

Improved Duplication Models for Proteome Network Evolution

Gürkan Bebek¹, Petra Berenbrink², Colin Cooper³,
Tom Friedetzky⁴, Joseph H. Nadeau⁵, and S. Cenk Sahinalp^{2,*}

¹ Department of EECS, Case Western Reserve University, Cleveland, OH 44106-7071 USA

² School of Computing Science, Simon Fraser University, Burnaby BC, V5A 1S6 Canada

³ Department of Computer Science, King's College, London WC2R 2LS, UK

⁴ Department of Computer Science, Durham University, Durham, DH1 3LE, UK

⁵ Genetics Department, Case Western Reserve University, Cleveland, OH 44106-4955 USA

Abstract. Protein-protein interaction networks, particularly that of the yeast *S. Cerevisiae*, have recently been studied extensively. These networks seem to satisfy the small world property and their (1-hop) degree distribution seems to form a power law. More recently, a number of duplication based random graph models have been proposed with the aim of emulating the evolution of protein-protein interaction networks and satisfying these two graph theoretical properties. In this paper, we show that the proposed model of Pastor-Satorras et al. does not generate the power law degree distribution with exponential cutoff as claimed and the more restrictive model by Chung et al. cannot be interpreted unconditionally. It is possible to slightly modify these models to ensure that they generate a power law degree distribution. However, even after this modification, the more general k -hop degree distribution achieved by these models, for $k > 1$, are very different from that of the yeast proteome network. We address this problem by introducing a new network growth model that takes into account the sequence similarity between pairs of proteins (as a binary relationship) as well as their interactions. The new model captures not only the k -hop degree distribution of the yeast protein interaction network for all $k > 0$, but it also captures the 1-hop degree distribution of the sequence similarity network, which again seems to form a power law.

1 Introduction

Protein-protein interactions play a central role in the execution of key biological functions of a cell. Such a relationship can be summarized in a *graph* (network) in which each *node* represents a protein and each (undirected) *edge* represents an interaction. A graph including *all* proteins in an organism and all possible interactions between these proteins can be called the *proteome network* of that organism.

The structure of the yeast proteome network seems to reveal two interesting graph theoretic properties [20, 35]: (i) The degree distribution of nodes (i.e. the

* Corresponding Author; cenk@cs.sfu.ca

proportion of nodes with degree k as a function of degree) approximates a *power-law* (i.e. is approximately ck^{-b} for some constants c, b). (ii) The graph exhibits the *small world effect*.

Small world phenomena and the power-law degree distributions have previously been observed in a number of naturally occurring graphs such as communication networks [14], web graphs [1, 4, 9, 11, 21, 22], research citation networks [29], human language graphs [15], neural nets [36] etc. These two properties can not be observed in the classical random graph models studied by Erdős and Rényi [13] in which edges between pairs of nodes are determined independently. However, it is possible to generate graphs that satisfy these properties by an iterative process that adds one new node to the graph at each step [1, 2, 6, 8, 9, 11, 21]. The new node is then connected to some b (b can be a constant or an independent random variable) of the existing nodes, each of which is chosen with probability proportional to its degree. Unfortunately such a *preferential attachment* model does not capture the essence of the genome evolution and hence can not be used to model proteome networks. According to Ohno’s model [25], the two underlying mechanisms for genome evolution is gene duplication and point mutations.⁶ Recent work, thus, has focused on random graph models that grow via node duplications and get modified by mechanisms that emulate point mutations.

Among these studies, the most promising one, which we call the *general duplication model*, was described independently in [26, 34, 7]. The general duplication model works in iterations; in each iteration t , one existing node (representing a gene or an associated protein) is chosen uniformly at random and is “duplicated” with all its edges. After the *duplication* step, to emulate mutations, also named as the *divergence* step, each edge of the new node is deleted with probability q . This is followed by inserting edges between the new node and every other node with probability r/t where t is the total number of nodes and r is a constant. With the right selection of parameters q and r , the general duplication model well approximates the degree distribution of the yeast proteome network.

The first serious study to formally analyze the degree distribution of the general duplication model was by Pastor-Satorras et al. who, in [26], claim that the distribution of both the general yeast proteome network and the duplication model is a “power law with exponential cut-off”. This means that the fraction of nodes with degree k among all nodes is independent of time and is approximated by $f_k = ck^{-b} \cdot a^{-k}$; here a, b, c are constants. However, they make a number of simplifying assumptions in their analysis to get this result. For instance, they approximate the probability for generating a node with degree k by the proba-

⁶ After a gene duplication event, one of the genes may accumulate deleterious mutations and be lost, or both copies of the gene may be retained. Two possible evolutionary reasons for keeping both copies can be (i) selection for increased levels of expression, or (ii) divergence of gene function[23, 30]. Functional divergence can be produced through complementary degeneration [16]. Although the duplicated regions of the genomes have been described and listed before (for instance *S. Cerevisiae*[31, 37]), there is no certain schema of how duplications formed the current shape of the genomes.

bility of duplicating a node with degree $k + 1$ only and subsequently deleting a single edge. This assumption also reduces the number of *singletons*. They further approximate this probability with a function linear in k .

A more recent analysis of the degree distribution of the general duplication model, for the special case that $r = 0$ is given by Chung et al. [10]. As per [10], we will refer to this special case as the *pure duplication model*. In contrast to [26], Chung et al. claim that the fraction of nodes with degree k is independent of time and is of the form $f_k = ck^{-b}$; here b is a function of $p = 1 - q$ and values of $b \leq 2$ are possible for some p . The pure duplication model creates *singleton* nodes, i. e. nodes that are not connected to any other node of the graph. Since, a node can only get a new edge if one of its neighbors is copied, a singleton will remain singleton during the whole graph generation process. Note that in this model all non-singleton nodes form one connected component.

In a separate work, van Noort et al. [24] show that the gene coexpression network in *S. Cerevisiae* have scale-free and small-world network properties. By using the homology relations between the genes in coexpression network, they present a model which can generate networks with similar scale-free and small-world properties. The model starts with a number of genes which have a number of transcription factor binding sites (TFBSs) and genes sharing a minimum number of TFBSs considered coexpressed. At every time step each gene can be duplicated or deleted with certain probabilities. Also, at every time step a TFBS of a gene can be deleted or a new TFBS from another gene can be acquired by a gene with certain corresponding probabilities. Different from many, van Noort et al. [24] consider deleting or inserting a TFBS of the gene which deletes a set of connections, or adds a set of links to the gene. Hence, in their approach the connections of genes were considered with groups. Van Noort et al. [24] claim that the model generates a degree distribution with a slope similar to the coexpression network of *S. Cerevisiae*⁷. Additionally, average clustering coefficient⁸, and shortest path length of the networks were compared. Although these are measures to understand the topology of a network, they are not sufficient to claim that two networks are similar at all.

