A Comparison Study on Protein-protein Interaction Network Models

Mingyu Shao*, Yi Yang*, Jihong Guan† and Shuigeng Zhou*
*School of Computer Science, Fudan University, Shanghai 200433, China.
{shaomy, yyang1, sgzhou}@fudan.edu.cn
†Department of Computer Science & Technology, Tongji University, Shanghai 201804, China
jhguan@tongji.edu.cn

***Abstract*—This paper presents a comprehensive comparison study on the performances of major existing models over two PPI datasets, by comparing the global and local statistical properties of the original PPI networks and the model-reproduced ones. Our experimental results show that the DD model has best fitting ability while iSite model and STICKY model also fit well with the PPI datasets over most statistical properties.**

***Keywords*-PPI network models; Performance comparison; Biological mechanism; Statistical properties.**

I. Introduction

In biological systems, most proteins play biological functions through binding each other together to form protein-protein interactions (PPIs). These PPIs and the corresponding proteins further form PPI networks. Therefore, PPI networks are fundamental to organisms. Modern biotechnology, such as two-hybrid (Y2H) assays [1], tandem-affinity purification and mass spectrometry (TAP/MS) [2, 3] has produced an increasing amount of PPI data, which contain a wealth of biological information, and thus have motivated the analysis and modeling of PPI networks. An appropriate PPI network model can help biologists better understand the underlying mechanisms of PPI network formation and evolution. As more and more PPI data are available, a number of models have been proposed to model PPI networks, such as Duplication-Divergence model [4] and geometric random graphs [5]. However, their fitting capabilities are quite different. Considering this situation, it is necessary to comprehensively and systematically compare different models by empirical study and identify which models can better describe the characteristics of the PPI networks. This will facilitate greatly biologists to choose proper PPI network models for their research work and to establish better models. However, little work has been done on this problem so far in the literature.

In this paper we comprehensively studied and compared the major existing PPI network models over large PPI datasets. We selected ten typical network models, of which eight models were proposed specially for PPI network modeling and the remaining two are more general models for network modeling. As yeast has relatively the most complete PPI data, two yeast PPI datasets were used for this study. The comparison is based on a bunch of network measures to characterize the global and local statistical properties of PPI networks. By comparing the statistical properties between the original PPI networks and the model-reproduced networks, we find that the DD model fits best with the PPI data while iSite model and STICKY model also fit well with the PPI datasets over most statistical properties. By analyzing the embedded mechanisms of these models, we speculate that PPI networks exhibit "degree-weighted" behavior and evolve by gene duplication and divergence.

II. Major PPI Network Models

PPI network modeling has been studied for more than a decade, and a dozen of models have been proposed to model PPI networks of yeast, human and many other species. Here, we selected ten typical models, eight of which are models proposed for PPI networks and the other two (the Erdős–Rényi random graph model and the Barabási–Albert model) are theoretical models used for comparison. Table I summarizes the models used in our study. Readers can refer to papers cited in the table for further details of these models.

III. Network Statistical Properties

Comparative analysis of networks (or graphs) could enable us to find the similarity and difference between networks generated from models and networks inferred from real data. Since direct comparison of networks involves solving the "subgraph isomorphism problem", which has no efficient approach so far, in this paper we compare two networks by comparing much easily computed statistical properties of the networks. The statistical properties of networks used in our paper roughly fall into two categories: global properties and local properties. Widely used global properties include *average degree*, *degree distribution*, *average clustering coefficient*, *average shortest path length* and *diameter*. Local properties include *GDD-agreement* and *RGF-distance*. Detailed description of the local statistical properties used in our paper can be referred to [5]. Definitions of the global properties can be found in [14].

IV. Experimental Evaluation

A. Datasets and Experimental setting

The latest versions or updates of two popular PPI datasets were used in our study: DIP (28.February.2012) [15] and INTACT (7.February.2012) [16]. Their details are given in

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

Table I: Summary on typical PPI network models.

