Protein Interactions & Molecular Networks

Generative Probabilistic Models for Protein-Protein Interaction Networks – The Biclique Perspective

Regev Schweiger[1,*], Michal Linial[2,3,*] and Nathan Linial[1,3,*]

[1]School of Computer Science and Engineering, The Hebrew University, Jerusalem, 91904 Israel. [2]Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, 91904 Israel. [3]The Sudarsky Center for Computational Biology, The Hebrew University, Jerusalem, 91904 Israel

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Much of the large-scale molecular data from living cells can be represented in terms of networks. Such networks occupy a central position in cellular systems biology. In the protein-protein interaction (PPI) network, nodes represent proteins and edges represent connections between them, based on experimental evidence. PPI networks are rich and complex, so that a mathematical model is sought to capture their properties and shed light on PPI evolution. The mathematical literature contains various generative models of random graphs. It is a major, still largely open question, which of these models (if any) can properly reproduce various biologically-interesting networks. Here we consider this problem where the graph at hand is the PPI network of Saccharomyces cerevisiae. We are trying to distinguishing between a model family which performs a process of copying neighbors, represented by the Duplication-Divergence (DD) model, and models which do not copy neighbors, with the Barabási-Albert (BA) preferential attachment model as a leading example.

Results: The property of the network that we observe is the distribution of maximal bicliques in the graph. This is a novel criterion to distinguish between models in this area. It is particularly appropriate for this purpose, since it reflects the graph's growth pattern under either model. This test clearly favors the DD model. In particular, for the BA model the vast majority (92.9%) of the bicliques with both sides ≥4 must be already embedded in the model's seed graph, whereas the corresponding figure for the DD model is only 5.1%. Our results, based on the biclique perspective, conclusively show that a naïve unmodified DD model can capture a key aspect of PPI networks.

Supplementary information: Supplementary data are available online.

1 INTRODUCTION

Many real-life systems can be modeled as complex networks, or graphs, of interacting components. Although the study of such large-scale networks is not new, there has recently been much renewed interest in this field. This is due to technological advances of two types: (i) In collecting data which depict large networks in high-resolution detail, and (ii) In the development of computational tools for the analysis of data. Among the well-studied examples of such networks are the World Wide Web, citation networks, neuronal connections, metabolic networks, ecological webs and more (for reviews see (Albert and Barabási, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003)).

One of the best studied network types is the Protein-Protein Interaction (PPI) Network. Protein interactions play a crucial role in the execution of biological functions in any model system. Accordingly, a systematic mapping of PPI on the scale of the whole proteome – the interactome – has become recently available for major model systems from yeast (Ito*, et al.*, 2001; Yu*, et al.*, 2008) to human (Gandhi*, et al.*, 2006). Although far from complete, PPI network mappings have revealed topological and dynamic features of the interactome that are common to numerous model systems (Gandhi*, et al.*, 2006). However, the overall size of interactomes differs substantially between human and other multicellular model organisms, including the Drosophila melanogaster of the Caenorhabditis elegans (Stumpf*, et al.*, 2008).

Protein-Protein Interactions are usually represented by a graph (network). Every node in such a network corresponds to a protein, and there is an edge between two nodes, if the two corresponding proteins interact physically. Such networks have been mapped for several organisms using high-throughput (HTP) techniques such as Yeast-2-Hybrid Systems (Y2H) (Ito*, et al.*, 2001; Krogan*, et al.*, 2006) and Co-Immunoprecipitation (Gavin*, et al.*, 2002; von Mering*, et al.*, 2002). High-throughput techniques are prone to a high rate of false positives and false negatives.

Saccharomyces cerevisiae (baker's yeast) is the organism with the most comprehensive, high-coverage interactome mapping. Other PPI networks, such as those of E. coli and H. pylori were investigated as well, but they are far from being complete (Rain*, et al.*, 2001). PPI networks human, Drosophila melanogaster and Caenorhabditis elegans are more complex, due to their multicellular nature. Specifically, interactomes are harder to define, as they vary between different cell types, and subsequently, harder to map.