There is also another study presented by Przulj et al. [28], in which a different approach to model these networks has been studied. Przulj et al. [28] claim that a random geometric model better captures the currently accepted protein-protein interaction networks. A geometric disc graph is formed by connecting two nodes of the graph with an edge, if their distance in the metric space is smaller than a certain threshold. Przulj et al. [28] argue that the scale-free property of the proteomes is a result of the noise in the available data at the moment and

⁷ Numerical results were not presented in [24]. Hence, the simulation results given draws certain amount of question about how close the degree distribution, i.e. the power-law exponent, was.

⁸ The clustering coefficient of a node is the ratio between the actual number of edges between neighbors of a node and the maximum possible number of edges between these neighbors. Average clustering coefficient of a network is the average of clustering coefficients over all units in the system. [36]

the degree distribution of such networks should follow the Poisson distribution. By counting the number of different motifs in the networks, they form a measure of local network structure and using this they compare different models with the available proteomes. According to the experiments they carry out, a three dimensional geometric disc graph with same number of nodes but six times larger edge count has similar number of motifs as the proteomes they worked on. Although, the network motifs considered capture local properties of the networks, in their work, Przulj et al. [28] (i) do not take into account Ohno's Theory [25] which states that, the proteome network should be generated through a process, which attributes the genome sequence growth and evolution to subsequent gene duplications followed by mutations on the gene sequences, (ii) do not consider global properties of the networks before drawing conclusions, such as the average degree or the degree distribution. Moreover, the work presented has vague descriptions on how scale free networks are formed. For instance, there are many models available that can generate scale free networks, but not every scale free network necessarily is generated by emulating proteome network growth i.e. duplication and divergence.

The most recent study that was presented by Ispolatov et al. [18] focuses on duplication-divergence models with completely asymmetric divergence. In a completely asymmetric divergence process, links are removed from the duplicated node only. In their study, Ispolatov et al. examines this model where the evolution is characterized by a single parameter, the link retention probability. They claim that, this single-parameter duplication-divergence network growth model can approximate the degree distribution of real protein-protein interaction networks. Although their model generates similar degree distributions, in reality the network lacks the local structure similarity. For instance, this model would not generate any triangular subgraphs (a clique of three in the network) since the duplication would generate cycles of even length or degree one nodes. However, cycles of any size exists in vast numbers in the real proteome network.

In most of these studies the protein-protein interactions identified by high-throughput yeast two-hybrid screens or inferred from mass spectrometry of coimmunoprecipitated protein complexes were considered. However, analysis based on the agreement of the interaction and expression data show that almost less than half of these interactions are biologically relevant [12]. In a recent study, Han et al. [17] showed that low coverage makes determination of the true topology of the network difficult. Han et al. also showed that sampling the real network through these experiments (since the experiments only reveal partial networks), regardless of the topology of the network that we are looking for, the topology of the sub network that is sampled would have a degree distribution similar to a power law. In other words, according to these experiments, it is not clear whether the proteome network has a power law degree distribution or not. However, in this paper, we assume that the proteome network should be generated through a process, which attributes the genome sequence growth and evolution to subsequent gene duplications followed by mutations on the gene sequences. Previously, it has been shown that this process would generate a network with power-law

degree distribution [26, 34, 7]. Moreover, we show that the degree distribution of the general duplication model is a power law.

We can summarize our contributions as follows. (i) We show that the degree distribution of the pure duplication model ($r = 0$) cannot be a power law as stated in [10]. (ii) We show that the degree distribution of the general duplication model can not be a power law with exponential cut-off as stated in [26]. In fact, for $r > 0$, it is simply a power law. It is also possible to slightly modify the pure duplication model so that it achieves a power law degree distribution but these details are left to a more complete version of the paper [5] due to space limitations. (iii) The (1-hop) degree distribution of a graph is the distribution of nodes with degree k as a function of k . A more general notion is the ℓ -hop degree distribution which is defined to be the distribution of nodes that can reach k nodes in at most ℓ -hops as a function of k . We observe in this paper that the general duplication model does not capture the ℓ -hop degree distribution of the yeast proteome network for $\ell > 1$. (iv) We describe a new model that takes into account the sequence similarity between protein pairs as a binary relationship in addition to their interactions. Our model accurately captures the ℓ -hop degree distribution of the yeast proteome network for all $\ell > 0$ and yields a good approximation to the degree distribution of the sequence similarity network.

Our specific contributions are as follows. We first show in Section 2 that the (expected) proportion of singletons generated by the pure duplication model ($r = 0$) grows in time. In fact, the only limiting (time independent) solution is $f_0 = 1$ and $f_k = 0$ for all $k > 0$. Note that for the case $p = q = 0.5$ the average degree of nodes in the pure duplication model does not change over time (see Lemma 3). Together with the fact that the fraction of singletons increases in time, this implies that (i) the average degree of non-singletons must increase in time and (ii) there is a single connected component of size $o(t)$ with increasing average degree. It is quite possible that this connected component of the network generated by the pure duplication model exhibits a power law with parameter $b \leq 2$, however this is difficult to establish.

In the rest of Section 2, we show that the degree distribution of the general duplication model (in fact, any random model based on duplications) is not a power law with exponential cut-off as claimed in [26]. We achieve this by showing a bound for the maximum degree of the general duplication model and contrasting it with that of a network which exhibits power law with exponential cut-off.

In [5] we proved that the general duplication model for $r > 0$ and a slightly modified version of the pure duplication model indeed achieve a power law degree distribution as per the yeast proteome network. (Due to space limitations we omit these proofs.) However, a more general measure for capturing the topological properties of a network is the ℓ -hop degree distribution for all $\ell > 0$. Under this measure (for $\ell > 1$), we show that the (modified) general duplication model is quite different from the yeast proteome network.

In Section 3, we finally present our *sequence similarity enhanced model* which is based on the observation that the interactions of sequence-wise similar pro-

teins are highly correlated. The model thus employs *sequence similarity edges* between pairs of nodes/proteins to better capture the mechanisms for updating the interactions after a duplication event. Our model not only captures the degree distribution of the yeast proteome network but also yields a much better approximation to its ℓ -hop degree distribution for $\ell > 1$. Moreover we have observed that the average clustering coefficients of networks generated by this model and the original proteome network are almost equal to each other.

