IDENTIFICATION OF ESSENTIAL PROTEINS USING A NOVEL MULTI-OBJECTIVE OPTIMIZATION METHOD

Chong Wu[⋆,*] *Houwang Zhang*[†] *Le Zhang*[‡] *Hanying Zheng*[†]

[⋆] City University of Hong Kong, Department of Electrical Engineering, Kowloon, Hong Kong
[†] China University of Geosciences, School of Automation, Wuhan, China
[‡] Tongji University, Department of Computer Science and Technology, Shanghai, China

ABSTRACT

Using graph theory to identify essential proteins is a hot topic at present. These methods are called network-based methods. However, the generalization ability of most network-based methods is not satisfactory. Hence, in this paper, we consider the identification of essential proteins as a multi-objective optimization problem and use a novel multi-objective optimization method to solve it. The optimization result is a set of Pareto solutions. Every solution in this set is a vector which has a certain number of essential protein candidates and is considered as an independent predictor or voter. We use a voting strategy to assemble the results of these predictors. To validate our method, we apply it on the protein-protein interactions (PPI) datasets of two species (Yeast and Escherichia coli). The experiment results show that our method outperforms state-of-the-art methods in terms of sensitive, specificity, F-measure, accuracy, and generalization ability.

Index Terms— Essential proteins, graph theory, multi-objective optimization, protein-protein interactions.

1. INTRODUCTION

To living organisms, proteins are indispensable components in cellular life activities, they perform varied functions like catalyzing metabolic, DNA replication reactions, and transporting molecules [1]. Among them, there is a kind of proteins called essential proteins, living organisms would die or be infertile if they lack them [2]. Some essential proteins have been found to be related to human disease genes, hence, the study of the identification of essential proteins is very necessary [3].

With the development of high-throughput technologies, a lot of protein-protein interactions (PPI) have been obtained which makes using computational methods to study the identification of essential proteins possible [4]. In general, protein-protein interactions are constructed to an undirected network which is called protein interaction network (PIN). Network-based methods are successfully used in the identification of essential proteins. According to wether they integrate the biological information or not, they can be divided into two classes: (1) topological characteristics based methods; (2) integration methods.

The topological characteristics based methods use the features of node or edge of network to search the vital nodes, and they are also widely used in the field of complex networks. Degree centrality (DC) is the most well-known and simplest one applied on the identification of essential proteins. Some studies have confirmed that proteins with high degree tend to be essential proteins [5]. Besides DC, other node-aided methods were also applied to the identification of essential proteins, such as eigenvector centrality (EC) [6], betweenness centrality (BC) [7], closeness centrality (CC) [8], *etc.* Additionally, a few of edge-aided methods [9, 10, 11] have also been proposed to identify essential proteins from PIN, the typical one is edge clustering coefficient (ECC) [10]. A centrality method which is based on ECC called new centrality method (NC) is proposed to identify essential proteins from PIN [11]. Besides above node-aided and edge-aided methods, some researchers proposed a centrality method which combines the node and the edge characteristics of the network (NEC) [12].

Whereas, the PPI data obtained by high-throughput technologies have high false positives and these topological characteristics based methods are very sensitive to the stability of the PIN which consists of the PPI data. Hence, the performance of these methods is limited. Considering this problem, some researchers try to combine the biological information into the topological characteristics based methods to reduce the effect of high false positives of PPI data and improve the prediction accuracy of essential proteins. Some researchers proposed PeC which combines gene expression data into the NC and achieves a higher prediction accuracy than NC [13].

However, the generalization ability of these methods above is not good. In this paper, to improve the generalization ability, we consider the identification of essential proteins as a multi-objective optimization problem and use the adaptive multi-objective black hole algorithm (AMOBH) [14] to solve it, the new method is called IMAMOBH. After the optimization, we get a Pareto solution set. Each solution in this set will be considered as a predictor or voter. Each predictor

[*] Corresponding author, e-mail: chongwu2-c@my.cityu.edu.hk.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

will give a list of essential protein candidates. Then we use a voting mechanism to assemble them and get a final list of essential protein candidates. To validate our method, we select two species' PPI datasets (Yeast and Escherichia coli) and apply IMAMOBH on them. In the comparison experiments, our method achieves better results compared to some state-of-the-art methods like BC, CC, DC, EC, LAC[15], NC, NEC, PageRank[16], and PeC. The contributions of this paper can be concluded as follows,

- This is the first attempt of applying multi-objective optimization into the identification of essential proteins.
- A method with satisfactory generalization ability is proposed for the identification of essential proteins.