The PPI network in yeast was refined over the years, followed by a critical assessment of data quality (Bader and Hogue, 2002) Chua, 2008 #19}. While the topology of the partial PPI network obtained from each of the main technologies is different (Yu*, et al.*, 2008),

*To whom correspondence should be addressed.

the combination of PPI data from such complementary experimental technologies resulted in a near complete map (Reguly, *et al.*, 2006). For most organisms, the interaction data are still partial and thus topological assessment of these networks is prone to sampling bias (Han, *et al.*, 2004).

It is not obvious how to infer direct pair-wise interactions from high-throughput techniques that focus on protein complexes (i.e., co-immunoprecipitation). This is especially challenging when complexes of several proteins are considered. To document and describe protein interactions, several databases have been compiled, offering curated data from various sources (Guldener, *et al.*, 2006; Salwinski, *et al.*, 2004; Stark, *et al.*, 2006; Xenarios and Eisenberg, 2001).

A random graph model is a probability space of graphs. It is often described in terms of a random process that generates graphs. Unfortunately, the most thoroughly studied classical random graph model - the Erdős-Rényi (ER) model (Erdös and Rényi, 1959) - does not capture the properties of PPI networks. This gave rise to numerous attempts at defining other random models that generate graphs which are more reminiscent of the PPI graphs encountered in nature. We focus here on two families of random models which have received considerable attention in recent literature, the preferential attachment model (or the Barabási-Albert (BA) model); and the Duplication-Divergence (DD) model.

How does one test the agreement between a random graph model and a given set of data? It appears to be computationally hard to determine the exact probability that a particular network is generated by a given model. Therefore, attempts to fit a network to a model are usually realized by calculating certain statistics of the network, and comparing it to predictions derived from a model. Much attention was given to the *degree distribution* of the network in question (Recall that the degree of a node is the number of neighbors it has in the graph). In particular, it was often claimed that the degree distribution is governed by a power-law (but see (D'Souza, *et al.*, 2007; Deeds, *et al.*, 2006; Khanin and Wit, 2006; Lima-Mendez and van Helden, 2009; Mitzenmacher; Reiko, *et al.*, 2005; Stumpf, *et al.*, 2005; Stumpf, 2005)). Namely, that $P_{deg}(k)$ - the probability that a node has k neighbors is proportional to $k^{-\gamma}$, where γ is a network-dependent constant. Other important parameters are counts of fixed subgraphs. This approach has a strong theoretical underpinning in recent work such as (Lovász and Szegedy, 2006). This general idea is materialized in a variety of concrete ways: Small connected subgraphs (Hormozdiari, *et al.*, 2007; Pržulj, 2004), dense subgraphs (Colak, *et al.*, 2009), tree subgraphs (Alon, *et al.*, 2008), local walks (Middendorf, *et al.*, 2004), and *k*-hop reachability (Hormozdiari, *et al.*, 2007). Other measures include centrality, betweenness and more.

Motivated by the apparent power-law behavior, the *preferential attachment* model (aka the *Barabási-Albert* (*BA*) model) was introduced (Barabási and Albert, 1999). In this generative model, we start with a seed graph, and iteratively add new nodes to the graph. Every new node is linked by m edges to previously occurring nodes, where this set of neighbors is selected non-uniformly. Concretely, the probability of connecting to a given existing node is proportional to the degree of that node (see Fig. 1A). As noted in (Bollobás and Riordan, 2005), this description of the model is still incomplete. We therefore adopt the formulation in (Bollobás and Riordan, 2005) as our realization of the *BA model*. This model indeed produces a degree distribution where $P_{deg}(k) \sim k^{-\gamma}$.

The *Copying model* or *Duplication-Divergence* (*DD*) model was first suggested in attempting to explain the structure of the world-wide-web graph (Kumar, *et al.*). It was later adopted for the analysis of PPI network (Bebek, *et al.*, 2006; Pastor-Satorras, *et al.*, 2003). Here, too, we start from a seed graph that undergoes a growth process as follows. At every step, we randomly and uniformly select an existing node, and duplicate it, i.e. create a new node with an identical set of neighbors. This is followed by a random modification: Each new edge is retained with probability p and is omitted with probability 1-p. In addition, we connect each existing node to the new node with a probability r (see Fig. 1B. also see Supporting Information for details on both models). This model yields a power-law degree distribution as well. Moreover, node duplication can represent gene duplication, and random edge insertions and deletions are analogous to random mutations (Ohno, 1970).