1.1 Preliminaries

We first define the general duplication model formally. The general duplication model grows iteratively in discrete time steps. Let $G(t-1)$ be the network at the end of time step $t-1$. In time step t exactly one new node is generated and will be denoted as v_t . For any node v_s , we will denote its degree (or expected degree if the context is clear) at time step $t \geq s$ by $d_s(t)$.

(i) At each time step t , the new node v_t is generated by picking one of the nodes w in $G(t-1)$ uniformly at random and “duplicating” it to create v_t ; i. e. v_t will initially be connected to all neighbors of w .

(ii) The edges incident to v_t are updated through the following random process. Each edge e is considered independently and is deleted with probability q ($= 1-p$). Then, each node u which is not connected to v_t is considered independently and an edge between u and v_t is created with probability r/t .

As mentioned earlier, when $r = 0$ we have the pure duplication model; we show in the next section that it can not achieve a power law degree distribution as stated in [10]. To address this problem the pure duplication model can be modified via a new step (3) where v_t is connected to a uniformly chosen random node (either at all times or only if it had become a singleton at the end of step (2)). As a result, v_t never has degree 0.

Let $\mathbf{F}_k(t)$ denote the number of nodes of degree k at the end of step t in the random process and let $\mathbf{F}(t) = (\mathbf{F}_0(t), \mathbf{F}_1(t), \dots)$ be the degree sequence. Also let $F_k(t) = \mathbf{E}\mathbf{F}_k(t)$ be the expected value, and $f_k(t) = F_k(t)/t$ the expected fraction of nodes of degree k . Finally let $\mathbf{e}(t)$ be the number of edges in $G(t)$ and $e(t) = \mathbf{E}\mathbf{e}(t)$; similarly let $\mathbf{h}(t)$ be the average degree of a node (averaged over all nodes) in $G(t)$, and $h(t) = \mathbf{E}\mathbf{h}(t)$. We say a model has a power law degree sequence if we can find $b, c > 0$ constant such that $f_k(t) \rightarrow f_k$ as $t \rightarrow \infty$ where $f_k = (1 + O(1/k))ck^{-b}$.

2 On the General Duplication Model

This section is on the previous studies on the analysis of the general duplication model. We first show in Section 2.1 that the fraction of singletons in the pure duplication model grows with time in such a way that $F_0(t) \rightarrow t$ is the only consistent limiting solution. This implies that, unless $f_k = 0$ for $k \geq 1$ then $F_k(t) \neq tf_k$, where f_k is a time independent solution for the limiting proportion of nodes of degree k . In fact, for the particularly interesting case that $p = q =$

1/2, we show that the expected number of non singletons at time step t is between $O(\sqrt{t})$ and $O(t/\log \log t)$. This contradicts the assumption in Eqn(6) of [10]. Thus, without some modification, the pure duplication model of [10] cannot have a power law degree distribution in the form $F_k(t) \sim ct k^{-b}$ for any constants c, b .

Section 2.2 is on the analysis in [26] which predicts the general duplication model to have a degree distribution of the form ‘power law with exponential cut-off’; i. e. there exists constants a, b, c such that, as $t \rightarrow \infty$, we have $f_k(t) \sim ck^{-b}a^{-k}$ for $k \rightarrow \infty$. We show that this cannot be true by demonstrating that the expected maximum degree for a power law with exponential cut-off is $O(\log t)$ whereas the general duplication model has expected maximum degree of $\Omega(t^p)$.

2.1 Properties of the pure duplication model

Lemma 1. *The expected proportion of singletons, $f_0(t)$, in the pure duplication model is a non-decreasing function of t and tends to a limit $f_0 \leq 1$. If also we have that $f_k(t) \rightarrow f_k$ for $k \geq 1$ then $f_0 = 1$ and $f_k = 0$ for $k \geq 1$.*

Proof. We have the following recurrence for singletons in the pure duplication model:

$$F_0(t+1) = F_0(t) + \sum_{k \geq 0} \frac{F_k(t)q^k}{t}.$$

Thus writing $F_k(t) = tf_k(t)$ we have

$$(t+1)(f_0(t+1) - f_0(t)) = \sum_{k \geq 1} f_k(t)q^k \geq 0,$$

and we see that $f_0(t+1) \geq f_0(t)$. As $f_0(t) \leq 1$ it follows that $f_0(t) \rightarrow f_0 \leq 1$ from below as $t \rightarrow \infty$.

Suppose next that for some $k \geq 1$, k constant, $f_k(t) \rightarrow f_k > 0$, then $\sum_{k \geq 1} f_k q^k = c > 0$. Thus there exists T such that for $t \geq T$, $\sum_{k \geq 1} f_k(t)q^k \geq c/2 > 0$ and

$$f_0(t+1) \geq f_0(t) + \frac{c}{2(t+1)}.$$

Iterating this we get

$$f_0(t) \geq \frac{c}{2} \log t/T + O(1/T) + f_0(T)$$

i. e. , $f_0(t) > 1$ for t large enough, which is impossible. □

This lemma excludes the existence of power law solutions $f_k \sim ck^{-b}$ for finite $k \geq 1$ (which are suggested in [10]), but we cannot exclude non-limiting degree distributions by this argument.

It is possible to obtain a tighter estimate on the proportion of singletons in the network for the particularly interesting case that $p = q = 1/2$. As per

Lemma 3 (see below), this case preserves the (expected) average degree of the nodes throughout the generation of $G(t)$. Thus, $e(t) = e(0) \cdot t$ (where $e(0)$ is the number of edges of $G(0)$).

Lemma 2. *Consider the case $q = 1/2$. Let $F^+(t) = t - \mathbf{F}_0(t)$, the number of non-singleton nodes at time t and $F^+ = \mathbf{E}F^+$. Then, there are constants $c_1, c_2 > 0$ such that $c_1\sqrt{t} \leq F^+(t) \leq c_2t/\log \log t$.*

Proof. We have the following recurrence:

$$F^+(t+1) = F^+(t) + \frac{1}{t} \sum_{k \geq 0} F_k(t)(1 - (1/2)^k) \quad (1)$$

Thus:

$$F^+(t+1) = F^+(t) + \frac{F^+(t)}{t} - \frac{F^+(t)}{t} \sum_{k \geq 1} \frac{F_k(t)}{F^+(t)} \frac{1}{2^k} \quad (2)$$

As $F_1(t) \leq F^+(t)$, one can easily check $F^+(t) \geq F^+(0)\sqrt{t}$ giving the lower bound.