	Basic principles	Biological background[a]	Empirical data[b]	Node-attribute[c]
ER [8]	Two nodes are linked with probability p	–*	–	–
BA [9]	Preferential attachment	–	–	–
DD [4]	Node duplication and link dynamics	Gene duplication and divergence	Heteromerization rate	–
Berg [10]	Node duplication and link dynamics	Gene duplication and divergence	Node duplication rate Edge addition rate	–
iSite [6]	Node duplication and link dynamics	Gene duplication and divergence	Subfunctionalization asymmetry Heteromerization rate	Binding sites
Solé [11]	Node duplication and link dynamics	Gene duplication and divergence	Edge addition and deletion rate	–
Thomas [12]	Linked if two nodes have complementary binding domains	Complementary binding domains	–	Binding domain
GEO [5]	Linked if two nodes within a radius r	–	–	–
STICKY [7]	More likely to link if two nodes have higher "stickiness"	–	Degree list	–
Two-step [13]	(i) Preferential depletion (ii) Similarity	–	α	–

[a] means the information related to biological mechanism or phenomenon.
[b] means the information extracted from empirical data (except the number of nodes and edges).
[c] means individual protein information.
* each "–" in the table indicates a negative answer.

Table II. For each of the datasets, we removed the redundant interactions by keeping just the interactions with the highest scores. Approximately, 0.5% and 24% of the interactions were removed from DIP and INTACT respectively.

We evaluated and compared the 10 models listed in Table I. In order to compare these models, we generate networks based on the datasets. The general rule of network construction is to keep the numbers of nodes and edges match with that of real datasets. Most models use only the numbers of proteins and interactions as input. Additional parameters were taken directly from the original papers in some cases, while some parameters of a few models were fine-tuned to keep the numbers of nodes and edges match with that of real datasets. To guarantee the robustness of the results, we repeated the network generation process 100 times for each model with the same parameter setting. The statistical properties of the networks were calculated using the source code from GraphCrunch [17] with our own modification and efficiency optimization.

B. Results and Analysis

1) Statistical Characteristics of Datasets: Table II compares the statistical properties of the datasets used in this paper. The number of nodes (proteins) approximates the expected size of the yeast genome ($\sim\ 6,000$) whereas the total number of interactions are much larger than the estimated number of yeast proteome ($18,000 \pm 4,500$ [18]), which may be due to the high false positive rate of high throughput screens. From Table II, we can see that clustering coefficients of the two datasets are generally at least one order of magnitude larger than that of random graphs, while the average shortest path lengths and diameters show no significant difference.

The findings shown above and the degree distributions shown in Fig. 1 indicate that both datasets have roughly

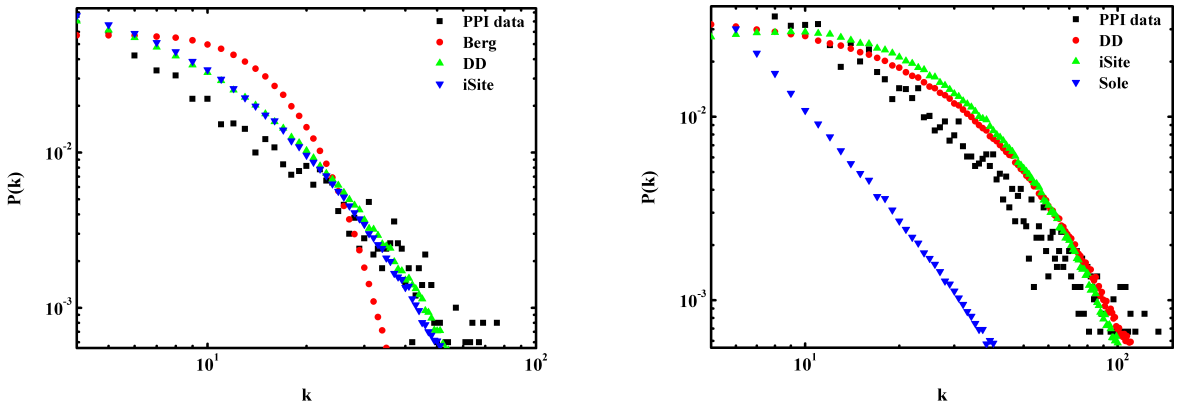
Table II: Datasets characteristics. For comparison, we also give statistical values of random graphs with similar average degrees to that of the PPI datasets. These statistical values are clustering coefficient, average shortest path length and diameter, corresponding to the rows C_{rand}, l_{rand} and D_{rand}, respectively.