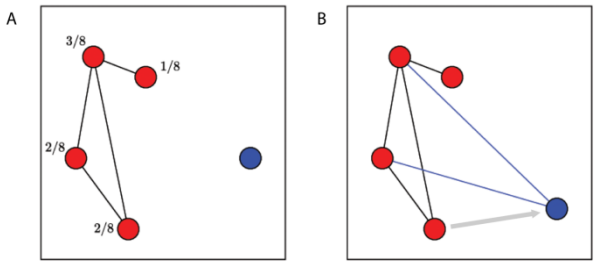

Fig. 1. Model illustrations. (A) Barabási-Albert model. A new node (blue) is added at each step. For each of its outgoing edges, the probability of connecting to a given existing node is proportional to the degree of that node. (B) Duplication-Divergence model. A new node (blue) is added at each step by duplicating an older node. This is followed by a random insertion and deletion of edges.

Here, we propose using counts of (non-induced) *maximal bipartite cliques*, or *maximal bicliques*, as a method to distinguish between *Duplication-Divergence*-type models and *Barabási-Albert*-type models. In the network that we consider, V is the set of nodes (proteins) and E is the set of edges (interactions). In this context, a *biclique (A,B)* is a subgraph consisting of two disjoint sets of nodes ($A, B \subseteq V$), where all edges between these two sets exist in the graph ($A \times B \subseteq E$). We do not impose any conditions on the existence or non-existence of edges between nodes in each set. A biclique is *maximal* if there is no other biclique containing it, i.e, there are no $A \subseteq A_2, B \subseteq B_2$, such that $A_2 \times B_2 \subseteq E$.

Bicliques are naturally related to the *DD* model. The introduction of a new node w that is a duplicate of an old node v, creates the biclique (*(v,w)*, *N(v)*). Further duplication of v or w, yields an even larger biclique. Even with random insertions and deletions of interactions, it is clear that essentially, it is the process of copying neighbors that gives rise to dense bipartite graphs, and specifically large bicliques. In contrast, in models like the *Barabási-Albert* model, where the neighbor sets of different nodes are uncorrelated, there is no apparent reason why large bicliques should occur. Indeed, the introduction of the *DD* model was motivated by the ubiquity of large bipartite graphs in the web graphs (Kumar, *et al.*). It was also noted that large dense bipartite graphs exist in the yeast PPI network (Bu, *et al.*, 2003), which were used to infer protein

interactions and find binding motifs (Li, *et al.*, 2006). For these reasons, bicliques are natural candidates for distinguishing between different models.

2 METHODS

2.1 Data of protein-protein interactions

Protein Interactions were extracted from DIP (Salwinski, *et al.*, 2004) version 30/12/2009. The database contains experimental data covering 5033 proteins of the *Saccharomyces cerevisiae* (baker's yeast) and 22,118 interaction edges, originating from 16,444 experiments. Note that ~90% of the interactions are reported by HTP experiment techniques of the Y2H and Co-immunoprecipitations. Additional methods of low throughput include X-ray crystallography, native gels, cross-linking study and various affinity chromatography technologies.

2.2 Parameter Enumeration and Optimization

Parameter optimization was performed in two stages: First a coarse search, and later a refinement of the better-performing parameters in higher resolution. Each parameter set was used to generate >20 graphs.