Now let $g(k) = 1/2^k$, which is convex and thus for any set of λ_k for which $\sum \lambda_k = 1$, we must have $\sum \lambda_k g(k) \geq g(\sum k \lambda_k)$. Now pick $\lambda_k = \frac{F_k(t)}{F^+(t)}$. We have $\sum k F_k(t) = 2e(t) = 2e(0)t$. Thus:

$$\sum_{k \geq 1} \frac{F_k(t)}{F^+(t)} \left(\frac{1}{2}\right)^k \geq \left(\frac{1}{2}\right)^{2e(t)/F^+(t)} \quad (3)$$

By substituting (3) into (2) and using $e(t) = e(0)t$ we get:

$$F^+(t+1) \leq F^+(t) + \frac{F^+(t)}{t} \left(1 - \left(\frac{1}{2}\right)^{2e(0)t/F^+(t)}\right).$$

This is only satisfied if $F^+(t) \leq c_2t/\log \log t$. This can be verified as follows. Let $c_2 = 4e(0) \log 2$. Either $F^+(t) \leq c_2t/\log \log t$, or if not we can substitute this lower bound into the exponent on the right hand side and iterate the recurrence on t to obtain a contradiction. \square

Lemma 3 (below) states that the expected number of edges is $e(t) = ct^{2p}$ and consequently the expected average degree is $h(t) = 2ct^{2p-1}$. Thus for $p < 0.5$ the average degree decreases over time and for $p > 0.5$ it increases. Only for $p = 0.5$ the average degree remains constant; however as the proportion of singletons is $\geq 1 - O\left(\frac{1}{\log \log t}\right)$ due to Lemma 2, the average degree of non-singletons (which all form a single connected component) is $\geq c \log \log t$.

Proposition 1. *The power law exponent b in [10] is given by the solution of $1 = bp - p + p^{b-1}$ and has the value 2 when $p = 1/2$. This is incompatible with $e(t) = 2e(0)t$ unless the connected component is of size $o(t)$.*

To see this, recall that $\sum kF_k(t) = 2e(t)$. Under the assumption that we have a power law degree distribution at $p = 1/2$, then $F_k(t) \sim ck^{-2}t$ and

$$e(t) = \frac{ct}{2} \sum_{k \geq 1} \left(1 + O\left(\frac{1}{k}\right)\right) k^{-1}.$$

However $\sum_{k=1}^{k^*} k^{-1}$ diverges as $k^* \rightarrow \infty$, and we cannot have $e(t) = 2e(0)t$, unless we truncate k^* at a finite value. Lemma 4 (below) sets the expected maximum degree in the pure model at $\Omega(t^p)$, and the power law assumption itself is not compatible with k^* being finite.

It is however still possible that a power law with exponent $b = 2$ holds for the connected component C . Putting $k^* = O(t^{1/2})$ we see that $\sum k^{-1} = O(\log t)$ which gives $e(t) = 2e(0)t$ provided $|C| = O(t/\log t)$, in accordance with the results of Lemma 2.

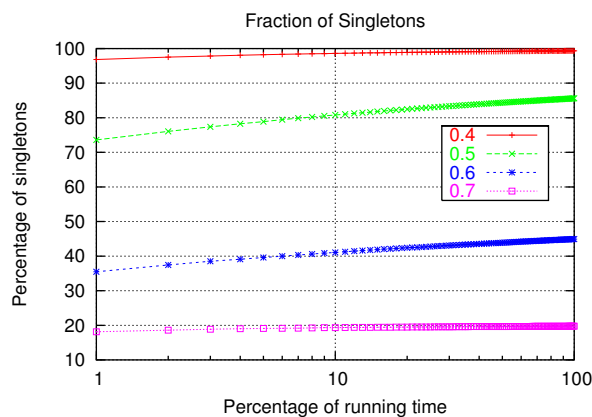


Fig. 1. Percentage of singletons in the pure duplication model as function of time (each curve is for a different value of p)

Lemma 3. *The expected total number of edges and the expected average degree of nodes at step t satisfy*

$$e(t) \sim e(0)t^{2p} \quad \text{and} \quad h(t) \sim h(0)t^{2p-1}$$

Proof. The number of edges at time $t + 1$ in terms of the number of edges at time t is

$$\mathbf{E}(e(t+1) \mid e(t)) = e(t) + \frac{1}{t} \sum_{s \leq t} pd_s(t).$$

The first term is trivial; the second term is obtained by considering the possibility that each given node v_s is duplicated at time t ; then $pd_s(t)$ would be the expected number of its edges retained. Because the sum of the degrees of all nodes is twice the number of edges, we have, taking expectations again, that

$$e(t+1) = \left(1 + 2\frac{p}{t}\right) e(t)$$

which has a solution $e(t) \sim e(0)t^{2p}$. \square

Figure 2.1 shows the percentage of the singletons in the network over the time for different values of p . The model was run until 1000000 non-singleton nodes were created. The plot uses a linear scale on the y-axis (percentage of singletons) and a logarithmic scale on the x-axis (running time).

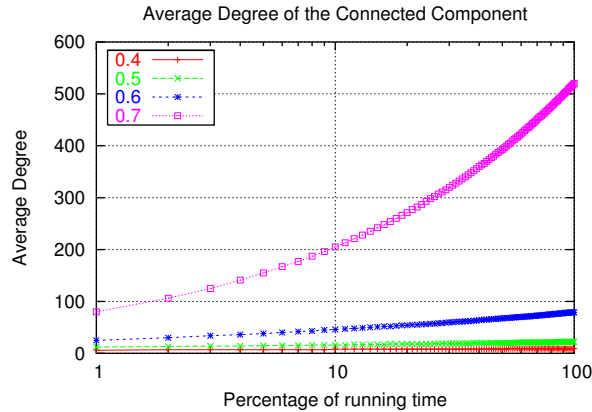


Fig. 2. Average degree of non-singleton nodes in the pure duplication model as function of time (each curve is for a different value of p)

Figure 2.1 shows the average degree over time for different values of p . Again, the model was run until 1000000 non-singleton nodes were created. The average degree of the network increases by time and the larger the value of p is, the larger is the increases of the average degree.