2. MATERIALS

The PPI data of Saccharomyces cerevisiae (Yeast) and Escherichia coli are downloaded from the DIP [17] database. The PPI dataset of Yeast contains 4,979 proteins and 22,061 interactions. The PPI dataset of Escherichia coli owns 2528 proteins and 11496 interactions.

The essential genes lists of Yeast and Escherichia coli are collected from OGEE [18]. The Yeast network consists of 1,209 essential proteins, 3,322 nonessential proteins, and 448 unknown proteins. The Escherichia coli network consists of 444 essential proteins, 1403 nonessential proteins, and 681 unknown proteins.

The gene expression datasets of Yeast and Escherichia coli are downloaded from GEO [19]. We use the Pearson correlation coefficient (PCC) to evaluate the gene expression similarity (GES) of two interacting proteins [13]. The gene ontology data used in this paper is collected from paper [9]. GO semantic similarity is based on the biological characteristics of genes. It is used to represent the genes functional similarity [20]. Using biological process category of GO, genes functional similarity (GFS) between two proteins can be calculated by the algorithm proposed in paper [21].

3. METHOD

3.1. Identification of Essential Proteins using Adaptive Multi-objective Black Hole Algorithm

The identification of essential proteins can be considered as a multi-objective problem (MOP), which has two objectives: gene expression similarity (GES) and genes functional similarity (GFS), as follows,

$$f1 = \sum_{i=1}^{n} NTE(i) * GES(i), \tag{1}$$

$$f2 = \sum_{i=1}^{n} NTE(i) * GFS(i), \tag{2}$$

where, n is the number of elements in one solution, $NTE(i)$ is the number of triangles consist of a certain edge includes protein i. The objectives of our methods consist of network topological feature NTE and biological information like GES and GFS. Here we choose NTE because it is highly correlated with GES and GFS. Hence, we construct objective functions like this type.

To solve above MOP, we use the adaptive multi-objective black hole algorithm (AMOBH) [14] which has several advantages: lower computational complexity, faster convergence rate, and better population diversity compared to state-of-the-art methods. The Pareto solution set of above MOP is corresponding to a set of different weighted combination of two objectives. The optimization method will guarantee the diversity of solution. Hence, we can avoid the subjective selection of the weights of two objectives. To assemble the results of solutions in the Pareto solution set, we build a voting system to select a certain number of solutions to form the final essential protein candidates. Every solution in Pareto solution set is considered as a voter. The larger number of votes of a protein obtained means it has the bigger probability to be chosen into the final essential protein candidates.

The brief pseudo code of AMOBH is as Algorithm 1 show.

Algorithm 1 Adaptive multi-objective black hole algorithm

```
Input: Size of population $N$, size of archive $M$
Output: Pareto solution set $Ar$
  Initialize the population set $Pop$
  Calculate the fitness values of each solution on two objective functions
  Initialize the black holes $B_h$ and save them into the $Ar$
  repeat
    Population $Pop$ updates
    if $rand < l$ (learning rate) then
      Eilte $Bh(j)$ mutation
    end
    Determine whether to accept the new solution $Pop(i)$ to $Ar$ or not
    Update the evolution statues based on Shannon entropy
    if $\|Pop(i) - Bh(j)\|_2 < \theta_1$ (threshold) then
      Reinitialize $Pop(i)$
    end
  until $error < \theta_2$ (threshold)
```

The original AMOBH is used to solve the continuous MOPs. The solution update formula is as follow,

$$Pop_{t+1}(i) = Pop_t(i) + rand(Bh(j) - Pop_t(i)), \tag{3}$$

where, $Pop(i)$ represents the solution i, $t + 1$ means current iteration, t means previous iteration, and $Bh(j)$ means one of black holes (elite solutions) from the black hole set. However, here is a discrete MOP. The update rule needs to be changed.

The new solution update rule is shown in Fig 1. As Fig 1 shows, a solution is a vector. It will get close to a certain black hole. At first, we get the different parts of two vectors (eg. solution i and black hole j) and call them PartA from the black hole j and PartB from the solution i respectively. Then we select several elements from PartA randomly and use them to replace the same number of elements in PartB. The maximum of selecting elements is the size of PartA. After that, we will get a new solution i which is much similar to the black hole j as Fig 1 shows since there are more similar elements between two vectors. What's more, the order of an element in a solution vector is not important in this problem.

