 | 0.43 | 0.47 | 0.43 |
| RRW | 0.61 | 0.24 | 0.34 | 0.35 |
| CMC | 0.63 | 0.27 | 0.38 | 0.40 |

From Tables 1 and 2, we can see that WN-PC outperforms the other methods for predicted complexes. Next, we will analyze overlaps of these complexes in detail.

Although some redundancy may be of biological importance, complexes overlapping to a very high extent in comparison to their expected density and size should be discarded. So, we perform some preprocessing operations for these predicted complexes by various algorithms. When quantifying the extent of overlap between each pair of complexes, a complex with a small expected density or size is discarded when the overlap score of the pair is above the threshold. In this paper, the overlap threshold is typically set as 0.8 (Nepusz *et al.*, 2012), where the overlap score of two complexes $A$ and $B$ is obtained according to Eq. (1).

We propose an evaluation criterion called OMMR to assess the quality of overlaps detected by various overlapping functional module prediction algorithms.

Given a benchmark modules set BM, SBM= $\{\mathrm{sbm}_1, \mathrm{sbm}_2, ..., \mathrm{sbm}_n\}$ is a subset of BM, in which each module matches at least one protein by predicted protein complexes. PM is a predicted protein complexes set, and

$$\mathrm{SPM}=\left\{\max_{1\le j\le|\mathrm{PM}|}\mathrm{OS}(\mathrm{sbm}_1,\mathrm{pm}_j),\ \max_{1\le j\le|\mathrm{PM}|}\mathrm{OS}(\mathrm{sbm}_2,\mathrm{pm}_j),\ \ldots,\ \max_{1\le j\le|\mathrm{PM}|}\mathrm{OS}(\mathrm{sbm}_n,\mathrm{pm}_j)\right\}$$

is a subset of PM. The OMMR can be defined as

$$\mathrm{OMMR}=\frac{1}{|\mathrm{SBM}\times\mathrm{SBM}|}\sum_{i=1}^{|\mathrm{SBM}|}\sum_{i=1}^{|\mathrm{SBM}|}\mathrm{spm}_i\cap\mathrm{spm}_j. \tag{5}$$

In Eq. (5), SBM×SBM is the overlaps set among SBM, and $|\mathrm{spm}_i\cap\mathrm{spm}_j|$ is the number of matched proteins in the overlap between $\mathrm{sbm}_i$ and $\mathrm{sbm}_j$. Table 3 lists the overlaps predicted by various algorithms.

In Tables 1 and 3, we can see that WN-PC predicts 485 functional modules and matches 157 benchmark modules. Among these 157 benchmark modules, there are 185 overlaps, 106 of which have been matched by WN-PC and average precision is 0.899. According to Eq. (5), the OMMR of WN-PC is 0.547, which is 86.05%, 19.43%, 782.26%, and 120.56% higher than that of ClusterONE, COACH, RRW, and CMC, respectively.

To make a comprehensive comparison, we analyze the average precision and OMMR of overlaps predicted by various algorithms according to their sizes, respectively (Figs. 1 and 2). The maximum size of overlaps in the benchmark set was 17. From Fig. 1 we can see that WN-PC, ClusterONE, and COACH archive higher average precision than the other methods.

From Fig. 2 we can see that WN-PC and COACH can still obtain high OMMRs, even if the number of predicted overlaps used in matching decreases sharply. For example, when the number of proteins in overlaps equaled 1, COACH matched 47 benchmark overlaps and the OMMR was 0.117. For WN-PC, when the size of overlaps was 3, only 18.87% of benchmark overlaps were matched, while the OMMR was 0.112. So, we believe that even when their overlaps are few in terms of the number of proteins, the OMMR is very likely to be high for WN-PC and COACH.

**Table 3 Comparison of overlaps predicted by various algorithms**

| Algorithm | Number of matched overlaps | Average precision | OMMR |
|---|---|---|---|
| WN-PC | 106 | 0.899 | 0.547 |
| ClusterONE | 69 | 0.802 | 0.294 |
| COACH | 101 | 0.802 | 0.458 |
| RRW | 23 | 0.333 | 0.062 |
| CMC | 60 | 0.768 | 0.248 |