	DIP	INTACT
Proteins	5004	5942
Interactions	22325	74097
$\langle k \rangle$	8.92	24.96
C/C_{rand}	0.096/0.002	0.29/0.004
l/l_{rand}	3.97/4.14	2.57/2.98
D/D_{rand}	10/7	7/4

similar network characteristics, and exhibits the properties of small world [19].

2) Comparison among the Models: For each dataset, we first compared the degree distributions of all models with that of the PPI datasets. Fig. 1 shows the degree distributions of two datasets in double logarithmic scale, where each subfigure shows the three models that have the highest correlation coefficients with the PPI dataset. We can see from Fig. 1 that both DD model and iSite model generate similar degree distributions to that of the two datasets, while Berg model and Sole model also give similar degree distributions to DIP and INTACT dataset, respectively.

As can be seen from Table II, there is no significant difference between the average shortest path length and diameter of the PPI datasets and that of the random graphs, which indicate that they may be not good indicators for distinguishing the models. Thus we compared the average clustering coefficient. Table III shows the differences of the average clustering coefficients between the models and the datasets. A better model has clustering coefficient closer to that of the real data. The closest three values of each



(a) DIP dataset (b) INTACT dataset

Figure 1: Degree distributions of PPI datasets and tested models. (a) The degree distributions of the DIP dataset (square) and the top three models: Berg (circle), DD (up triangle) and iSite model (down triangle); (b) The degree distributions of the INTACT dataset (square) and the top three models: DD (circle), iSite (up triangle) and Sole model (down triangle).

Table III: Differences of clustering coefficients between model-reproduced networks and PPI data inferred networks.

	DIP	INTACT
BA	**-0.086**	−0.267
Berg	−0.093	−0.280
DD	**0.082**	**-0.128**
ER	−0.094	−0.281
GEO	0.497	0.313
iSite	0.260	**0.058**
Solé	−0.093	−0.281
STICKY	**-0.072**	**0.134**
Thomas	−0.094	−0.279
Two-step	0.437	0.359

dataset are in boldface. We can see that the BA model, DD model and STICKY model occupy the top three places of the clustering coefficient over DIP dataset, while the DD model, iSite model and STICKY model have clustering coefficients closest to that of the INTACT dataset. The good performance of the BA model indicates that PPI networks have high clustering coefficients comparable to that of scale-free networks.