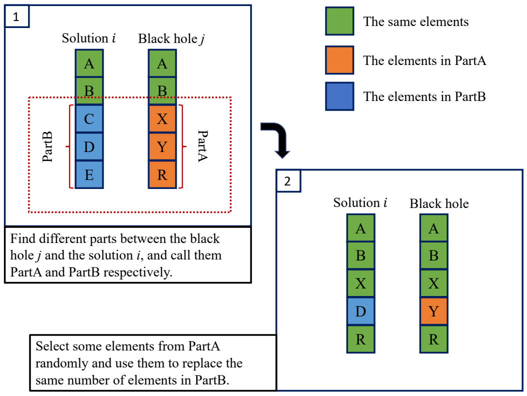

1
Solution *i*
Black hole *j*
A
B
C
D
E
A
B
X
Y
R
PartB
PartA
Find different parts between the black hole *j* and the solution *i*, and call them PartA and PartB respectively.
2
Solution *i*
Black hole
A
B
X
D
R
A
B
X
Y
R
Select some elements from PartA randomly and use them to replace the same number of elements in PartB.
The same elements
The elements in PartA
The elements in PartB

Fig. 1. The process of solution update. Green box represents the common element which is in both black hole j and solution i, orange box represents the unique element which is only in black hole j, and blue box represents the unique element which is only in solution i. After the update, the number of green boxes in two vectors becomes larger which means the similarity becomes higher.

After using AMOBH to solve above MOP, we will get a Pareto solution set Ar. Every Pareto solution represents a possible essential protein candidate list provided by a certain weighted combination of two objectives. It is considered as a voter. To assemble the results of different voters and maintain a good generalization ability, we adopt a voting strategy. If a protein i is in the Pareto solution j, it gets a vote from j. The more number of votes a solution obtained means it is more possible to be selected as an essential protein candidate.

3.2. Computational Complexity

The computation of core AMOBH algorithm is $O(MN^2)$. M is the number of objectives, and N is the size of archive. The values of $NTE * GES$ and $NTE * GFS$ of all proteins are calculated before the optimization, and max computation of this step is $O(K^2)$. K is the total number of proteins. However, because of the property of small-world, the computation of this step is much smaller than $O(K^2)$. Thus the computation of core AMOBH algorithm dominates the computation of IMAMOBH.

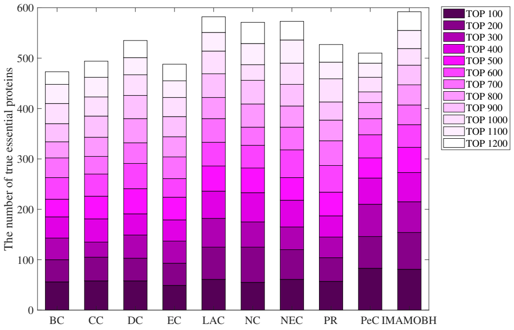


Fig. 2. Comparison of the number of true essential proteins identified from Yeast PPI dataset (different colors means different top ranked proteins intervals).

