Communications in Physics, Vol. 22, No. 1 (2012), pp. 7-14

EVOLUTION OF PROTEIN-PROTEIN INTERACTION NETWORKS IN DUPLICATION-DIVERGENCE MODEL

BUI PHUONG THUY
Nam Dinh University of Technology Education
TRINH XUAN HOANG
Institute of Physics, VAST

Abstract. *Protein interacts with one another resulting in complex functions in living organisms. Like many other real-world networks, the networks of protein-protein interactions possess a certain degree of ordering, such as the scale-free property. The latter means that the probability P to find a protein that interacts with k other proteins follows a power law, $P(k) \sim k^{-\gamma}$. Protein interaction networks (PINs) have been studied by using a stochastic model, the duplication-divergence model, which is based on mechanisms of gene duplication and divergence during evolution. In this work, we show that this model can be used to fit experimental data on the PIN of yeast Saccharomyces cerevisae at two different time instances simultaneously. Our study shows that the evolution of PIN given by model is consistent with growing experimental data over time, and that the scale-free property of protein interaction network is robust against random deletion of interactions.*

I. INTRODUCTION

Proteins are molecules that play crucial roles in almost every biological processes [1]. Their functions include catalysis, transport of ions, cell signaling, and activity in the immune system. Many diseases are found because of a disorder in processing of proteins or a lack of activity by a certain protein. For example, type I diabetes is known to be related to the inability of the pancreas to produce enough insulin to properly control blood sugar levels, whereas type II diabetes is due to a lack of proper function of insulins – or insulin resistance. The functionality of a protein pertains to its three-dimensional structure. Protein structures can be measured experimentally by NMR or X-ray crystallography, and to some extents, at least, can be partially predicted by computer models [2]. The prediction of protein function is more difficult. Until now, only the functionalities of very few proteins are known.

It has been demonstrated in experiments that proteins may bind together when functioning [3]. The knowledge of interactions between proteins is needed to understand protein activities in complex processes such as signal processing, DNA duplication, protein synthesis and cell division. Thanks to advanced experimental techniques, especially the new two-hybrid screening method [4], protein-protein interactions are now routinely determined in laboratories. A map of protein-protein interactions is called a protein interaction network (PIN) [5]. The PINs are large-scale and complex networks. For example, the PIN of a simple organism, the single-celled yeast Saccharomyces cerevisae, has nearly 6000 proteins and thousands of interactions [5–7].

Protein interaction networks are generally expressed in form of a graph in which each node denotes a protein and a connection between two nodes is present if there is a physical interaction between the two proteins. The number of connections, k, that a node possesses is called the degree of the node. The distribution of k is call the degree distribution. Like many other real-world networks [8], the degree distributions of PINs are found to follow a power law, $P(k) \sim k^{-\gamma}$. The exponent γ is found to be between 2 and 3 for various types of networks. Networks of this property are called scale-free [9]. Such a degree distribution is very different from that of random graphs, which have a Poisson distribution [10, 11]. The discoveries in studies of complex networks have led to the developments of graph generating models for the purpose of having an universal theory of scale-free networks (for a review see [8]).

The scale-free property of protein interaction networks has been successfully reproduced by the duplication-divergence model [12]. This model is based on two key mechanisms of biological evolution: 1) duplication of genes during cell division and 2) random genetic mutation between the gene duplicates. In this model, a network starts from small size having a few nodes and a few links, and grows in size as new elements are added. The copying of nodes mimics the gene duplication and the random assignments and removals of links mimic the random mutation. A similar model has been introduced independently by Pastor-Satorras *et al.* [13]. These models have demonstrated that evolution can produce scale-free networks.

In this study, we focus on the duplication-divergence model in light of the newest experimental data on the S. cerevisae protein interaction network. These data have been growing in time as new interactions were discovered. We ask the question whether the networks produced by the duplication-divergence model are consistent with experimental data at two different time instances: the PIN data in 2000 and those in 2008. By assuming that the PIN in 2000 is a sub-net of the PIN in 2008, the former can be obtained from the latter by deletion of a certain number of links. Our study shows that by using a single set of parameters in the model, one can fit very well the degree distribution obtained by the simulations with the 2008 experimental data, and then by random deletion of links, one can recover the distribution of the 2000 experimental data.

