

## RESEARCH ARTICLE

# On the statistical significance of protein complex

Youfu Su<sup>1</sup>, Can Zhao<sup>1</sup>, Zheng Chen<sup>1</sup>, Bo Tian<sup>1</sup> and Zengyou He<sup>1,2,\*</sup>

<sup>1</sup> School of Software, Dalian University of Technology, Dalian 116024, China

<sup>2</sup> Key Laboratory for Ubiquitous Network and Service Software of Liaoning, Dalian 116024, China

\* Correspondence: zyhe@dlut.edu.cn

Received August 30, 2017; Revised March 15, 2018; Accepted April 15, 2018

**Background:** Statistical validation of predicted complexes is a fundamental issue in proteomics and bioinformatics. The target is to measure the statistical significance of each predicted complex in terms of  $p$ -values. Surprisingly, this issue has not received much attention in the literature. To our knowledge, only a few research efforts have been made towards this direction.

**Methods:** In this article, we propose a novel method for calculating the  $p$ -value of a predicted complex. The null hypothesis is that there is no difference between the number of edges in target protein complex and that in the random null model. In addition, we assume that a true protein complex must be a connected subgraph. Based on this null hypothesis, we present an algorithm to compute the  $p$ -value of a given predicted complex.

**Results:** We test our method on five benchmark data sets to evaluate its effectiveness.

**Conclusions:** The experimental results show that our method is superior to the state-of-the-art algorithms on assessing the statistical significance of candidate protein complexes.

**Keywords:** predicted complex; statistical significance testing; subgraph mining; community detection

**Author summary:** The detection of protein complexes is equivalent to finding interesting sub-networks from protein-protein interaction networks. One such interesting measure is the statistical significance of protein complexes in terms of  $p$ -values. Here a new yet simple  $p$ -value calculation method is presented, which can outperform existing methods consistently and significantly.

## INTRODUCTION

With the recent developments of high-throughput experimental technologies, people have collected many high-quality, large-scale protein-protein interaction (PPI) data sets and interaction networks [1–3]. Protein complex is a group of proteins which interact with each other at the same location and time. It is a unit to carry out metabolic functions in the cell or tissue [4]. Hence, detecting protein complexes is a fundamental problem in bioinformatics and proteomics research, which is of primary importance to gain a better understanding on the hierarchy and organization of biological processes and cellular components [5].

To date, many computational approaches have been

proposed to detect protein complexes. The details of these existing methods and the challenges of protein complex detection have been summarized and discussed in many reviews from different angles [6–10]. These methods could be classified into two categories according to whether the overlap between protein complexes is allowed. In the first category, there are no overlapping proteins between the predicted complexes. However, one protein may belong to more than one complex to perform multiple functions, so algorithms in the second category have been proposed to cope with this problem for detecting overlapping protein complexes [3,5,11,12].

Due to the lack of consensus on the definition of protein complex, formulating metrics for evaluating the quality of

a complex is also a challenging task. Many attempts have been made to evaluate the goodness of a protein complex. Based on whether using additional reference database, the evaluation methods can be classified into two categories. In the first category, the predicted result and the gold standard dataset are used as the input, *i.e.*, ground-truth protein complexes are assumed to be known in these methods [13,14]. If the database is complete and all complexes are valid, this type of methods will be accurate. However, most databases are not complete and some errors might be included. Alternatively, evaluation methods in the second class assess the detected complexes using only the information encoded in the target PPI network. Towards this direction, many metrics have been proposed, which provide a numeric value for evaluating the goodness of each protein complex or all detected protein complexes. However, most of these methods do not address the issue of statistical significance of protein complexes. In other words, how to evaluate if the protein complex obtained by an algorithm is real based on rigorous statistical significance testing procedures. Such testing-based approaches provide at least one major advantage over the other methods in that the results can be quantified in terms of the  $p$ -value, which is a universally understood measure between 0 and 1. In contrast, quantitative numerical values generated from other approaches are generally data-dependent, which is hard for people to interpret and determine a universal threshold across all data sets.

During the past decades, the issue of testing the statistical significance of protein complexes has only been discussed in a few papers [15–18]. Even in the context of community detection and dense subgraph mining, this issue is less addressed as well [19,20]. In this paper, we focus on the assessment of statistical significance of each predicted protein complex.