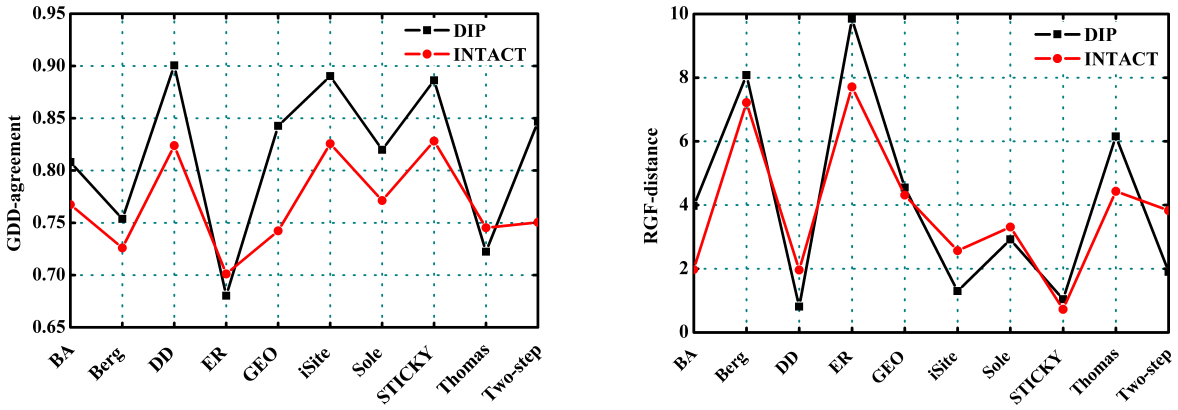
In addition to global statistical properties, local properties, GDD-agreement and RGF-distance, were also compared among the models. If two networks are similar, they will have larger GDD-agrement and smaller RGF-distance. Fig. 2 shows the comparison results of GDD-agreement and RGF-distance. From Fig. 2, we can see that the DD model, iSite model and STICKY model have the largest GDD-agreement and all of the three models have RGF-distances smaller than 3.0 for both datasets, while the BA model also has small RGF-distances for the INTACT dataset.

Considering all properties compared above, DD model performs best among all models. The good fitting ability of DD model indicates that PPI networks possibly evolve by gene duplication and divergence. The iSite model is based on the DD model, and it has good performance on most statistical properties. However, it does poorly in terms of clustering coefficient. It is probably because during the edge deletion step, only interactions from progenitor (progeny) node can be deleted, which increases the probability to keep edges between the neighbors of the progeny (progenitor) node compared to DD model. The good fitting ability of STICKY model may be attributed to two reasons: the large amount of information it uses as input (i.e., degree list) and the edge addition principle it adopts. Given the degree lists of networks generated by other kinds of models, the STICKY model could only fit well with networks that are generated involving the degrees of its nodes (data is not shown due to space limit). This suggests that the edge addition principle may agree with the behavior of PPI networks. We call this "degree-weighted" behavior. Thus, it seems that our PPI data has "degree-weighted" behavior.

V. Discussion and Conclusion

In this paper, we compared ten different network models by using two different yeast PPI datasets. The empirical results show that the DD model, followed by iSite model and STICKY model fits best with the two PPI datasets. By analysis of the three models, we speculate that yeast PPI networks show the "degree-weighted" behavior and evolve possibly by gene duplication and divergence.

It is also meaningful to have a deep look at the other models evaluated in this paper. The Berg model and Solé model use the same biological mechanism (i.e. gene duplication and divergence) to reproduce PPI networks as DD model does, yet they perform differently. The difference between the two models and DD model lies in that the former also introduce *de novo* interactions, whose ubiquity in protein



(a) GDD-agreement

(b) RGF-distance

Figure 2: Comparison of local statistical properties.

interaction network evolution was questioned [20]. Another possible reason of poor performance of these two models is that the parameters are not appropriate. The results might be better if parameters are properly estimated. The GEO model shows very interesting properties in the comparison. It has similar degree distribution to ER model, very high clustering coefficient, and large average shortest path length. It seems that the GEO model neither belongs to random graphs nor has "small world" properties. Thus, we guess that the GEO model has a more regular structure than the PPI networks. The poor performance of two-step model and Thomas model may suggest that the mechanisms underlying these two models do not dominate the evolution of PPI networks.

Table I also gives us some insight about good models. The information of respective proteins does not seem to lead the formation of PPI networks. A good model benefits from reasonable biological mechanism and appropriate estimation of the empirical data. Currently, the yeast PPI network is far from complete, hence it is possible that there also exist other mechanisms shaping the PPI network. With the rapid development of biotechnology, more accurate and complete data will be available. We can expect more appropriate models for PPI networks to appear in the near future.