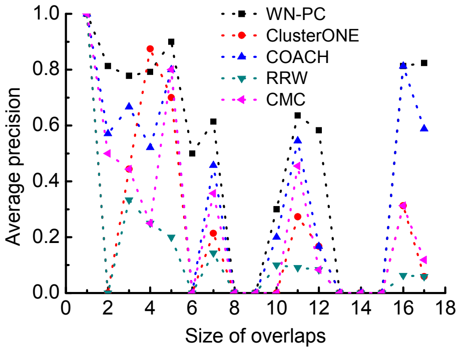

**Fig. 1 Comparison of average precision according to the size of overlaps**

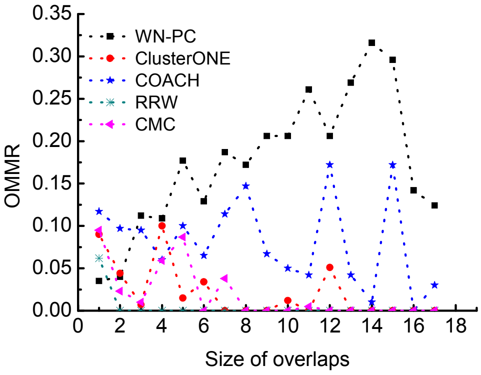

**Fig. 2 Comparison of OMMR according to the size of overlaps**

## 2.3 Comparison of essential proteins in overlaps

To reveal the relationship between essential proteins and overlaps of protein complexes and demonstrate the effectiveness of OMMR, we analyze the essentiality of proteins in protein complexes and overlaps predicted by these methods. Among all the 1855 proteins in the Gavin PPI network, 714 proteins are essential. Table 4 shows a comparison of the number of essential proteins in protein complexes and overlaps predicted by various algorithms.

**Table 4 Numbers of essential proteins in protein complexes and overlaps predicted by various algorithms**

| Algorithm | $N_{pc}$ | $N_{epc}$ | $N_{po}$ | $N_{epo}$ |
|---|---|---|---|---|
| WN-PC | 1364 | 591 | 850 | 399 |
| ClusterONE | 1624 | 652 | 351 | 166 |
| COACH | 1393 | 598 | 648 | 321 |
| RRW | 908 | 422 | 79 | 39 |
| CMC | 1095 | 515 | 184 | 110 |

$N_{pc}$: number of proteins in complexes; $N_{epc}$: number of essential proteins in complexes; $N_{po}$: number of proteins in overlaps; $N_{epo}$: number of essential proteins in overlaps

From Table 4 we can see that WN-PC, ClusterONE, and COACH have identified more essential proteins than the other methods, and WN-PC predicts the largest number of essential proteins in overlaps. We also calculate the percentage of essential proteins in overlaps among total proteins and total predicted essential proteins by various algorithms (Table 5).

As shown in Table 5, WN-PC has the highest frequencies of essential proteins in predicted protein complexes among the five methods. The top three in terms of the performance of identifying essential proteins are WN-PC, COACH, and ClusterONE, respectively. The results accord with Tables 1–3, and verify the effectiveness of OMMR. The results also indicate that overlaps in protein complexes play important roles in identifying essential proteins.

**Table 5 Percentages of essential proteins in overlaps among total proteins and predicted essential proteins by various algorithms**

| Algorithm | $P^*$ (%) | $P^{**}$ (%) |
|---|---|---|
| WN-PC | 29.25 (399/1364) | 67.51 (399/591) |
| ClusterONE | 10.22 (166/1624) | 25.46 (166/652) |
| COACH | 23.04 (321/1393) | 53.68 (321/598) |
| RRW | 4.30 (39/908) | 9.24 (39/422) |
| CMC | 10.05 (110/1095) | 21.36 (110/515) |

$P^*$: percentage of essential proteins among total proteins; $P^{**}$: percentage of essential proteins among predicted essential proteins

## 2.4 Visualization and statistics of overlaps

To obtain an overall view of overlaps by various algorithms, visualization and centrality statistics of predicted overlaps are given in this section. Fig. 3 shows the visualization of overlaps predicted by WN-PC, COACH, ClusterONE, and CMC. RRW correctly matched only one overlap, which is far fewer than those of other algorithms. For this reason, RRW is not included in this section.

Fig. 3 illustrates that overlaps predicted by WN-PC and COACH have more interactions among themselves. The essentiality of proteins in the network depends on many factors, so we made statistics for degree centrality (DC) and betweenness centrality (BC) of proteins in overlaps predicted by various methods, to provide some more information for further essential proteins study. Tables 6 and 7 list statistics for DC and BC, respectively.