2.2 On the degree distribution of the general duplication model

The next lemma shows that the degree distribution of the general duplication model can not be a power law with exponential cut-off as suggested in [26].

Lemma 4. *Let $a, b, c > 0$ be constants. The degree distribution of the general duplication model cannot be in the form $F_k(t) \sim ctk^{-b}a^{-k}$ as claimed in [26].*

Proof. Denote by k_{max} , the expected maximum degree in $G(t)$. Assume an exponential cut-off i. e. $F_k(t) \sim tck^{-b}a^{-k}$. Then $\sum_{k \geq k_0} F_k(t) = o(1)$ for $k_0 > \log t / \log a$, and so $k_{max} = O(\log t / \log a)$.

On the other hand consider the expected degree of the node v_s at time $t + 1$, which is a non-decreasing function of t . Even in the worst case situation ($r = 0$) we have:

$$d_s(t + 1) = d_s(t) + \frac{d_s(t)}{t}p \quad (4)$$

as the degree of v_s can only increase if one of its neighbors is picked at time t and the edge is retained. Thus:

$$d_s(t + 1) = d_s(t) \left(1 + \frac{p}{t}\right) = d_s(s) \left(1 + \frac{p}{s}\right) \cdot \left(1 + \frac{p}{s+1}\right) \dots \left(1 + \frac{p}{t}\right)$$

Since $\log(1 + x) = x - O(x^2)$ we have

$$\exp\left(\sum_{\tau=s}^t \log(1 + p/\tau)\right) \sim \exp\left(p \sum_{\tau=s}^t 1/\tau\right) = e^{p \log(t/s)}$$

which implies that $d_s(t + 1) = \Omega(d_s(s)(t/s)^p)$ and that $k_{max} = \Omega(t^p)$ contradicting the claim. \square

We finally prove that for $r > 0$ there are no degenerate limiting solutions of the form $f_0 = 1, f_k = 0, k \geq 1$ for the general model of [26].

Lemma 5. *For any $r > 0$ constant, the general model does not have a degenerate limiting solution of the form $f_0 = 1, f_k = 0, k \geq 1$.*

Proof. We have the following recurrence for the expected number of singletons:

$$F_0(t + 1) = F_0(t) + \sum_{k \geq 0} \frac{F_k(t)}{t} q^k \left(1 - \frac{r}{t}\right)^t - \frac{r}{t} F_0(t).$$

Assuming the existence of a limiting solution $F_k(t) = f_k t$ we have (after taking limits):

$$(1 + r - e^{-r}) \cdot f_0 = e^{-r} \sum_{k \geq 1} f_k q^k.$$

If $f_0 = 1$ then $\sum_{k \geq 1} f_k q^k = 0$, but $1 + r - e^{-r} > 0$ for $r > 0$ contradicting this. \square

3 An Enhanced Duplication Model Based on Protein Sequence Similarity

The general duplication model well approximates the degree distribution of the yeast proteome network as observed previously in [26]. (In fact, we have shown in [5] that this degree distribution is simply a power law for $r > 0$; due to space

limitations this proof is omitted.) In Figure 3 we compare the degree distribution of the yeast proteome network from the Database of Interacting Proteins (DIP) [38]⁹ to that of the (modified) general duplication model with the best fitting¹⁰ parameters $p = 0.465$ and $r = 0.08$. Although the DIP database is incomplete and includes several interactions which are not commonly observed, it still provides the most comprehensive protein-protein interaction data for the yeast *S. Cerevisiae*. As observed earlier, the degree distribution of the yeast proteome network is very similar to that of the general model with the above parameters.

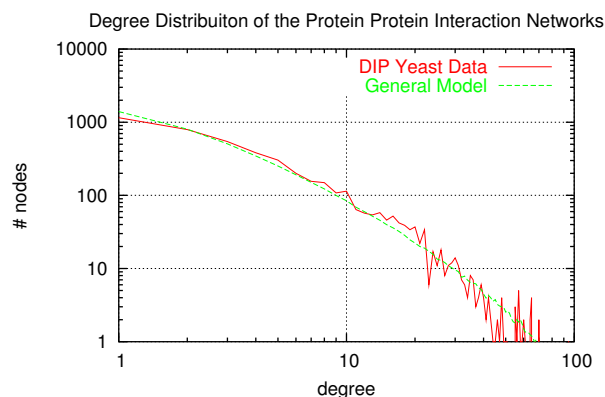


Fig. 3. The degree distribution of the proteome interaction network of the yeast and that of the general model with parameters $q = 0.535$, $r = 0.08$

The degree distribution is one possible measure for testing the structural similarity of two networks. Unfortunately structurally very different networks can have identical degree distributions. For example in an (infinite) *2-dimensional grid* all nodes have degree 4, similar to a collection of cliques of size 5. The grid obviously forms a single connected component whereas the 5-cliques are not connected at all. Thus it is desirable to use additional measures for testing the similarity of two networks more accurately.

A more refined measure of structural similarity is achieved by comparing the ℓ -hop degree distribution of the general duplication model and the yeast proteome network. In a given network, the ℓ -hop degree of a node is defined to be the total number of unique nodes it can reach in at most ℓ hops. Clearly the 1-hop degree of a node is its own degree.

⁹ The DIP yeast data has ≈ 15000 interactions among 6700 known yeast proteins. The DIP network has only 4700 of the proteins present in the network, which also means that there are about 2000 singletons in the network.

¹⁰ For all plots, the fits were achieved by calculating the average slope in both curves.

In Figure 8 we plot the average ℓ -hop degree of nodes as a function of their degree, both for the general duplication model and the yeast proteome network. By definition, the 1-hop degree distribution is a straight line with slope 1. Notice that for $\ell > 2$ the ℓ -hop degree distribution of the yeast proteome network is very different from that of the general duplication model. In fact, for $\ell > 2$, the number of nodes that can be reached by a typical node in the yeast proteome network is much higher than that observed in the general duplication model. We observed this qualitative difference for the general duplication model with all parameter choices we tested.

In order to capture the ℓ -hop degree distribution of the yeast proteome network for $\ell > 2$, we develop a more refined model that aims to emulate the divergence mechanisms in proteome network evolution more accurately. This model, which we call the *sequence similarity enhanced model*, exhibits a degree distribution very similar to that of the yeast network while also capturing its ℓ -hop degree distribution as seen in Figure 8. We provide the details of our enhanced model in the next section.