4. RESULTS AND DISCUSSION

4.1. Validation Metrics

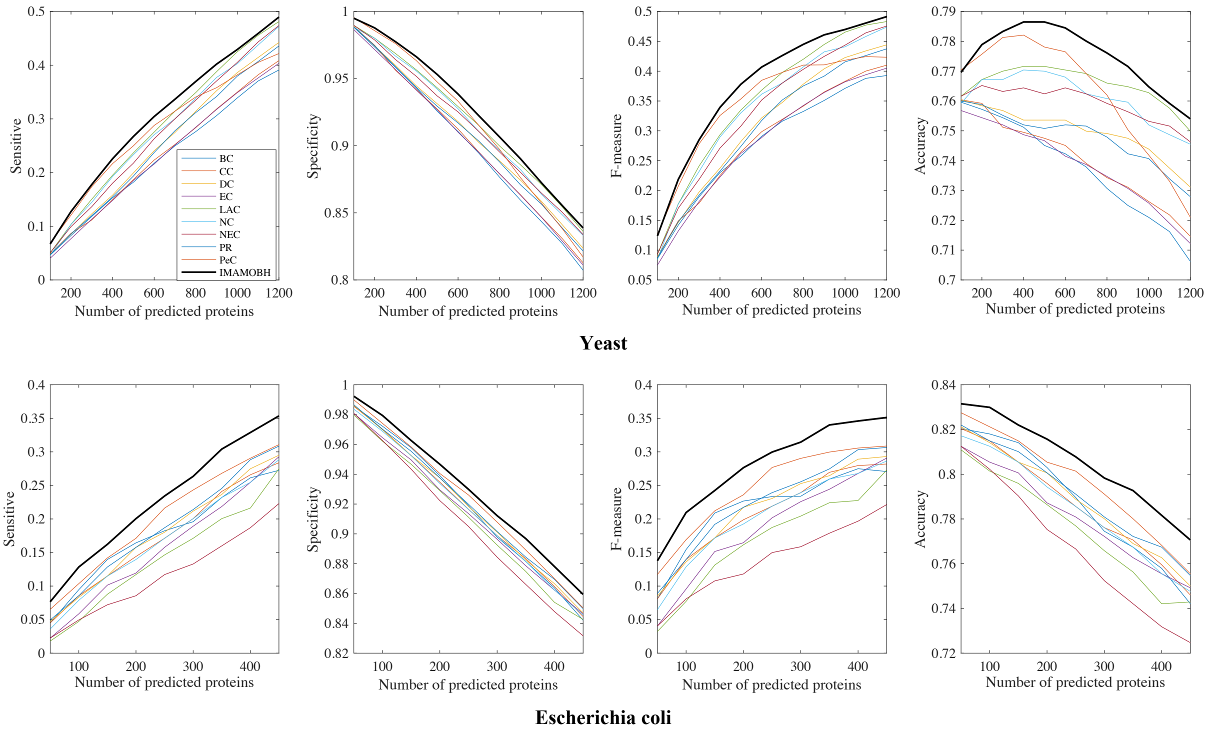
To verify the proposed method, in this paper, we select several most frequently used validation metrics: sensitive, specificity, F-measure, and accuracy [4].

4.2. Performance Analysis

We applied IMAMOBH on the PPI datasets of Yeast and Escherichia coli and compared its performance with several state-of-the-art methods: BC, CC, DC, EC, LAC, NC, NEC, PR (PR is the abbreviation of PageRank), and PeC. All methods adopt the default parameters and all experiments are run on a personal computer with Windows 10 OS, Intel Core i7 2.3GHz CPU, and 8GB memory. As most of validation methods for the identification of essential proteins, we also ranked all proteins by using each essential protein search method and selected a certain number of top ranked proteins as the essential protein candidates. Considering the number of true essential proteins in the PPI data of Yeast and Escherichia coli, we set the range of essential protein candidates of Yeast from top 1% to top 24% and the range of essential protein candidates of Escherichia coli from top 1% to 18%[1].

Fig 2 shows the comparison of the number of true essential proteins identified from Yeast PPI dataset using BC, CC, DC, EC, LAC, NC, NEC, PR, PeC, and IMAMOBH. From Fig 2 we can see that, our method identifies the most number of true essential proteins in almost all essential protein candidates. Fig 3 shows the results of 4 evaluation metrics (sensitive, specificity, F-measure, and accuracy) obtained by

[1] All the work (data and codes of our proposed method) can be downloaded on: https://github.com/ProfHubert/IMAMOBH.



Yeast

Escherichia coli

Fig. 3. Comparison of sensitive, specificity, F-measure, and accuracy obtained by all methods on the Yeast and Escherichia coli PPI dataset. Upper row is the results from Yeast PPI dataset. Bottom row is the results from Escherichia coli PPI dataset.

all identification methods on the PPI network of Yeast. We can see that IMAMOBH outperforms rest methods in terms of all evaluation metrics in all top ranked proteins.

Fig 4 shows the comparison of the number of true essential proteins identified from Escherichia coli PPI dataset using BC, CC, DC, EC, LAC, NC, NEC, PR, PeC, and IMAMOBH. It can be seen clearly that our method identifies more true essential proteins against rest methods in all essential protein candidates. And Fig 3 shows the results of 4 evaluation metrics obtained by all identification methods on the PPI network of Escherichia coli. It can be seen clearly that IMAMOBH achieves the best results of all evaluation metrics in all top ranked proteins. What's more, we can see that some methods like NC, LAC, and NEC achieve good results on Yeast PPI dataset. However, when they are applied on Escherichia coli PPI dataset, their performance is largely degraded. This proves that the generalization ability of our method is better than other state-of-the-art methods used in this paper.