II. THE PROTEIN INTERACTION NETWORK OF YEAST SACCHAROMYCES CEREVISAE

Saccharomyces cerevisae is one of the most simple single-celled organisms. It was the first eukaryote to have its entire genome sequenced. Since then, it remains in the forefront of genetic research. The genome of S. cerevisae has over 12 million base pairs and around 6000 genes. The protein interaction network of S. cerevisae, however, is very complex. So far, the data on protein-protein interactions of S. cerevisae has not been fully complete and is being supplemented each year by laboratories worldwide.

The experimental methods for determining specific protein-protein interaction are sophisticated, often have to be realized outside living organisms (in vitro). Most of the binary protein interaction data currently available was generated by large-scale yeast two-hybrid screening (Y2H) method. In pioneering experiments in 2000, Uetz *et al.* [5] have mapped out 957 pair-wise interactions between 1004 proteins in S. cerevisae. In 2008, Yu

et al. [7] have combined data from different sources and built up a network that contains 2930 binary interactions among 2018 proteins. We will call these data sets as 2000 and 2008 experimental data sets. It was estimated that these interactions represent only about 20% of the whole yeast binary interactome [7].

The experimental data in both years 2000 and 2008 show that the PIN of S. cerevisae is scale-free with an exponential cut-off (Fig. 1). The degree distribution can be fitted by the formula:

$$P(k) \sim e^{-k/k_c}(k + k_0)^{-\gamma} , \tag{1}$$

where k_c and k_0 are parameters for the exponential cutoff and degree shift, respectively. For large k one recovers the simple power law dependence $P(k) \sim k^{-\gamma}$. The exponent γ is found to be about 2.4 for the data in both years.

III. DUPLICATION-DIVERGENCE MODEL

Biological basis of the model

The discovery of DNA structure in 1953 had a major impact on the rapid development of molecular biology as a biological science and has opened an entirely new era of genetic research. Since then, the molecular mechanisms for biological heredity and evolution have been largely understood. Two key mechanisms for evolution are gene duplication and DNA base pair mutation. During the process of DNA replication, some gene can be accidentally duplicated creating an two identical genes in the genome. Sometimes, the duplication or replication can be erroneous and the new gene can be different from the original one. These random mutations in the genome sequence set out for evolution to take place in concert with natural selection. When the duplicated gene is much different from the original one or when mutations affect some key sites, a new protein is created whose function can be similar or different from the original one.

The duplication-divergence model defines an evolving network with a dynamics akin to the above mechanisms. The network starts from small size with a few nodes and a few links. Each node in the network represents a protein that is expressed by a gene, and each link represents an interaction between two proteins. The network grows in two repeated steps: duplication and divergence. In duplication step, a node is duplicated characterizing a new protein to be born. The duplicated protein is identical to the original one, thus it will interact with all proteins that interact with the parent protein. In the divergence step, the characteristics of the duplicated node and the parent node are changed with some probability. These can be modeled by random addition or removal of links on the network. The divergence step corresponds to base pair mutations that lead to a change in activity or function of a certain protein during the gene duplication process.

Description of the model

The duplication-divergence model was introduced by Vazquez el al. [12] in 2003. They formalized the evolving network according to the following rules:

(a) Duplication step: A node i is selected at random. A new node i', being a copy of i, is created. i' is linked to all the nodes that i is neighbored to. A link between i and i' is established with probability p.

(b) Divergence step: For each of the nodes j linked to i and i' one chooses randomly one of the two links, (i,j) and (i',j), and remove it with probability q.

In the above rules, p is a parameter that models the creation of an interaction between the duplicates of a self-interacting protein and its possible loss due to the divergence of the duplicates. The other parameter q represents the loss of interactions between the duplicates and their neighbors due to the divergence of the duplicates. Available experimental evidences indicate that q is very large and p is much smaller than q. The network growth can be simulated on computer. Vazquez *et al.* have found that $p = 0.1$ and $q = 0.7$ give the best fit to the PIN of S. cerevisae. In the simulations, the growth starts from an initial network of two nodes linked to each other.

In an independent work, Pastor-Satorras *et al.* [13, 14] introduced a growth model similar to the model of Vazquez *et al.* and based on the same biological basis. Pastor-Satorras *et al.* considered several scenarios for the rules of network growth. In addition, they have provided a mean-field treatment of the model that leads to an asymptotic behavior of the degree distribution as given by Eq.(1). This formula fits well experimental data on the PIN of S. cerevisae.