3.1 Sequence Similarity Distribution in the Yeast Proteome

A mathematical model for capturing proteome network evolution should take into account Ohno’s theory, which attributes the genome sequence growth and evolution to subsequent gene duplications followed by mutations on the gene sequences. The general duplication model implements gene duplications through a uniformly random node selection process. The mutations are implemented through random edge deletions and insertions. A more refined mutation model may take into account the sequence composition of genes and their associated proteins, and also their pairwise similarity levels.

Given two protein sequences A and B , one way to measure their similarity level through the use of their global alignment score $S(A, B)$ based on the BLOSSOM62 scoring matrix - the default scoring matrix for the best used protein alignment tools. The normalized similarity score of A and B can then be defined as

$$SN(A, B) = \frac{S(A, B)}{S(A, A) + S(B, B) - S(A, B)}.$$

Clearly $SN(A, A) = 1 = 100\%$. Note that $1 - SN(A, B)$ forms a metric, which turns out to be quite useful for our purposes.

Once the similarity between two proteins is determined via the above measure, one can depict how protein sequences relate to each other by plotting the distribution of their pairwise similarities. Such plots are provided for the yeast proteome in Figures 3.1, 5. The yeast genome was downloaded from *Saccharomyces Genome Database*[3] and the pairwise alignment of the ≈ 6700 protein coding sequences were computed via *FASTA align*[27] with default parameters. In Figure 3.1 we display the number of protein pairs whose normalized similarity score is in the range $x\% + 0.05$ for varying values of x . One can observe that the pairwise similarity distribution has a peak value at $\sim 50\%$ followed by a very sharp drop.

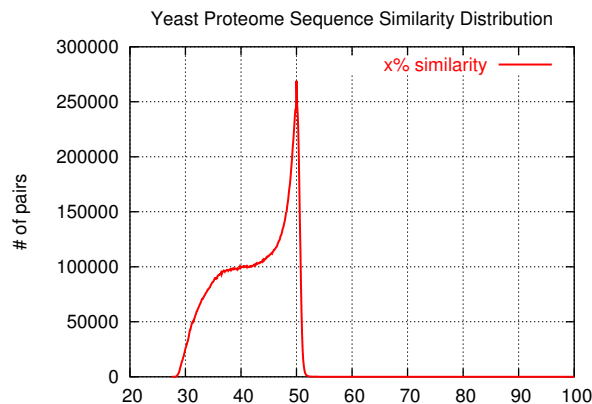


Fig. 4. Distribution of sequence similarity between pairs of yeast proteins (granularity: 0.1%)

The same distribution is depicted in a different perspective in Figure 5. Here the number of protein pairs whose normalized similarity score is *at least* $x\%$ is plotted for $x \in [20, 100]$. Observe that most pairs have a similarity score below a threshold value $\sim 50\%$ and comparatively very few pairs have a similarity score above that threshold value. The step function behavior of the normalized similarity score suggests that the pairs of proteins can be divided into two classes: protein pairs which are *similar* are the ones whose similarity scores are *above* the threshold; the other protein pairs are *dissimilar*.

Through an investigation of the the yeast proteome network we observed that sequence-wise “similar” proteins have similar interaction patterns.

More specifically, we considered all pairs of proteins A, B for which there is another protein C that is sequence-wise similar to A and which interacts with C . Among these protein pairs, the frequency of those which interact is 21 times the frequency of protein pairs that interact among *all* protein pairs.

Observation 1 *Given three proteins A, B, C , if A and B are sequence-wise similar and A interacts with C , then the chance that B interacts with C is ~ 21 times of that between arbitrary pair of proteins.*

Another observation we made was on the correlation between sequence similarities of protein triplets:

Observation 2 *Given three proteins A, B, C , if $A - B$ and $B - C$ are pair-wise similar, then with $\sim 65\%$ chance $A - C$ are similar.*

This observation is not very surprising as the normalized similarity score above forms a metric and the number of protein pairs whose similarity score is above

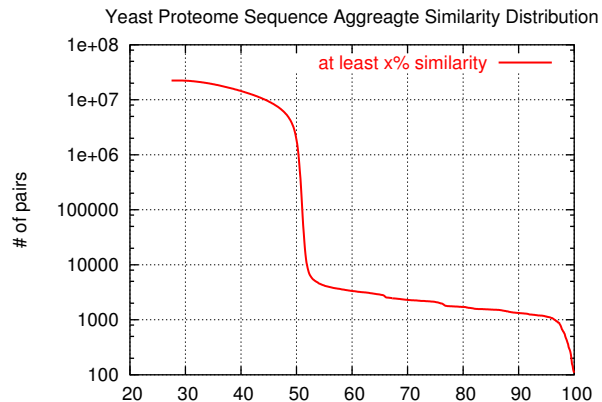


Fig. 5. Aggregate distribution of sequence similarity between yeast protein pairs. (aggregation performed from right to left).

the threshold value is distributed uniformly over the range [50% – 100%]. Nevertheless it will be quite useful in establishing our enhanced proteome growth model which we describe in the next section.

3.2 Enhanced Model Based on Sequence Similarity

Based on our observations on the sequence similarity and its implications on protein-protein interactions we develop a more refined network generation model below. Our new model, which we call the *sequence similarity enhanced model*, modifies the step for updating the interaction edges of a duplicated node through the use of additional edges indicating sequence similarity. Thus the new model has two types of edges: *interaction edges* connecting proteins that interact with each other, and *sequence similarity edges* connecting proteins that are similar.

As per the general duplication model, our sequence similarity enhanced model works in discrete time steps. Let $G(t-1)$ be the network at the end of time step $t-1$. At each time step t , a new node v_t is generated, again by picking one of the nodes w in $G(t-1)$ uniformly at random and “duplicating” it to create v_t ; i. e. v_t will initially be connected to all *similarity* neighbors and the *interaction* neighbors of v_t . The new node v_t will also be connected to w by a similarity edge. The following random process updates the similarity edges of v_t :

1. The similarity edge between v_t and w is deleted with probability δ .
2. Each remaining similarity edge is considered independently and is deleted with probability q' ($= 1 - p'$).
3. For each pair of similarity edges (v_t, u) and (u, u') , a similarity edge (v_t, u') is created with probability $(p')^2$.

Now the interaction edges of v_t are updated:

1. Each interaction edge is considered independently and is deleted with probability $q (= 1 - p)$.
2. For each node u , which is not initially connected to v_t , a new edge (u, v_t) is created independently with probability r/t .
3. For each interaction edge (v_t, u) and each similarity edge (u, u') , a new interaction edge (v_t, u') is created with probability .03 (~ 21 times the chance of having an interaction edge between an arbitrary pair of nodes - following Observation 1).