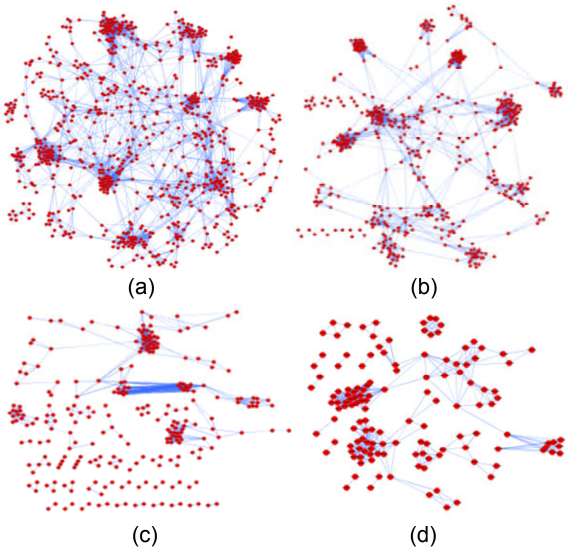

**Fig. 3 Visualization of overlaps by various algorithms**
(a) WN-PC; (b) COACH; (c) ClusterONE; (d) CMC

**Table 6 Statistics of DC by various algorithms**

| Algorithm | Percentage (%) | | | |
|---|---|---|---|---|
| | [1, 5] | [6, 10] | [11, 20] | >20 |
| WN-PC | 11.04 | 34.72 | 39.88 | 14.36 |
| ClusterONE | 35.85 | 24.91 | 23.77 | 15.47 |
| COACH | 3.67 | 24.04 | 48.62 | 23.67 |
| CMC | 0 | 15.22 | 47.10 | 37.68 |

**Table 7 Statistics of BC by various algorithms**

| Algorithm | Percentage (%) | | | |
|---|---|---|---|---|
| | [0, 1000] | (1000, 5000] | (5000, 10000] | >10000 |
| WN-PC | 27.85 | 24.29 | 13.13 | 34.73 |
| ClusterONE | 29.81 | 24.91 | 18.49 | 26.79 |
| COACH | 24.59 | 26.06 | 13.94 | 35.41 |
| CMC | 13.04 | 18.84 | 18.84 | 49.28 |

## 2.5 Gene ontology analysis of overlaps

To further validate the overlaps detected by various methods, we employ functional enrichment of GO terms to investigate the biological significance of overlaps of predicted complexes. To determine whether any gene ontology (GO) terms annotate a specified list of genes at a frequency greater than what would be expected by chance, GO::TermFinder calculates a P-value using hyper geometric distribution (Boyle *et al.*, 2004).

A low P-value of a predicted overlap indicates that those proteins in the overlap do not occur merely by accident, so the overlap achieves high statistical significance. Generally an overlap is considered to be significant with P-value<0.01 (Hu *et al.*, 2005).

Maraziotis *et al.* (2007) pointed out that the proportion of significant modules over all detected ones can be used to assess overall performance of various methods. In addition, the P-score is used as an effective evaluation measure, defined as

$$\mathrm{P}_{\text{score}} = \frac{1}{n}\sum_{i=1}^{n} -\lg(\mathrm{P}_{\text{value}_i}) \quad (\mathrm{P}_{\text{value}_i} < T), \qquad (6)$$

where $T$ is set to 0.01 as mentioned above. Fig. 4 shows the P-score of predicted overlaps by various algorithms. It is observed that WN-PC and COACH obtain higher P-scores than the other methods. The result is also consistent with the comparison result of OMMR and essential proteins, which verifies the effectiveness of OMMR.

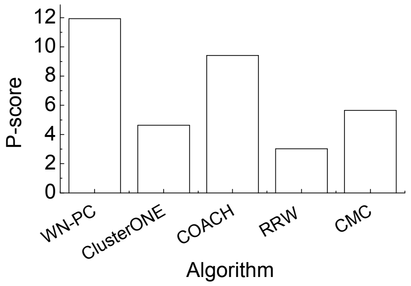

**Fig. 4 P-scores of predicted overlaps by various algorithms**

### 2.6 MIPS gold standard

In this section, we will analyze the results using the reference set derived from MIPS (Mewes *et al.*, 2006) instead of CYC2008. This benchmark set consists of 428 complexes. We test all the algorithms mentioned above on Gavin 2006 data. Table 8 contains the detailed results, where the MIPS set was used as a gold standard.