To date, the definition of the statistical significance of a single complex is still not clear. Some methods used the  $p$ -value of a one-sided or two-sided Mann-Whitney U test performed on the in-weights and out-weights of the vertices of a subgraph (*i.e.*, a protein complex) in order to quantify its significance [21,22]. Spirin and Mirny calculated the  $p$ -value through generating 1,000 random graphs in which the number of interactions for each protein is preserved in the null model [18]. OSLOM [17] used the probability of finding the cluster in a random graph generated by the specified configuration model to quantify the significance of a single complex. SiDeS [16] considered two reference models for quantifying the behavior of identified complexes and deriving significance measures. Overall, these existing methods mainly differ from each other with respect to their underlying assumptions, random null graph models and  $p$ -value

calculation methods.

In this paper, we propose a novel method for assessing the statistical significance of each candidate protein complex. The null hypothesis is that there is no difference between the number of edges in target protein complex and that in the random null model. Accordingly, the  $p$ -value is defined as the probability that we can find complexes with the same set of vertices in the random graph that are more dense than the target complex to be evaluated. We use the results generated by ClusterONE [3] and MDS [5] to test our evaluation method and some other methods on five datasets. Experimental results show that our method has better performance than the other methods.

The rest of this article is organized as follows. First, the Section of Results presents the experimental results. Some conclusions are provided subsequently. At last, we describe how to formulate the significance measure to evaluate protein complexes in detail.

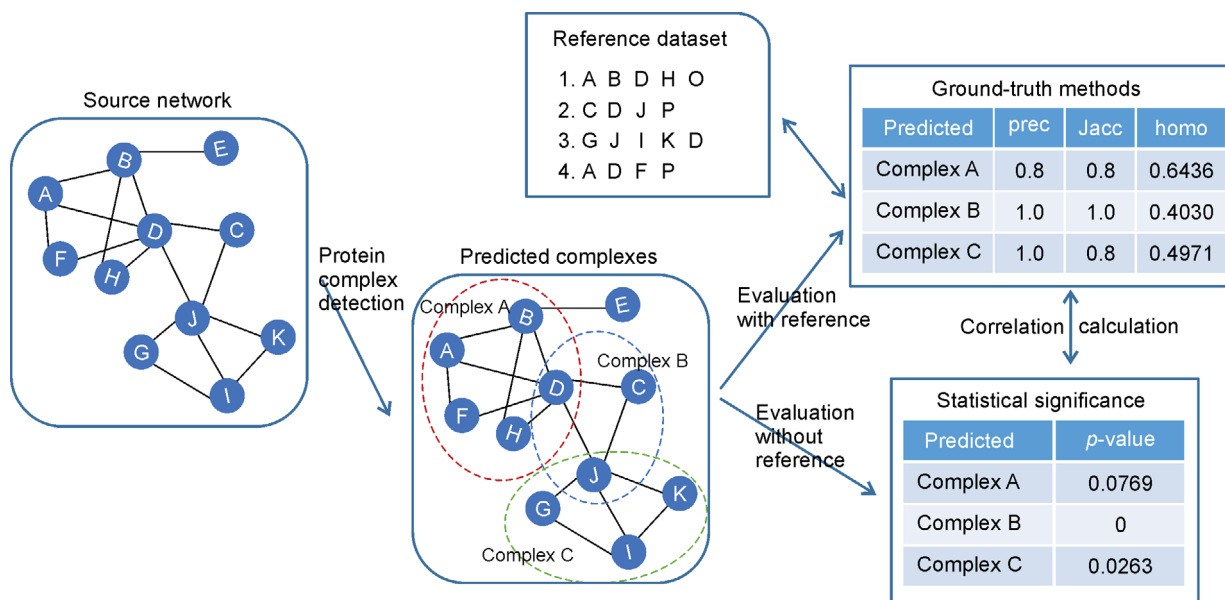
## RESULTS

We apply our algorithm to five large scale yeast PPI data sets using two protein complex detection methods and compare its performance with five existing methods. The main workflow of our experiments is shown in Figure 1. Firstly, we could obtain the predicted protein complexes through some protein complex detection methods (ClusterONE and MDS are used in the experiments), and then, we evaluate the predicted results with and without reference data sets of ground-truth complexes. Finally, we calculate the Pearson's correlation coefficient between the  $p$ -values (from our algorithm and other competing algorithms) and the goodness scores (from evaluation methods that use ground-truth protein complexes).

### Data sets

In the experiment, we use five data sets: Collins [23], Gavin [2], Krogan\_core [24], Krogan\_extend [24], and BioGRID [25]. The PPI networks in the first four data sets are weighted networks, we transform these networks into unweighted ones. More details on these data sets can be found in [3]. Note that all the input networks were stripped from self-interactions and isolated proteins.