2.3 Seed graph models and parameters.

Geometric model: m_0 points in R^d are sampled at random, each coordinate independently, from the standard normal distribution, with $x_j^{(i)} \sim N(0,1)$, for $i \in \{1,...,n\}, j \in \{1,...,d\}$. Each node i in the seed graph corresponds to a point x_i. There is an edge between nodes i and j if their corresponding points are closer than a radius ρ: $\|x_i - x_j\| \le \rho$.
Inverse geometric model: Similar to the geometric model, except that nodes are connected if the distance between their corresponding points is larger than a radius R: $\|x_i - x_j\| \ge R$.
Erdős-Rényi model: There are m_0 nodes, and each possible edge between two nodes exists uniformly at random with a probability p. For a schematic representation of the seed graph models see Fig. 4.
Seed graph sizes were tested for 10, 20, 50, 80, and 100. Seed graph size was allowed a change of ±10 at the later optimization stage. For the *geometric* and *inverse geometric models*, dimensions were tested for 2, 4, 6, 8 and 10, and were allowed a change of ±1 at later optimization stage. Radii were tested for 1, 2, 3, 4 and 5, and were allowed a change of ±0.5. For the *Erdős-Rényi model*, p was for values between 0 and 1 (in intervals of 0.1). For earlier connections between PPI graphs and geometric graphs, see (Pržulj, 2004).

2.4 Models

Duplication-Divergence model. We follow the formalization of (Bebek, *et al.*, 2006). To create a graph of n nodes, we begin with a seed graph of size m_0. We then perform n-m_0 iterations. In every iteration t, we uniformly select a node v from the previous nodes, [0,…,t-1]. For each of the nodes w in N(v), the set of neighbors of v, we create an edge between w and the new node t with probability p, or delete it with probability 1-p. Then, for each of the nodes u in [0,…,t-1], we connect u and t with a probability r/t. We then merge parallel edges.
Barabasi-Albert Model. We follow the formalization of (Bollobás and Riordan, 2005). To create a graph of n nodes, we begin with a seed graph of size m_0. We then perform n-m_0 iterations. In every iteration t, we add m edges from t sequentially. At each edge addition, we assign a probability of $deg(v)/C$ for all vertices v in [0,…,t-1], and $(deg(v)+1)/C$ for the node t itself, where C is a normalizing factor to make a valid probability distribution, and the degree included all edges previously introduced. Parallel edges are merged.

2.5 Comparison between distributions

For two biclique distributions p and q, and sizes 2≤n≤m, we define their *l2-distance* as

$$d(p,q) = \sqrt{\sum_{2 \le n \le m} \left(\log_{10}\left(p_{(n,m)}+1\right) - \log_{10}\left(q_{(n,m)}+1\right)\right)^2}$$

The +1 correction is applied to allow comparison between distributions with different biclique sizes.

2.6 Parameter enumeration

For DD models: p was tested for values between 0 and 1 (in intervals of 0.02); r was also tested for these values (although r could potentially be larger than 1, we did not test for these values as they generate too many edges, and introduce too much randomness into the graph which reduces that impact of the essence of the DD model). For *BA* models: m was tested for sizes from 1 to 40. For *DD* and *BA* models, sets of parameters giving less than 50,000 edges in average were discarded.

2.7 Enumeration of bicliques and implementation

Biclique exhaustion and enumeration was performed with the LCM-MBC algorithm and the FP-MBC software (Li, *et al.*, 2007). Graph generation, biclique enumeration and parameters optimization were all performed in C++, and run on a linux cluster of 250 processors with Sun Grid Engine, for approximately a week per model.

2.8 Biological inference

Bicliques were separated to left and right set (example in Fig. 2). The association of biological terms for with each set was performed according to the BioAsosciation protocol (Jenssen, *et al.*, 2001). Sets of proteins were tested as a group in an enrichment protocol for GO annotation enrichment using DAVID (Huang da, *et al.*, 2009). Only statistical significant annotations are reported (p-value <0.05).

3 RESULTS

3.1 Bicliques in the PPI networks are of biological relevance

We analyzed PPI network of *Saccharomyces cerevisiae* through the perspective of bicliques. Bicliques, especially the larger ones, often have a clear biological significance. Consequently, they capture an important essence of the entire PPI network (Sharan, *et al.*, 2007; Ulitsky and Shamir, 2007).
To illustrate: Fig. 2 shows a biclique of size (7,7) from the yeast PPI graph. In this case, there are no edges inside each set. As a first examination, we can check for GO annotation keyword enrichment in each set (Ashburner, *et al.*, 2000).
Statistical analysis based on literature gene associations (Jenssen, *et al.*, 2001) of the left set indicates enrichment for keywords of "recombination" (corrected p-value of 9.86e-33) and "sporulation" (corrected p-value of 3.22e-148). Actually, the proteins are strongly related to DNA repair and to events that take place at mitosis and meiosis. Proteins in the right set participate in folding and chaperone activity. Specifically, included are proteins related to stress dependent folding, peptide transport to mitochondria and complex formation. All these proteins are highly abundant and participate in very fundamental generic cell processes.