In this study, we use a variant of the model given by Pastor-Satorras *et al.* which provides the best fit to experimental data. This model is described in details as follows:

(a) The network growth starts from a set of n nodes with connectivities forming a close ring.
(b) Duplication step: a node i is selected at random and duplicated. The duplicated node i' is linked to all the nodes that i has interaction.
(c) Divergence step: The links from the duplicated node i' are removed with probability q. New links (not previously present after duplication step) are created between i' and all the rest of the nodes with probability p.
(d) The duplication and divergence steps are repeated until a desired number of nodes N is achieved.

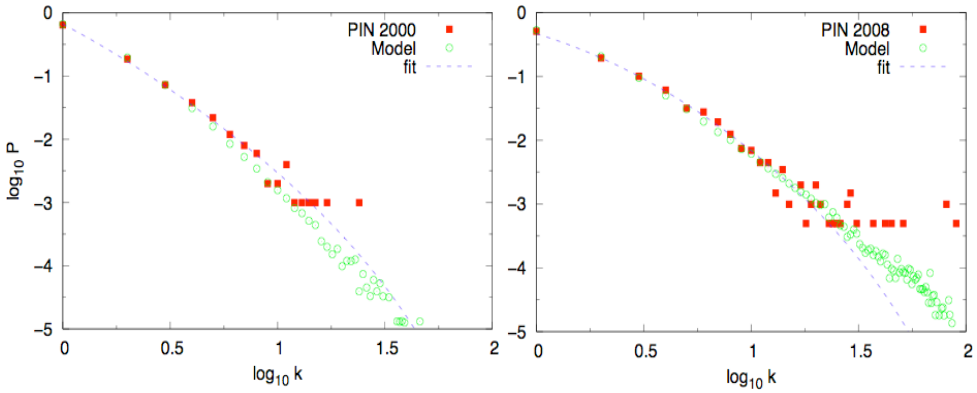
The model of Pastor-Satorras *et al.* is more general than the model of Vazquez *et al.* First, in latter the growth starts from a network of two nodes while in the former the number of initial nodes is a variable n. Second, Vazquez et al. allow only one new link to be established between i' and i (with probability p) while in the model of Pastor-Satorras the new node i' can have link to any other node.

IV. RESULTS AND DISCUSSION

Comparison of the model with experiments

Previous studies of Vazquez *et al.* and Pastor-Satorras *et al.* have considered the network size N to be ≈ 2000, equal to the size of the PIN revealed by experiments for S. cerevisae at the time of their publications. Thus, they have not grown the network to the size of the entire genome. The total number of active proteins in S. cerevisae has been estimated to be about 5800.

In this paper, we fix the network size to be $N = 5800$ like in the full network of S. cerevisae. The purpose is to check how other parameters like p and q change with



(A) (B)

Fig. 1. Log-log plots of the degree distribution $P(k)$ obtained by the duplication-divergence model and from experimental data for the protein interaction network of S. cerevisae. (A) Comparison between the model and experimental data collected by Uetz *et al.* [5] in year 2000. (B) Comparison between the model and experimental data collected by Yu *et al.* [7] in year 2008. The data of the model and experiments are shown in open circles and filled squares, respectively, as indicated. Experimental data are fitted by the formula given by Eq. (1) (dashed line) with $\gamma = 2.4$, $k_c = 15$, $k_0 = 0.4$ for the data in 2000 and $\gamma = 2.4$, $k_c = 16$, $k_0 = 1.5$ for the data in 2008.

different experimental data sets. We expect that the number of undiscovered protein-protein interaction is reflected in parameter q, the probability to remove links to the new duplicated node. We found that the 2000 experimental data can be best fitted with the model by using $n = 3$, $p = 0.1$ and $q = 0.75$, and the 2008 experimental data can be fitted with $n = 7$, $p = 0.12$ and $q = 0.7$. Note that because there are more interactions in the 2008 experimental data we got a slight increase in p and a slight decrease in q compared to those of the 2000 experimental data.

In order to find the best values of parameters, we scanned the p and q parameter space in 0.01 increments, and tried several small values of n. For each set of parameters, we run the model 100 times by computer simulations to obtain 100 final independent networks. The degree distribution is made averaged over the 100 networks and compared to the distribution from experimental data. Fig. 1 compares the degree distributions $P(k)$ between the model and experiments for two data sets in 2000 and 2008. Note that the agreements are very good, especially at low k. For high k, the experimental data are dispersed because of low statistics.