We compare the detected complexes with two ground-truth complex sets: MIPS and SGD. The former one contains every MIPS ComplexCat category with at least 3 and at most 100 members. The golden standard protein complexes in SGD were derived from the Gene Ontology (GO) annotations in the *Saccharomyces Genome Database*.



**Figure 1.** The main workflow of our experiments. The experimental procedure contains three major steps: protein complex detection, assessing the predicted complexes and correlation calculation.

## Parameter setting

We choose ClusterONE [3] and MDS [5] as the protein complex detection methods, and their software packages are publicly available. We run these two methods with their default parameter setting.

## Experiment

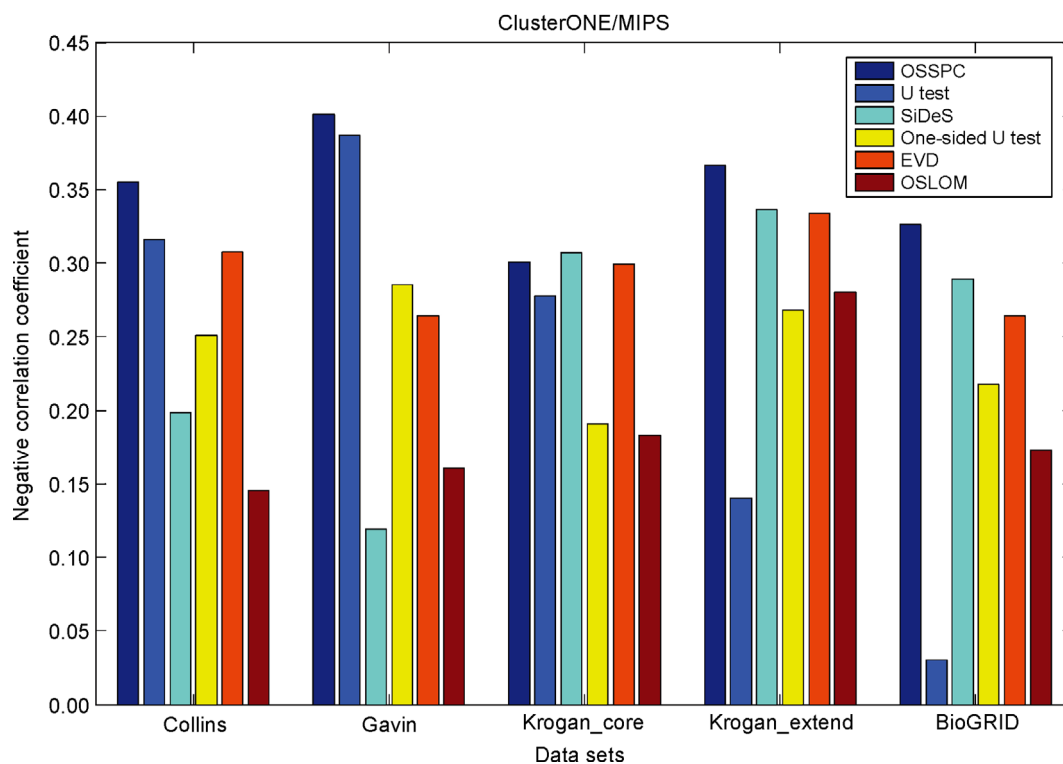
Given the reference set of ground-truth protein complexes, we are able to calculate the precision, Jaccard similarity coefficient and homogeneity for each complex returned from a protein complex detection method. We use the sum of these three measures to generate a new synthetic measure as the “ground-truth” performance measure (denoted by  $g$ -value). To check the performance of a significance testing algorithm in which ground-truth protein complexes are assumed be unknown and not used, we can calculate the Pearson’s correlation coefficient between  $p$ -values and  $g$ -values. Obviously, if one algorithm has better performance, it will have lower negative correlation coefficient values. For the ease of illustration and description, we will use the negative of Pearson’s correlation coefficient as the final performance indicator in the subsequent parts. That is, larger correlation values indicate better performance.