ACKNOWLEDGMENTS

This study was supported by China 863 Program (grant No. 2012AA020403) and 973 program (grant No. 2010CB126604). Jihong Guan was also supported by NSFC (grant No. 61173118).

REFERENCES

[1] Peter Uetz, Loic Giot, Gerard Cagney, *et al.*, *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*, Nature, 403(6770): 623-627, 2000.

[2] Anne-Claude Gavin, Markus Bösche, Roland Krause, *et al.*, *Functional organization of the yeast proteome by systematic analysis of protein complexes*, Nature, 415(6868): 141-147, 2002.

[3] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, *et al.*, *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*, Nature, 415(6868): 180-183, 2002.

[4] Alexei Vázquez, Alessandro Flammini, Amos Maritan, *et al.*, *Modeling of protein interaction networks*, Complex Syst., 1(cond-mat/0108043): 38-44, 2001.

[5] Nataša Pržulj, Derek G. Corneil, Igor Jurisica, *Modeling interactome: scale-free or geometric?*, Bioinformatics, 20(18): 3508-3515, 2004.

[6] Todd A. Gibson, Debra S. Goldberg, *Improving evolutionary models of protein interaction networks*, Bioinformatics, 27(3): 376-382, 2011.

[7] Nataša Pržulj, D.J. Higham, *Modelling protein-protein interaction networks via a stickiness index*, J R Soc Interface, 3(10): 711-716, 2006.

[8] P. Erdős, A. Rényi, *On random graphs, I*, Publicationes Mathematicae (Debrecen), 6: 290-297, 1959.

[9] Albert-László Barabási, Réka Albert, *Emergence of scaling in random networks*, Science, 286(5439): 509-512, 1999.

[10] Johannes Berg, Michael Lässig, Andreas Wagner, *Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications*, BMC Evolutionary Biology, 4(1): 51, 2004.

[11] Ricard V. Solé, Romualdo Pastor-Satorras, Eric D. Smith, *et al.*, *A model of large-scale proteome evolution*, Advances in Complex Systems, 5(1): 43-54, 2002.

[12] Alun Thomas, Rob Cannings, Nicholas A. M. Monk, *et al.*, *On the structure of protein-protein interaction networks*, Biochem Soc Trans., 31(6): 1491-1496, 2003.

[13] Christian M. Schneider, Lucilla de Arcangelis, Hans J. Herrmann, *Modeling the topology of protein interaction networks*, Phys Rev E Stat Nonlin Soft Matter Phys, 84(1 Pt 2): 016112, 2011.

[14] M. E. J. Newman, *The structure and function of complex networks*, SIAM REVIEW, 45(2):167-256, 2003.

[15] Lukasz Salwínski, Christopher S. Miller, Adam J. Smith, *et al.*, *The Database of Interacting Proteins: 2004 update*, Nucleic Acids Research, 32(Database issue): 449-451, 2004.

[16] Samuel Kerrien, Bruno Aranda, Lionel Breuza, *et al.*, *The IntAct molecular interaction database in 2012*, Nucleic Acids Research, 40(Database issue): 841-846, 2012.

[17] Oleksii Kuchaiev, Aleksandar Stevanović, Wayne Hayes, *et al.*, *GraphCrunch 2: Software tool for network modeling, alignment and clustering*, BMC bioinformatics, 12(1): 24, 2011.

[18] Haiyuan Yu, Pascal Braun, Muhammed A. Yıldırım, *et al.*, *High-quality binary protein interaction map of the yeast interactome network*, Science, 322(5898): 104-110, 2008.

[19] Duncan J. Watts, Steven H. Strogatz, *Collective dynamics of 'small-world' networks*, 393(6684): 440-442, 1998.

[20] Todd A. Gibson, Debra S. Goldberg, *Questioning the ubiquity of neofunctionalization*, PLoS computational biology, 5(1): e1000252, 2009.