A biological interpretation for the two sets can be suggested. Specifically, it shows that genes acting in the control of DNA repair and in mitosis (left set, Fig. 2) are linked to stress-dependent chaperones and protein import machinery (right set, Fig. 2). Cellular mechanisms activated by DNA damage and by protein misfolding are interconnected and share common elements (Kultz, 2005).

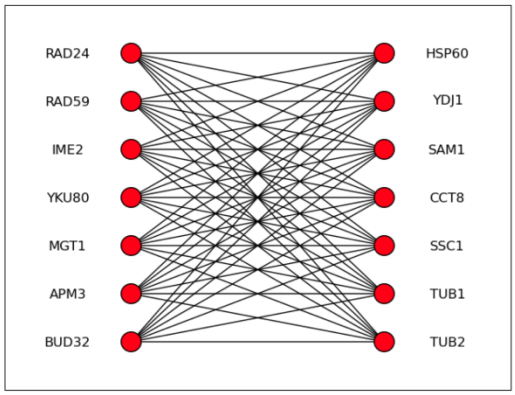

Fig. 2. An example of a biclique from the yeast PPI graph. The biqlique of size 7,7 is shown. Proteins are indicated by their Gene names. Proteins on the left set are related to DNA repair and to events that take care at mitosis and meiosis. Proteins on the right set participate in folding and chaperone activity. Additional bicliques of size 7,7 are analyzed in Supporting information, Table S1.

We performed GO term enrichment analysis for additional bicliques of similar properties (size 7,7; 24 bicliques). Most of them could indeed be associated with biological phenomena. Sets of genes enriched in DNA related activities (e.g., DNA repair, Cell cycle, Telomere maintenance) and sets of genes involved in protein maintenance (e.g., Unfolded protein binding, Stress response, Protein refolding. Chaperone) consist the two sides in the analyzed bicliques (Supporting Information, Table S1). Such crosstalk also prevails in the analysis of the gene expression programs of *Saccharomyces cerevisiae* under multiple environmental changes (Gasch*, et al.*, 2000).

3.2 The Yeast PPI has many bicliques of various sizes

We analyzed the distribution of bicliques in the PPI network of *Saccharomyces cerevisiae*, which consists of 5033 proteins and 22,118 interactions, obtained from the DIP database (Salwinski*, et al.*, 2004). Each biclique is uniquely defined by the sizes of its two disjoint node sets. So for each possible pair of sizes n,m ($2 \le n \le m$), we counted the number of maximal bicliques of size (n,m) in the graph.

A few observations are evident. The graph indeed contains many bicliques – a total of 126,584 distinct bicliques of all sizes (Figs. 3A,3B and Supporting Information Fig. S1). These bicliques range in size from (12,12) to (2,70). The most abundant size is (5,6), with 6676 (5.2% of total) bicliques of that size.

3.3 *DD* model yields a better fit than *BA* with a minimal seed

We next proceeded to generate graphs from the *DD* model and from *BA*, with a variety of parameters, and compare the distribution of maximal bicliques in these graphs to that of the real graph. To eliminate effects resulting from the seed graph, we ran both models with a minimal seed consisting of a complete graph on 3 nodes.

In principle, it is easy to create as many bicliques as desired, through an appropriate choice of (extreme) parameters for the models. If in the *DD* model, we let p be very small and r very large, dense graphs will result, with many bicliques. The same applies to a very large value of m in the *BA* model. It is therefore prudent to limit the possible range of parameters to values that yield a graph with a number of edges that is in the same order of magnitude of the real PPI graph. We therefore limit ourselves only to parameters that generate graphs with an average number of 50,000 edges (roughly twice the number of edges in the yeast PPI graph). The rationale behind such choice is the following: We expect that with the improvement of experimental techniques, more protein-protein interactions will be discovered, including those that are fragile and transient. In other words, each PPI graph should be considered as only an approximation of the true full PPI graph. However, our results seem quite insensitive to this specific choice of the limit of the number of edges (ranging from 10,000 to 100,000) for both models.