We compare our method (denoted by OSSPC) with five methods in the literature: Mann-Whitney U test [21], SiDeS [16], one-sided Mann-Whitney U test [22], the method based on FisherTippett extreme value distribution

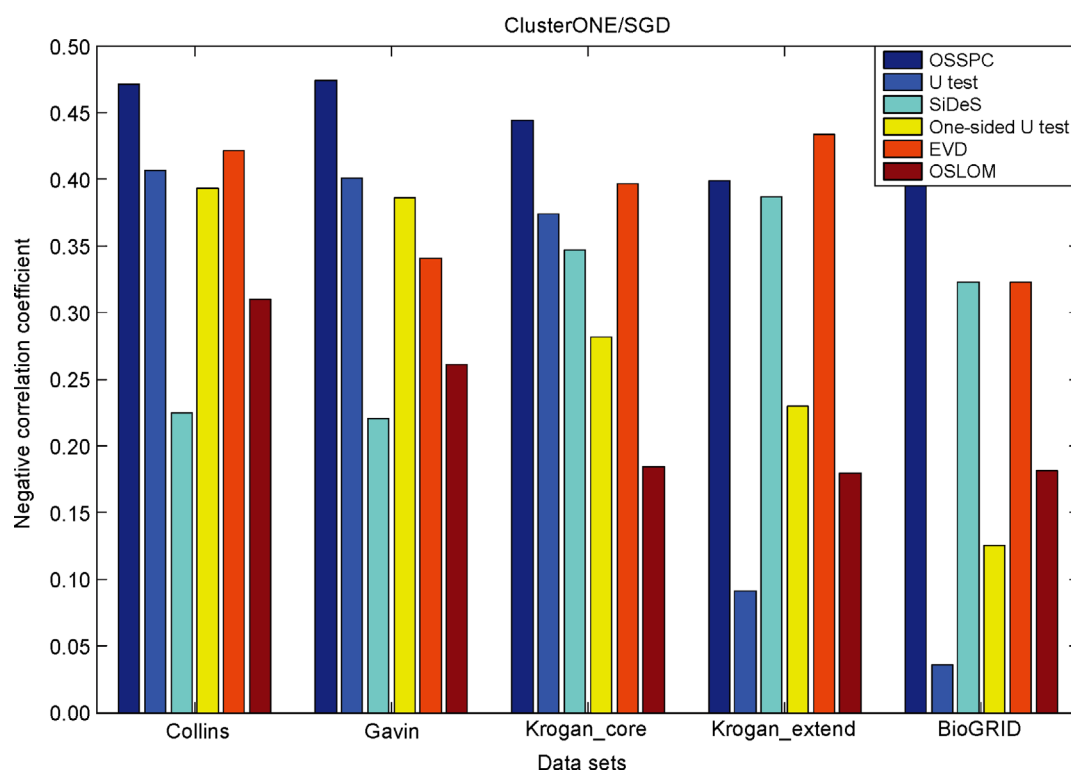
(EVD) in Ref. [18] and OSLOM [17] with their default parameters. The comparison results are presented in Figures 2–5, where ClusterONE is used as the protein complex detection method in the first two figures, and MDS is used as protein complex detection method in the last two figures.

Figures 2 and 3 show the comparison results when the reported complexes by ClusterONE are used as the input. From the comparison results in the two figures, we have the following observations. Firstly, OSSPC performs the best 4 times in both Figures 2 and 3. This indicates that our method can achieve better performance than other statistical significance testing methods in the case that MIPS is used as the gold standard. Secondly, our method performs the second best on the Krogan\_core data set in Figure 2. Note that SiDeS has the best performance on the Krogan\_core data set, which is only slighter better than our method. Therefore, our method could provide a more reasonable  $p$ -value to quantify the statistical significance of a protein complex.