There are a greater number of overlaps in MIPS than CYC2008. In Table 8, it is obvious that three methods with the best performance are WN-PC, COACH, and ClusterONE, when MIPS instead of CYC2008 was adopted as a gold standard.

### 2.7 Results using Krogan and BioGRID data

To further investigate the quality of overlaps predicted by various methods, we also run these algorithms on Krogan data and BioGRID data. The matching results of each algorithm on Krogan data and BioGRID are shown in Table 9. The RRW method cannot run on the BioGRID network because of memory exception.

**Table 8 OMMR of overlaps predicted by various algorithms using the MIPS gold standard**

| Algorithm | Number of matched overlaps | Average precision | OMMR |
|---|---|---|---|
| WN-PC | 617 | 0.747 | 0.442 |
| ClusterONE | 357 | 0.506 | 0.144 |
| COACH | 528 | 0.674 | 0.367 |
| RRW | 188 | 0.305 | 0.038 |
| CMC | 321 | 0.487 | 0.127 |

Comparing Table 9 with Tables 3 and 8, we can see that algorithms achieve a higher performance using Krogan and BioGRID data than Gavin data, due to their reliabilities. Results for Krogan and BioGRID data are also consistent with results for Gavin data. The top three for the performance using Krogan and BioGRID data are WN-PC, COACH, and ClusterONE, which accords with the result for Gavin and verifies the effectiveness of OMMR.

**Table 9 OMMR of various algorithms using Krogan and BioGRID data**

| Algorithm | Number of matched overlaps | Average precision | OMMR |
|---|---|---|---|
| WN-PC$^{a,1}$ | 118 | 0.933 | 0.719 |
| ClusterONE$^{a,1}$ | 79 | 0.907 | 0.506 |
| COACH$^{a,1}$ | 106 | 0.917 | 0.619 |
| RRW$^{a,1}$ | 57 | 0.762 | 0.161 |
| CMC$^{a,1}$ | 52 | 0.769 | 0.462 |
| WN-PC$^{a,2}$ | 677 | 0.811 | 0.509 |
| ClusterONE$^{a,2}$ | 312 | 0.606 | 0.214 |
| COACH$^{a,2}$ | 631 | 0.724 | 0.418 |
| RRW$^{a,2}$ | 237 | 0.244 | 0.052 |
| CMC$^{a,2}$ | 370 | 0.562 | 0.181 |
| WN-PC$^{b,1}$ | 210 | 0.975 | 0.946 |
| ClusterONE$^{b,1}$ | 46 | 0.872 | 0.451 |
| COACH$^{b,1}$ | 201 | 0.966 | 0.892 |
| CMC$^{b,1}$ | 172 | 0.910 | 0.675 |
| WN-PC$^{b,2}$ | 870 | 0.908 | 0.782 |
| ClusterONE$^{b,2}$ | 345 | 0.569 | 0.209 |
| COACH$^{b,2}$ | 827 | 0.820 | 0.632 |
| CMC$^{b,2}$ | 682 | 0.643 | 0.355 |

$^{a}$ Using Krogan data; $^{b}$ using BioGRID data. $^{1}$ Using benchmark set CYC2008; $^{2}$ using benchmark set MIPS

## 3 Conclusions

Protein complexes are the basic units of macromolecular organizations and help us to understand the cell's mechanism. Recent developments in experiments supply a large amount of protein-protein interaction data, which provides a stepping stone for finding protein complexes. Some overlapping protein complexes prediction methods have been developed. Our previous research shows that overlapping protein complexes, especially their overlaps, play important roles in identifying essential proteins. In this paper, we propose a measure called OMMR to evaluate overlaps of essential modules. The experimental results indicate the importance of overlaps and reveal the relationship between overlaps and identification of essential proteins.