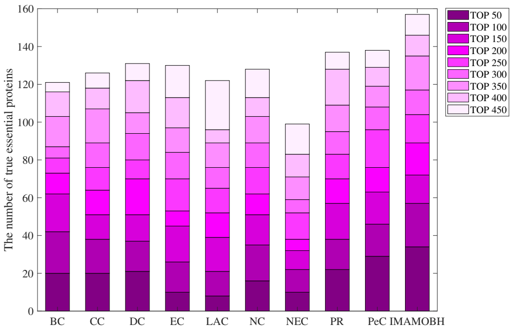


Fig. 4. Comparison of the number of true essential proteins identified from Escherichia coli PPI dataset (different colors means different top ranked proteins intervals).

5. CONCLUSION

In this paper, we consider the identification of essential proteins as a multi-objective optimization problem and use AMOBH algorithm to solve it. We call this identification method IMAMOBH. Our method avoids the subjective selection of weights and achieves a better generalization ability. The validation experiments on the PPI data of Yeast and Escherichia coli show that our method achieves better performance in terms of sensitive, specificity, F-measure, and accuracy compared to some state-of-the-art methods. In future, we will validate our method on more different species' PPI datasets and sought to introduce many-objective optimization methods to the identification of essential proteins.

6. REFERENCES

[1] B. Xu, J. Guan, Y. Wang, and Z. Wang, "Essential protein detection by random walk on weighted protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 2, pp. 377–387, March 2019.

[2] M. L. Acencio and N. Lemke, "Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information," *BMC Bioinformatics*, vol. 10, no. 1, pp. 290, 2009.

[3] Y. Zhu and C. Wu, "Identification of essential proteins using improved node and edge clustering coefficient," in *2018 37th Chinese Control Conference (CCC)*. IEEE, 2018, pp. 3258–3262.

[4] M. Li, P. Ni, X. Chen, J. Wang, F. Wu, and Y. Pan, "Construction of refined protein interaction network for predicting essential proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1386–1397, July 2019.

[5] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks?," *PLoS Genetics*, vol. 2, no. 6, pp. e88, 2006.

[6] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[7] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the Yeast protein interaction network," *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[8] S. Wuchty and P. F. Stadler, "Centers of complex networks," *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 45–53, 2003.

[9] Y. Wang, H. Sun, W. Du, E. Blanzieri, G. Viero, Y. Xu, and Y. Liang, "Identification of essential proteins based on ranking edge-weights in protein-protein interaction networks," *PLoS ONE*, vol. 9, no. 9, pp. e108716, 2014.

[10] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[11] J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1070–1080, 2012.

[12] Y. Q. Jiang, Y. Wang, G. S. Wang, G. Ou, C. Su, and L. Huang, "Essential protein identification by a bootstrap K-nearest neighbor method based on improved edge clustering coefficient," *Computational Intelligence in Industrial Application*, pp. 145–149, 2015.

[13] M. Li, H. Zhang, J. Wang, and Y. Pan, "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data," *BMC Systems Biology*, vol. 6, no. 1, pp. 15, 2012.

[14] C. Wu, T. Wu, K. Fu, Y. Zhu, Y. Li, W. He, and S. Tang, "AMOBH: Adaptive multiobjective black hole algorithm," *Computational Intelligence and Neuroscience*, vol. 2017, pp. Article ID 6153951, 2017.

[15] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, "A local average connectivity-based method for identifying essential proteins from the network level," *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 143–150, 2011.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web.," Tech. Rep., Stanford InfoLab, 1999.

[17] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg, "DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[18] W. H. Chen, P. Minguez, M. J. Lercher, and P. Bork, "OGEE: an online gene essentiality database," *Nucleic Acids Research*, vol. 40, no. D1, pp. D901–D906, 2011.

[19] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[20] Y. Jiang, Y. Wang, W. Pang, L. Chen, H. Sun, Y. Liang, and E. Blanzieri, "Essential protein identification based on essential protein–protein interaction prediction by integrated edge weights," *Methods*, vol. 83, pp. 51–62, 2015.

[21] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 296–304.