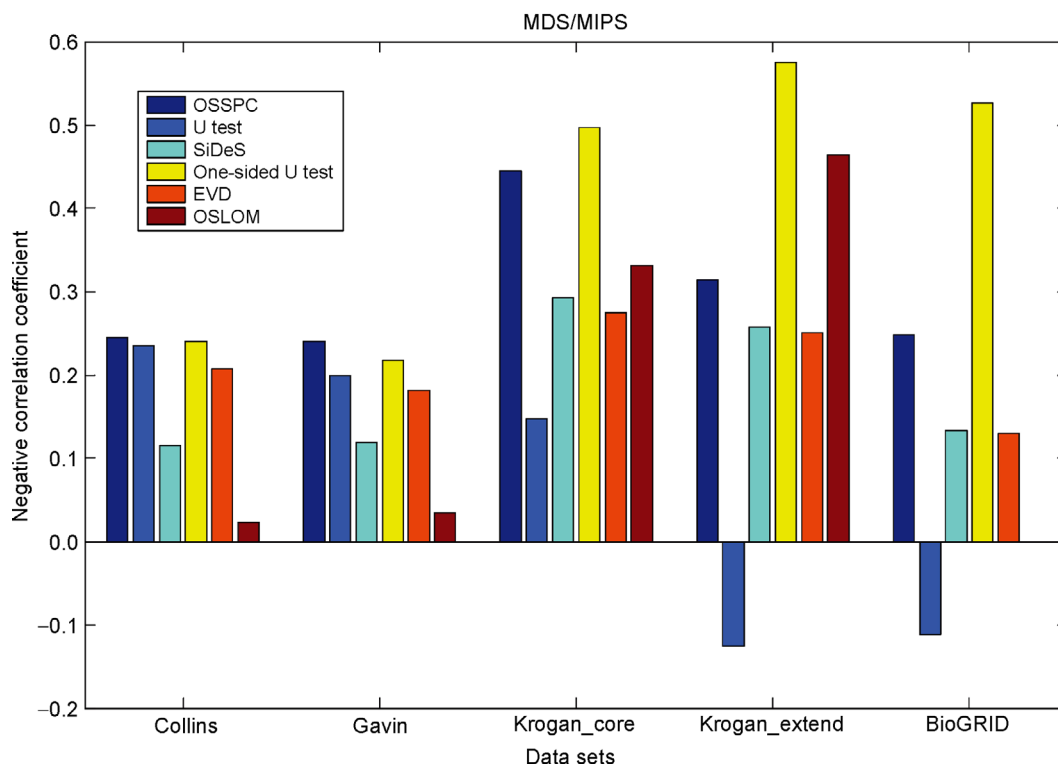
Figures 4 and 5 show the comparison results when the reported complexes by MDS are used as the input. In this experiment, the result of OSLOM on BioGRID is not included since it cannot finish within 12 hours. It can be observed that OSSPC achieves the highest value on two data sets and the second highest value on other two data sets in Figure 4. In general, our method does not always perform the best when we use MDS as the protein complex detection method. Anyway, our method at least has comparable performance to the competing algorithms



**Figure 2.** The performance comparison of different significance testing algorithms. Here we use ClusterONE as the protein complex detection method and MIPS as the golden standard.



**Figure 3.** The performance comparison of different significance testing algorithms. Here we use ClusterONE as the protein complex detection method and SGD as the golden standard.



**Figure 4.** The performance comparison of different significance testing algorithms. Here we use MDS as the protein complex detection method and MIPS as the golden standard.

when MIPS is used as the golden standard.

It is necessary to explain why the performance of our method equipped with MDS is worse than that of using ClusterONE. The reason is probably that the way for computing the  $p$ -value might favor some protein complex detection algorithms like ClusterONE.

The above experiments show that the use of different PPI networks will lead to different evaluation results as well. That is, different significance testing algorithms yield different behaviors across different networks. Furthermore, different golden standard sets of complexes can affect the evaluation results as well.

## CONCLUSIONS

In this paper, we tackle the problem of evaluating a single protein complex from a statistical significance perspective and propose a novel method for calculating the  $p$ -value. According to experimental results, our method has better performance than those competitive methods.

In the future work, we will incorporate the proposed significance testing procedure into the protein complex detection process for mining statistically significant complexes.