## References

Adamcsek, B., Palla, G., Farkas, I.J., *et al.*, 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**(8):1021-1023. [doi:10.1093/bioinformatics/btl039]

Bader, G.D., Hogue, C.W.V., 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.*, **4**:2.1-2.27. [doi:10.1186/1471-2105-4-2]

Boyle, E.I., Weng, S., Gollub, J., *et al.*, 2004. GO::TermFinder—open source software for accessing gene ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**(18):3710-3715. [doi:10.1093/bioinformatics/bth456]

Chen, B., Shi, J., Zhang, S., *et al.*, 2013. Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy. *Proteomics*, **13**(2):269-277. [doi:10.1002/pmic.201200336]

Cherry, J.M., Adler, C., Ball, C., *et al.*, 1998. SGD: Saccharomyces Genome Database. *Nucl. Acids Res.*, **26**(1):73-79. [doi:10.1093/nar/26.1.73]

Dezső, Z., Oltvai, Z.N., Barabási, A.L., 2003. Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. *Genome Res.*, **13**:2450-2454. [doi:10.1101/gr.1073603]

Enright, A.J., van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.*, **30**(7):1575-1584. [doi:10.1093/nar/30.7.1575]

Gavin, A.C., Aloy, P., Grandi, P., *et al.*, 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**:631-636. [doi:10.1038/nature04532]

Han, J.D., Bertin, N., Hao, T., *et al.*, 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, **430**:88-93. [doi:10.1038/nature02555]

Hart, G.T., Lee, I., Marcotte, E.M., 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinform.*, **8**:236.1-236.11. [doi:10.1186/1471-2105-8-236]

Hu, H., Yan, X., Huang, Y., *et al.*, 2005. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, **21**(suppl 1):i213-i221. [doi:10.1093/bioinformatics/bti1049]

Jiang, P., Singh, M., 2010. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, **26**(8):1105-1111. [doi:10.1093/bioinformatics/btq078]

Krogan, N., Cagney, G., Yu, H., *et al.*, 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**:637-643. [doi:10.1038/nature04670]

Lei, X., Wu, S., Ge, L., *et al.*, 2013. Clustering and overlapping modules detection in PPI network based on IBFO. *Proteomics*, **13**(2):278-290. [doi:10.1002/pmic.201200309]

Leung, H.C.M., Xiang, Q., Yiu, S.M., *et al.*, 2009. Predicting protein complexes from PPI data: a core-attachment approach. *J. Comput. Biol.*, **16**(2):133-144. [doi:10.1089/cmb.2008.01TT]

Liu, G., Wong, L., Chua, H.N., 2009. Complex discovery from weighted PPI networks. *Bioinformatics*, **25**(15):1891-1897. [doi:10.1093/bioinformatics/btp311]

Macropol, K., Can, T., Singh, A.K., 2009. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinform.*, **10**:283.1-283.10. [doi:10.1186/1471-2105-10-283]

Maraziotis, I.A., Dimitrakopoulou, K., Bezerianos, A., 2007. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinform.*, **8**:408.1-408.15. [doi:10.1186/1471-2105-8-408]

Mewes, H.W., Frishman, D., Mayer, K.F.X., *et al.*, 2006. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucl. Acids Res.*, **34**(suppl 1):D169-D172. [doi:10.1093/nar/gkj148]

Nepusz, T., Yu, H., Paccanaro, A., 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**(5):471-472. [doi:10.1038/nmeth.1938]

Ni, W.Y., Xiong, H.J., Zhao, B.H., *et al.*, 2013. Predicting overlapping protein complexes in weighted interactome networks. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(10):756-765. [doi:10.1631/jzus.C13b0097]

Palla, G., Derényi, I., Farkas, I., *et al.*, 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**:814-818. [doi:10.1038/nature03607]

Pu, S., Wong, J., Turner, B., *et al.*, 2009. Up-to-date catalogues of yeast protein complexes. *Nucl. Acids Res.*, **37**(3):825-831. [doi:10.1093/nar/gkn1005]

Shih, Y.K., Parthasarathy, S., 2012. Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics*, **28**(18):i473-i479. [doi:10.1093/bioinformatics/bts370]

Stark, C., Breitkreutz, B.J., Reguly, T., *et al.*, 2006. BioGRID: a general repository for interaction datasets. *Nucl. Acids Res.*, **34**(suppl 1):D535-D539. [doi:10.1093/nar/gkj109]

Wu, M., Li, X., Kwoh, C.K., *et al.*, 2009. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinform.*, **10**:169.1-169.16. [doi:10.1186/1471-2105-10-169]

Zhang, R., Lin, Y., 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucl. Acids Res.*, **37**(suppl 1):D455-D458. [doi:10.1093/nar/gkn858]

Zhao, B., Wang, J., Li, M., *et al.*, 2014. Prediction of essential proteins based on overlapping essential modules. *IEEE Trans. NanoBiosci.*, **13**(4):415-424. [doi:10.1109/TNB.2014.2337912]