In general, it is evident that graphs generated by the *DD* model indeed contain a large number of bicliques, and those bicliques tend to have relatively large sizes. In contrast, graphs generated by the *BA* model have fewer bicliques, and those tend to be smaller. These observations are in line with our understanding of the emergence of bicliques, as previously discussed.

We compared the distribution of bicliques in model-generated graphs to that of the real yeast PPI graph. The distance between two biclique distributions was defined as the *l2* distance of the distributions of the base-10 logarithm of the values. We searched for sets of parameters for each model that generate a distribution as similar as possible to that of the real graph.

For the *BA* graph, we find that the best parameter is $m=10$, giving 49,691±40 edges (mean±std). It is evident that despite the large number of edges, the sizes of bicliques grow more slowly than in the real yeast PPI graph, reaching only (5,7) on average (albeit a longer tail for graphs with a smaller size – up to (2,157)). The l2 distance to the real graph is 23.39±0.44 (Fig. 3C).

For the *DD* graph, the best parameters were found to be $p=0.6$ and $r=0.3$, giving 43,251±13,379 edges. Although this graph has roughly the same number of edges as the BA graph, we find many more bicliques, whose sizes now vary up to (9,10), a wider range, closer to the real graph (also with a long tail, up to (2,265)). The l2 distance to the real graph in this case is 12.94±0.62, which is significantly closer (Fig. 3D).

We conclude that both the sizes and the number of bicliques in the *DD* model substantially exceed those of the *BA* model. Moreover, with an optimal choice of parameters for both models, the *DD* model fits the real yeast PPI graph significantly better than the *BA* graph. Since we operate here with a degenerate, minimal seed, it seems that this property is inherent in the graph evolution method itself, and not with the choice of seed.