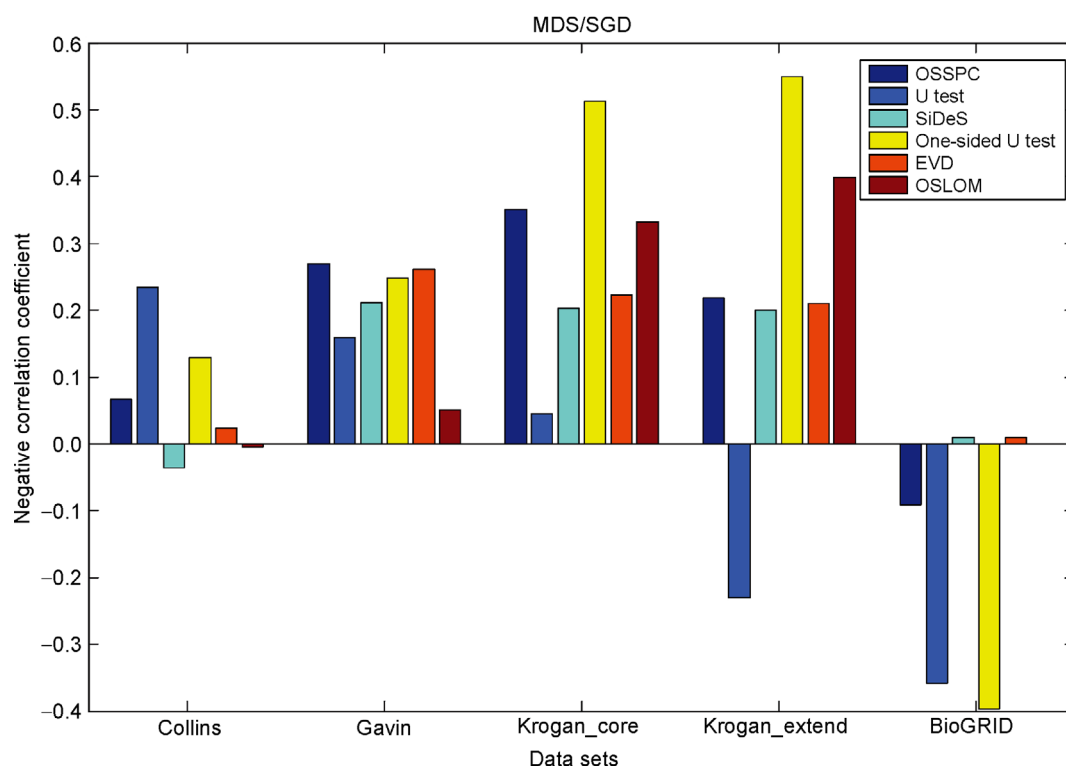
## MATERIALS AND METHODS

### Problem definition

Let  $G(V, E)$  be an undirected graph with a vertex set  $V$  and an edge set  $E \subseteq V \times V$ . The graph is assumed to be simple (without self loop or parallel edges). Let  $S = \{S_i | S_i \subseteq V\}$  be the set of protein complexes generated from a protein complex detection method. The problem to be solved is to design a  $p$ -value calculation method to accurately measure the statistical significance of each given protein complex  $S_i$ .

### Statistical significance

The null hypothesis in our model is that there is no difference between the number of edges in the target protein complex and that in the random null model. Note that a true protein complex must be a connected subgraph. Hence, we only consider connected subgraphs in the corresponding random graphs as well. The number of possible connected subgraphs with  $n$  vertices and  $k$  edges can be calculated using the following formula [26,27]:



**Figure 5.** The performance comparison of different significance testing algorithms. Here we use MDS as the protein complex detection method and SGD as the golden standard.

$$q_{n,k} = \begin{cases} 0, & \text{if } k < n-1 \text{ or } k > n(n-1)/2, \\ n^{n-2}, & \text{if } k = n-1, \text{ and otherwise,} \\ \binom{n(n-1)/2}{k} - \sum_{m=0}^{n-2} \binom{n-1}{m} \sum_{p=0}^k \binom{(n-1-m)(n-2-m)/2}{p} q_{m+1,k-p}. \end{cases} \quad (1)$$

The  $p$ -value is the probability that we can find subgraphs with the same set of vertices in the random graphs that are denser than the target complex. So the statistical significance in terms of  $p$ -value for a given complex can be expressed as:

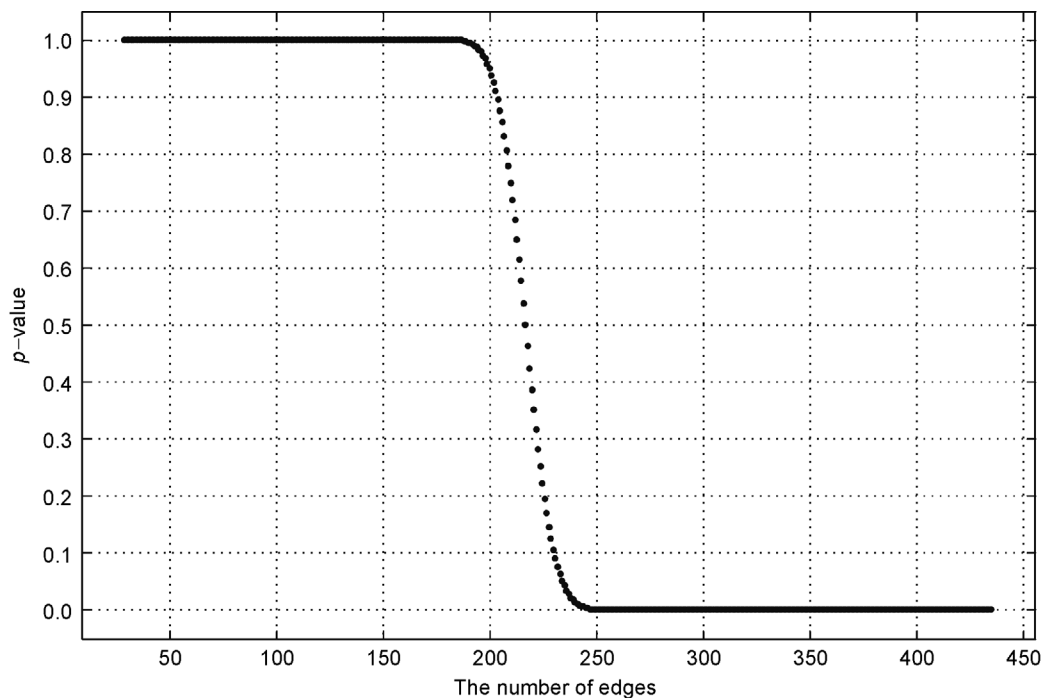
$$p\text{-value} = \frac{q_{n,e+1} + q_{n,e+2} + \cdots + q_{n,n(n-1)/2}}{\sum_{i=n-1}^{n(n-1)/2} q_{n,i}}, \quad (2)$$