Following Vazquez *et al.* [12], we rank order the nodes in the network according to their degrees. The most highly connected node has rank $r = 1$, the second highest connected node has rank $r = 2$ and so on. The plot of k versus r gives more information on the high-k regime than the $P(k)$ plot. Fig. 2 compares the $\log(k)$ versus $\log(r)$ plots between the model and experiments. It is shown that the agreement between the model and experiments are remarkably good. Note that the agreement shown here is much

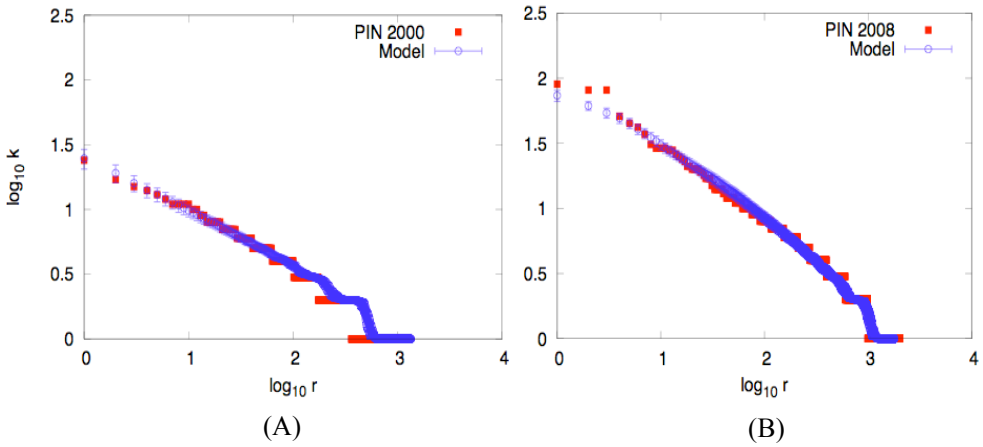


Fig. 2. Log-log plots of the dependence of degree k of the nodes on their rank r. The figure compares the results from the model and the 2000 experimental data (A) and 2008 experimental data (B).

better than the one shown in the paper of Vazquez *et al.* We found that the value of n, the number of initial nodes, affects a lot the high-k part of the plot. It is crucial that the initial network has more than two nodes before the evolution starts.

Comparison with experiments after random deletion of interactions

So far, the experimental results of protein interaction network of S. cerevisae have not been completed. It means there are still some interactions between proteins that have not been discovered yet. The networks revealed by experimental data in 2000 and 2008 are sub-nets of the full network of S. cerevisae. Similarly, the network formed by the data in 2000 is a sub-net of the network formed by the data in 2008, neglecting possible experimental inconsistency.

In this study, we try to check by using the duplication-divergence model whether the network in 2000 can be recovered from the network in 2008 by random deletion of links. With $n = 7$, $p = 0.12$ and $q = 0.7$, the model gives an average of about 3000 links of the resulting networks from 100 runs. For each run, after the network has been fully grown, we try to remove the links randomly so that only 957 links remains (the same number of interactions is found in 2000 experimental data). Fig. 3 shows the comparison between the modeled networks after link removal and the real network from the 2000 experimental data. The figure shows that the agreement is very good, at the same quality as the agreement shown on Figs. 1A and 2A. This is a very interesting result, which shows that the model is consistent with experimental data at two different times, using the same set of parameters.

Next, we want to see how the scale-free property changes by random removal of interactions. Fig. 4 shows the $P(k)$ plots for the networks after links' removal. The number of remained interactions (or links) are 2500, 1500, 1000, and 500 as indicated. The figure shows that the scale-free property persists even when a large portion of the links were removed. The fits shown in the figure are all with $\gamma = 2.4$.