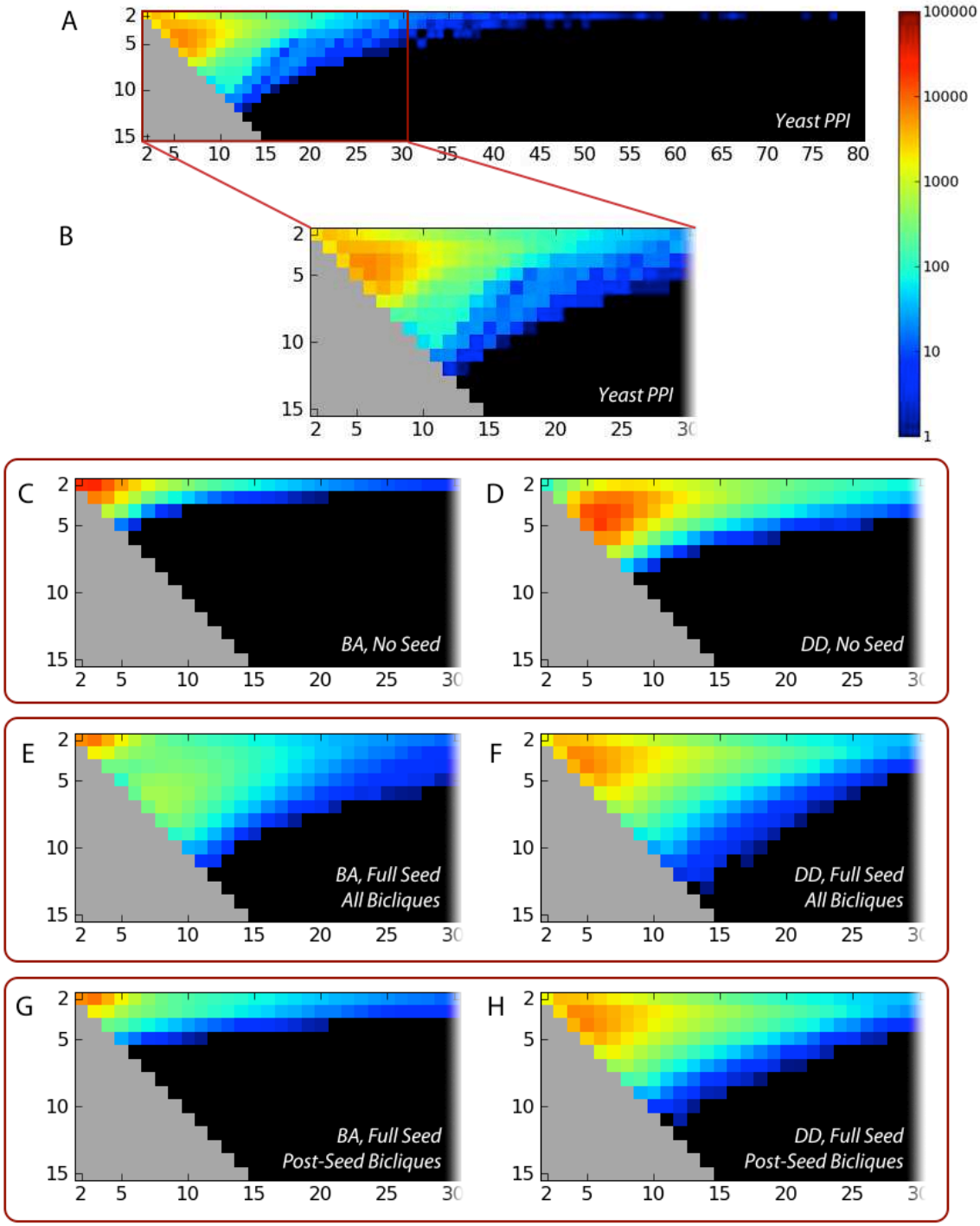

Fig. 3. Biclique distributions. A biclique is uniquely defined by the sizes of its two disjoint node sets. For each possible pair of sizes n,m ($2 \leq n \leq m$), the number maximal bicliques of size (n,m) exist in the graph is shown at a log-scale. The gray area corresponds to n>m, which is null by definition. This histogram is shown for: (A-B) The real Protein-Protein Interaction graph of Saccharomyces Cerevisiae; (C) The best fit of the Barabási-Albert (BA) model, with a degenerate seed graph; (D) The best fit of the Duplication-Divergence (DD) model, with a degenerate seed graph; (E) The best fit of the BA model, with a full seed graph; (F) The best fit of the DD model, with a full seed graph; (G-H) The same as E-F, except that bicliques that are fully contained in the seed graph are omitted from the count (see also Supporting Information, Fig. S1).

3.4 *DD* model yields a better fit than *BA* with a larger seed

Having established that with a small seed, *DD* outperforms *BA*, we checked the possibility of using a more complicated seed. It follows quite easily from the definition of the two models, that their results are highly dependent on the initial conditions. We therefore set out to see whether a better choice of a seed graph will improve the fit for any of the models.

We only considered seed graphs of up to ~100 nodes. Three different methods were used for the generation of seed graphs: (1) A *random geometric* graph. Here random points are sampled from a standardized normal distribution on R^d. Each node corresponds to a point, and two nodes are connected in the graph if the corresponding two points are at distance smaller than a parameter ρ. In this model the neighbor sets of two conncted vertices are positively correlated (Fig. 4A); (2) An *inverse geometric model,* which is similar to the geometric model, but where two nodes are connected when the corresponding points are at distance R or above. This model tends to create large induced initial bipartite graphs (Fig. 4B); (3) An *ER* graph. In this model edges are independent of each other (Fig. 4C).

We repeated the previous scheme of finding the parameters for both *DD* and *BA*. In both cases, the best results were achieved using the *inverse geometric model*, possibly due to the large number of bicliques it inherently contains.

We find that for the *BA* graph, the best parameters are $m=6$, $d=6$ and $R=4.25$, with a seed size of 110, giving 47,286±9092 edges (mean±std). The l2 distance to the real graph is 12.35±3.08 (Fig. 3E). Although larger bicliques emerge, only bicliques of smaller sizes make substantial contributions to the total count. The *BA* algorithm does not tend to expand bicliques. Indeed, for bicliques in the size range of $4\leq n\leq m$, the fraction of bicliques that are already fully contained the in seed graph is 92.9% (Fig. 3G). It appears then, that in order to achieve a similar distribution of bicliques, it is crucial to start from an extremely large seed, and even so, most of the larger bicliques do not expand beyond their original size in the seed graph.