where  $n$  is the number of vertices and  $e$  is the number of edges in the target complex. For example, given a protein complex with 5 vertices and 8 edges, there will be 728 connected random subgraphs based on the same set of 5 vertices. Among these random subgraphs, there are 11 subgraphs that has more than 8 edges: one subgraph with 10 edges and 10 subgraphs that have 9 edges. So the  $p$ -value of this complex is  $(10 + 1)/728 = 0.0151$ .

## Function fitting

Since the Equation (2) is too complicated and time-consuming to compute in a straightforward manner when the number of vertices is large, we have to find a fast method to compute the  $p$ -value approximately. When the number of vertices is less than 10, we will still use Equation (2) to calculate the  $p$ -value. Otherwise, we will adopt some data fitting methods to construct a function to estimate this value. As shown in Figure 6, the  $p$ -values with the different numbers of edges followed a “S” shape’s distribution when the number of vertices  $n$  in the subgraph is fixed. Therefore, we introduced a sigmoid function to fit the data:

$$f(x) = a + \frac{b}{1 + e^{-\frac{(x-t)}{s}}}. \quad (3)$$



**Figure 6.** The  $p$ -value curve. We choose a subgraph with 30 vertices, where the number of edges is ranged from 29 to 435.

There are 4 parameters to be estimated:  $\beta = \{a, b, t, s\}$ . And the parameters  $t(n)$  and  $s(n)$  are derived as a function of complex size  $n$ . Given a fixed number of vertices and a set of  $l$  pairs:  $\{(x_i, y_i) | 1 \leq i \leq l\}$ , where  $x_i$  is the number of edges and  $y_i$  is the corresponding  $p$ -value, we can estimate  $\beta$  by minimizing the sum of the squares of the deviations  $H(\beta)$ :

$$H(\beta) = \sum_{i=1}^l [y_i - f(x_i, \beta)]^2. \quad (4)$$

Therefore, we can get a series of  $H(\beta)$ s by varying the number of vertices. The Levenberg–Marquardt algorithm (LMA) is an iterative procedure, which is usually used to solve non-linear least squares problems. In each iteration step, the parameter vector,  $\beta$ , is replaced by a new estimate,  $\beta + \delta$ . To determine  $\delta$ , the functions  $f(x_i, \beta + \delta)$  are approximated by their linearizations:

$$f(x_i, \beta + \delta) \approx f(x_i, \beta) + J_i \delta, \quad (5)$$

where  $J_i = \frac{\partial f(x_i, \beta)}{\partial \beta}$  is the gradient of  $f$  with respect to  $\beta$ .

More details on LMA can be found in Refs. [28,29].

Through solving a series of formulas, the parameters of in Equation (3) are specified as follows, when the number of vertices is no less than 10:

$$a = 1, \quad b = 1,$$

$$t = \frac{n(n-1)-2}{4}, \quad s = \frac{n}{5}.$$

So the  $p$ -value can be calculated approximately using the following expression:

$$p\text{-value} \approx 1 - \frac{1}{1 + e^{-5 \frac{4x - n(n-1) + 2}{4n}}}. \quad (6)$$

To ensure the accuracy of the  $p$ -value, Equation (6) is used when the number of vertices is no less than 10. Otherwise, we will adopt Equation (2) to calculate the  $p$ -value of the predicted complex.