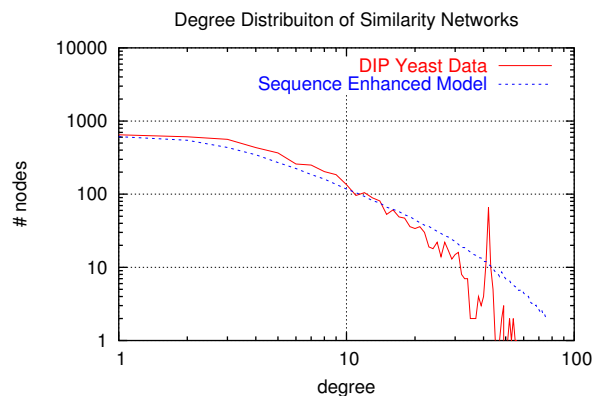


Fig. 6. The degree distribution of the proteome sequence similarity network of the yeast and that of the enhanced model with parameters $\delta = 0.7$ and $p' = 0.225$

At the time of duplication, v_t and w are sequence-wise identical and thus each similarity edge (u, w) is duplicated as (u, v_t) . Step (i) of the similarity edge update process maintains the edge (w, v_t) with probability $1 - \delta$. Here, the parameter δ is the measure of divergence. In other words, the mutation events that occurred after the duplication event are reflected on the new edge by the deletion parameter *delta*. Step (ii) maintains every other similarity edge (u, v_t) with probability p' . Finally, Step (3) imposes Observation 2 on the constructed network. The interaction edge update process, in particular Steps (i) and (ii), works similar to that in the general duplication model. The only difference is in Step (iii) where similarity edges are used to update interaction edges in order to impose Observation 1 on the constructed network.

The sequence similarity edges in the network are determined by two parameters, δ and p' . It is possible to estimate the values of δ and p' in the Yeast proteome network by fitting the *sequence similarity* degree distribution of the model with the *sequence similarity* degree distribution of the yeast proteome network. The best fitting sequence similarity degree distribution is achieved for $\delta = 0.7$ and $p' = 0.225$ and is given in Figure 6.

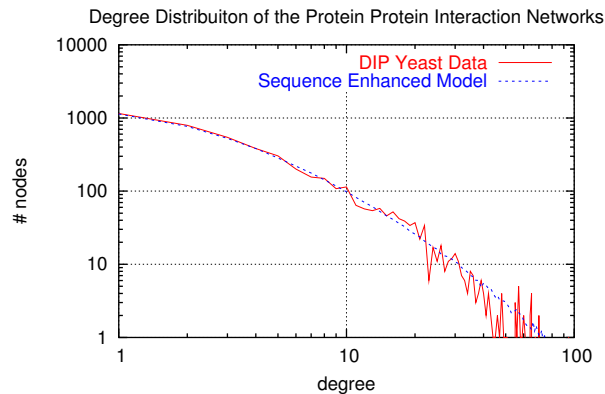


Fig. 7. The degree distribution of the proteome interaction network of the yeast and that of the enhanced model with parameters $q = 0.6$, $r = 0.1$, $\delta = 0.7$ and $p' = 0.225$.

Based on the above values of δ and p' , it is possible to estimate the other two parameters, r and p that determine the interaction edges. The best fitting interaction degree distribution of the model to that of the yeast proteome network is achieved at $q = 0.6$ and $r = 0.04$ and is given in Figure 7.

The average clustering coefficients of the Yeast proteome network, the general model and the enhanced model can be seen in Table 1. The average clustering coefficients for the models are calculated using the resulting networks that have the best fitting degree distributions with the Yeast proteome network. In Figure 8,

Table 1. The average clustering coefficients of the DIP Data, and the models

	<i>Clustering Coefficient</i>
DIP Data	0.39
General Model	0.33 ± 0.01
Enhanced Model	0.37 ± 0.01

we finally compare the ℓ -hop degree distributions of the sequence similarity enhanced model, the general duplication model, and the yeast proteome network. Our sequence similarity enhanced model accurately captures the ℓ -hop degree distribution of the yeast proteome network for all values of ℓ .

4 Conclusion

The paper first shows that the degree distribution of the pure duplication model ($r = 0$) cannot be a power law as stated in [10]. Then it shows that the degree

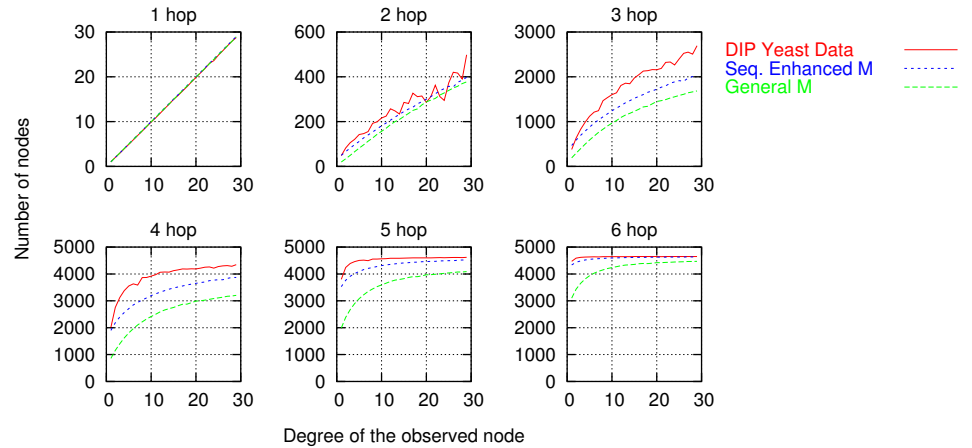


Fig. 8. The k -hop degree distribution of (i) the Yeast proteome network, (ii) the general duplication model, and (iii) the sequence similarity enhanced model. A typical node reaches all nodes in the network in 7 hops, thus ℓ -hop degree distribution for $\ell \geq 7$ is not very meaningful.

distribution of the general model can not be a power law with exponential cut-off as stated in [26]. These two problems have been addressed in [5] where the general duplication model for $r > 0$ is established to have a power law degree distribution. Unfortunately, in this paper, we observe that the general duplication model does not capture the more general ℓ -hop degree distribution of the yeast proteome network for $\ell > 1$. Thus, a new model, which takes into account the sequence similarity between protein pairs as a binary relationship, in addition to their interactions is introduced. This new model is shown to accurately capture the ℓ -hop degree distribution of the yeast interaction network for all $\ell > 0$ in addition to yielding a good approximation to the degree distribution of the yeast similarity network.