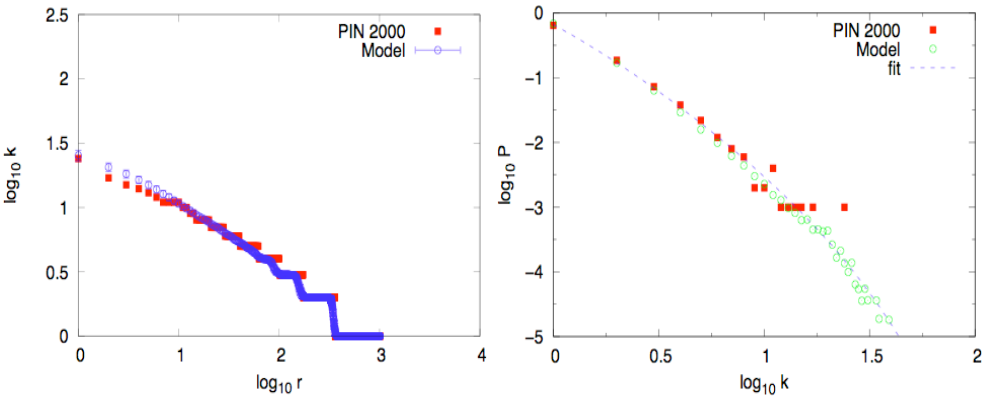


Fig. 3. Log-log plots of order-rank dependence (A) and degree distribution (B) for the networks obtained by random removal of links from the simulation data for year 2008. The model is compared to experimental data (filled squares) for year 2000 as indicated.

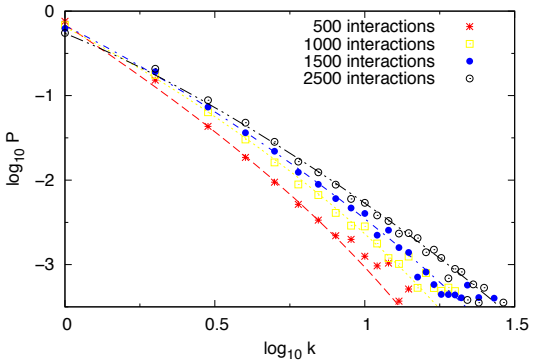


Fig. 4. Log-log plots of the degree distributions of the networks obtained by random removal of links from 2008 simulation data. The number of interactions left on the networks are indicated. The data are fitted by formula given by Eq. (1) using $\gamma = 2.4$ and letting k_0 and k_c varied.

V. CONCLUSIONS

We have shown that the evolution of protein interaction networks can be described quite accurately by the duplication-divergence model. By using an optimized set of parameters, $n = 7$, $p = 0.12$, $q = 0.7$ and $N = 5800$, we show that the modeled networks are scale-free and in very good agreement to the 2008 experimental data for yeast S. cerevisae. On the other hand, by random removal of interactions, from the simulation data for 2008 we obtained results that are in excellent agreement with the 2000 experimental data. This result suggests that the full interaction network of S. cerevisae can be described by the duplication-divergence model, and as the number of interactions increases one should use

a larger p and smaller q in the model to get agreement with experimental data. Our study also shows that the scale-free property of protein interaction networks is robust against random removal of interactions. This robustness may be crucial for the survival of species during evolution.

This work was supported by National Foundation for Science and Technology Development(NAFOSTED).

REFERENCES

[1] T. E. Creighton, *Proteins: Structures and Molecular Properties* (W. H. Freeman and Company, New York, 1993).
[2] P. Bradley, K. M. S. Misura, D. Baker, *Science* **309**, 1868-1871 (2005).
[3] E. Harlow, P. Whyte, S. J. Elledge, *Cell* **75**, 805-816 (1993).
[4] S. Fields, and O. Son, *Nature* **340**, 245-246 (1989).
[5] P. Uetz *et al.*, *Nature*, **403** (2000) 623-627.
[6] H. Jeong, S. Mason, A. L. Barabasi, Z. N. Oltvai, *Nature* **411** , 41 (2001)
[7] H. Yu *et al.*, *Nature*, **322** (2008) 104-110.
[8] R. Albert, A.-L. Barabasi, *Rev. Mod. Phys.* **74**, 47 (2002).
[9] A.-L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
[10] P. Erdos, and A. Rényi, *Publ. Math. (Debrecen)* **6**, 290 (1959).
[11] B. Bollobás, *Discrete Math.* **33**, 1 (1981).
[12] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, *ComPlexUs*, **1** (2003) 38-44.
[13] R. Pastor-Satorras, E. Smith, R.V. Sole, *Journal of Theoretical Biology*, **222** (2003) 199-210.
[14] R. V. Solé, R. Pastor-Satorras, E. D. Smith, T. Kepler, *Adv. Comp. Syst.* **5**, 43 (2002).

Received 30 September 2010.