## AUTHOR CONTRIBUTIONS

Youfu Su drafted the manuscript. Zheng Chen and Bo Tian performed the implementations. Can Zhao and Zengyou He conceived the study and finalized the manuscript. All authors read and approved the final manuscript.

## ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (No. 61572094), the Fundamental Research Funds for the Central Universities of China (Nos. DUT2017TB02 and DUT14QY07). Additionally, we want to thank the academic support received from Mr. Ben Teng and Dr. Xiuli Ma.

# COMPLIANCE WITH ETHICS GUIDELINES

The authors Youfu Su, Can Zhao, Zheng Chen, Bo Tian and Zengyou He declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

# REFERENCES

1. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403, 623–627
2. Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440, 631–636
3. Nepusz, T., Yu, H. and Paccanaro, A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, 9, 471–472
4. Teng, B., Zhao, C., Liu, X. and He, Z. (2015) Network inference from AP-MS data: computational challenges and solutions. *Brief. Bioinform.*, 16, 658–674
5. Ma, X., Zhou, G., Shang, J., Wang, J., Peng, J. and Han, J. (2017) Detection of complexes in biological networks through diversified dense subgraph mining. *J. Comput. Biol.*, 24, 923–941
6. Chen, B., Fan, W., Liu, J. and Wu, F.-X. (2014) Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Brief. Bioinform.*, 15, 177–194
7. Ji, J., Zhang, A., Liu, C., Quan, X. and Liu, Z. (2014) Survey: functional module detection from protein-protein interaction networks. *IEEE Trans. Knowl. Data Eng.*, 26, 261–277
8. Li, X., Wu, M., Kwok, C.-K. and Ng, S.-K. (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, 11, S3
9. Wang, J., Li, M., Deng, Y. and Pan, Y. (2010) Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, 11, S10
10. Bhowmick, S. S. and Seah, B. S. (2016) Clustering and summarizing protein-protein interaction networks: a survey. *IEEE Trans. Knowl. Data Eng.*, 28, 638–658
11. Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I. and Vicsek, T. (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22, 1021–1023
12. Palla, G., Derényi, I., Farkas, I. and Vicsek, T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435, 814–818
13. Brohée, S. and van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7, 488
14. Song, J. and Singh, M. (2009) How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, 25, 3143–3150
15. Traag, V. A., Krings, G. and Van Dooren, P. (2013) Significant scales in community structure. *Sci. Rep.*, 3, 2930
16. Koyutürk, M., Szpankowski, W. and Grama, A. (2007) Assessing significance of connectivity and conservation in protein interaction networks. *J. Comput. Biol.*, 14, 747–764
17. Lancichinetti, A., Radicchi, F., Ramasco, J. J. and Fortunato, S. (2011) Finding statistically significant communities in networks. *PLoS One*, 6, e18961
18. Spirin, V. and Mirny, L. A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA*, 100, 12123–12128
19. Chakraborty, T., Dalmia, A., Mukherjee, A. and Ganguly, N. (2017) Metrics for community analysis: A survey. *ACM Comput. Surv.*, 50, 1–37
20. Zhang, P. and Moore, C. (2014) Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proc. Natl. Acad. Sci. USA*, 111, 18144–18149
21. Csardi, G. and Nepusz, T. (2006) The Igraph software package for complex network research. *Inter. Journal Complex Systems*, 1695, 1–9
22. Nepusz, T., Yu, H. and Paccanaro, A. Clusterone cytoscape plugin. [http://www.paccanarolab.org/static\\_content/clusterone/c1l-cytoscape3-1.0.html](http://www.paccanarolab.org/static_content/clusterone/c1l-cytoscape3-1.0.html)
23. Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S. and Krogan, N. J. (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics*, 6, 439–450
24. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440, 637–643
25. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) Biogrid: a general repository for interaction datasets. *Nucleic Acids Res.* 34, Suppl 1, D535–D539
26. “How many connected graphs over v vertices and e edges?” <http://math.stackexchange.com/questions/689526/how-many-connected-graphs-over-v-vertices-and-e-edges>
27. Shor, P. W. (1995) A new proof of cayley’s formula for counting labeled trees. *J. Com. Theory*, 71, 154–158
28. Marquardt, D. W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.*, 11, 431–441
29. Moré, J. (1977) The levenberg–marquardt algorithm: Implementation and theory. In *Conference on Numerical Analysis*. Dundee, UK