References

1. Aiello W., Chung F., Lu L., A random graph model for power law graphs, *Proc. ACM STOC*, pp 171-180, 2000.
2. Aiello W., Chung F., Lu L., Random evolution in massive graphs, *Proc. FOCS*, pp 510-519, 2001.
3. Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C. L., Binkley, G., Lane, C., Schroeder, M., Sethuraman, A., Dong, S., Weng, S., Miyasato, S., Andrada, R., Botstein, D., and Cherry, J. M. "Saccharomyces Genome Database" <ftp://ftp.yeastgenome.org/yeast/> (date of access).

4. Barabási, A.-L., Albert, R. A., Emergence of scaling in random networks, *Science* **286**, pp 509-512, 1999.
5. Bebek G., Berenbrink P., Cooper C., Friedetzky T., Nadeau J., Sahinalp S.C., The degree distribution of General Duplication Models, *Simon Fraser University Technical Report*, 2004.
6. Berger N., Bollobás, B., Borgs C., Chayes J., Riordan O., Degree distribution of the FKP network model, *Proc. ICALP*, LNCS 2719, pp 725-738, 2003.
7. Bhan A., Galas D. J., & Dewey T. G., A duplication growth model of gene expression networks, *Bioinformatics*, **18**, pp 1486-1493, 2002.
8. Bollobás, B., Borgs C., Chayes J., Riordan O., Directed scale-free graphs, *Proc. ACM-SIAM SODA*, pp 132-139, 2003.
9. Bollobás, B., Riordan, O., Spencer, J., and Tusanády, G., The degree sequence of a scale-free random graph process, *Random Structures and Algorithms*, **18**, pp 279-290, 2001.
10. Chung, F., Lu L., Dewey T.G., Galas D.J., Duplication models for biological networks, *Journal of Computational Biology*, **10**, pp 677-687, 2003.
11. Cooper C., Frieze A., A general model of webgraphs, *Random Structures and Algorithms*, **22(3)**: pp 311-335, 2003.
12. Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D., Protein interactions: Two methods for assessment of the reliability of high-throughput observations *Molecular and Cellular Proteomics* **1**:349-356, 2002.
13. Erdős, P., Rényi, A., On random graphs I, *Publicationes Mathematicae Debrecen*, **6**, pp 290-297, 1959.
14. Faloutsos M., Faloutsos P., Faloutsos C., On Power-Law Relationships of the Internet Topology, *SIGCOMM*, 1999.
15. Ferrer i Cancho, R., Janssen, C., The small world of human language, *Procs. Roy. Soc. London B*, **268**, pp 2261-2266, 2001.
16. Force A., Lynch M., Pickett F.B., Amores A., Yan Y., Postlethwait J., Preservation of duplicate genes by complementary degenerative mutations. *Genetics*, **151**, pp 1531-1545, 1999.
17. Han, J.D., Dupuy, D., Bertin, N., Cusick, M., and Vidal M., Effects of sampling on the predicted topology of interactome networks, *Nature Biotechnology* **23**, 839 - 844, (2005)
18. Ispolatov, I., Krapivsky, P.L., Yuryev, A., Duplication-divergence model of protein interaction network, *Physical Review*, E **71**, 061911, 2005.
19. Ito, T. et al., A Comprehensive two-hybrid analysis to explore the yeast protein interactome, *PNAS*, **vol. 98, no 8** pp 4569, 2001.
20. Jeong, H., Mason, S., Barabasi, A.-L. & Oltvai, Z. N., Lethality and centrality in protein networks, *Nature*, **411**, pp 41, 2001.
21. Kleinberg, J., Kumar, R., Raphavan, PP, Rajagopalan, S. and Tomkins, A., The Web as a graph: Measurements, models and methods, *Proc. COCOON*, Tokyo, Japan, pp 1-17, 1999.
22. Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., Upfal E., Stochastic models for the web graph, *Proc. FOCS* pp 57-65, 2000.
23. Nadeau, J.H., Sankoff D., Comparable Rates of Gene Loss and Functional Divergence After Genome Duplications Early in Vertebrate Evolution, *Genetics*, **147**, pp 1259, 1997.
24. van Noort, V., Snel, B., Huymen, M. A., The yeast coexpression network has a small-world scale-free architecture and can be explained by a simple model, *EMBO Reports*, **Vol. 5**, No. 3, 2004.

25. Ohno, S., *Evolution by gene duplication*. Berlin: Springer, 1970.
26. Pastor-Satorras, R., Smith, E., and Sole, R.V., Evolving protein interaction networks through gene duplication, *J. Theor. Biol.* **222**, pp 199-210, 2003.
27. Pearson, W. R., Lipman, D. J. "Fasta" <ftp://ftp.virginia.edu/pub/fasta/> (data of access).
28. Przulj, N., Corneil, D. G., and Jurisica, I., Modeling Interactome: Scale-Free or Geometric?, *Bioinformatics*, **vol.20**, number 18, pages 3508-3515, 2004.
29. Redner, S., How Popular is Your Paper? An Empirical Study of the Citation Distribution, *Eur. Phys. Jour.* **B 4**, pp 131-134, 1998.
30. Seoighe C., Wolfe K.H., Yeast genome evolution in the post-genome era. *Current Opinion in Mol. Biol.*, **2**, pp 548-554, 1999.
31. Seoighe C., Wolfe K.H., Updated map of duplicated regions in the yeast genome. *Gene*, **238(1)**, 253-61, 1999.
32. Simon, H. A., On a class of skew distribution functions, *Biometrika*, **42**, pp 425-440, 1955.
33. Uetz, P. L. et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces Cerevisiae*, *Nature*, **403**, pp 623-7, 2000.
34. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A., Modelling of protein interaction networks, *Complexus* **1**, 38-44, 2003.
35. Wagner, A., The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.* **18**, pp 1283-1292, 2001.
36. Watts, D. J. & Strogatz, S. H., Colective dynamics of small-world networks, *Nature*, **393**, pp 440-442, 1998.
37. Wolfe K.H., Shields D.C., Molecular evidence for an ancient duplication of the entire yeast genome, *Nature*, **387**, pp 708-713, 1997.
38. Xenarios, I. et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.* **30**, pp 303-305, 2002.