For the *DD* graph, the best parameters were found to be $p=0.3$ and $r=1.05$, $d=2$, $R=1.5$ with a 40 nodes seed 40, giving 19,463.76±1307.77 edges, close to the number of edges in the real PPI network. Here the l2 distance to the real graph is significantly smaller: 7.15±1.90, (Fig. 3F).

Compared with its *BA* counterpart, only 5.1% of bicliques in the range $4\leq n\leq m$ reside fully inside the seed (Fig. 3H). This is in agreement with the observation that the majority of subgraphs are generated in the duplication process. Thus, a much closer distribution is achieved, with a smaller seed graph, and much more of the graph structure is generated by the process rather than being embedded in the seed. A comparison of the different statistics for the *BA* and *DD* models is shown in Table 1.

Table 1. Comparison of different statistics for best BA and DD models, with and without a seed.

	# Edges, Best BA Model (sd)	# Edges, Best DD Model (sd)	Dis-PPI, Best BA Model (sd)	Dis-PPI, Best DD Model (sd)
No Seed	49,691 (40)	43,251 (13,379)	23.39 (0.44)	12.94 (0.62)
With Seed	47,286 (9092)	19,463 (1307)	12.35 (3.08)	7.15 (1.9)

Shown are number of edges; l2 distance to the yeast PPI graph; and the percentage of bicliques that fully reside inside the seed. The yeast PPI graph contains 22,118 edges. Dis-PPI, l2 distance to yeast PPI. BC, bicliques sd, standard deviation.

4 DISCUSSION

The proper modeling of PPI networks is a major challenge from both perspective: the theoretical-mathematical and the biological one. In the search for the best model to explain experimental data, one should be aware of several pitfalls originating in the methodology or in the data itself.

A key source of difficulty is the quality of available data. In particular, PPI data originates from several different sources of varying levels of quality and reliability. The systematic curation of such data provides at least a partial remedy. Much of the yeast PPI network information originates from Y2H experiments. This method suffers from a high rate of false positive (Bader and Hogue, 2002; Uetz, *et al.*, 2000). However, with advances in experimental design (Vermeulen, *et al.*, 2008; Yu, *et al.*, 2008), con-

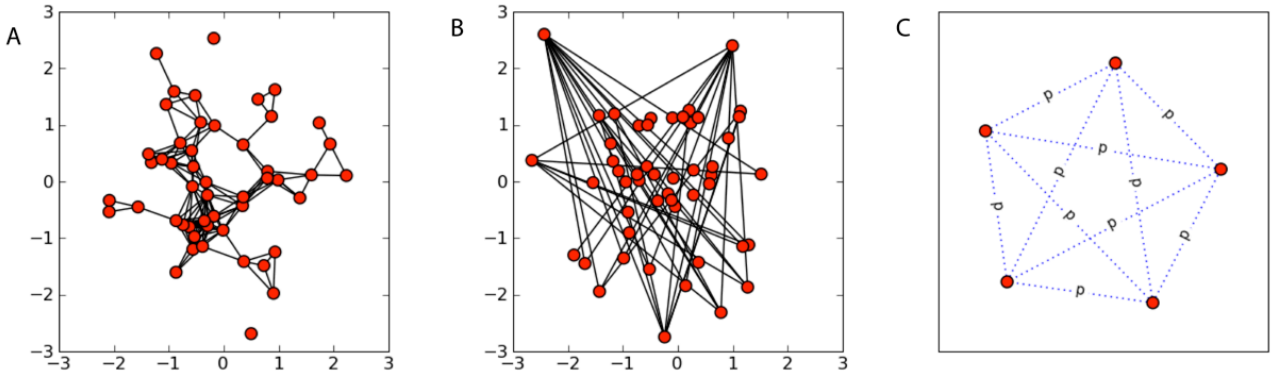

Fig. 4. ***Seed model illustrations.*** (A) *Random Geometric model*. Random points are sampled from a standardized normal distribution on R^d (here $d=2$). Each node corresponds to a point, and two nodes are connected in the graph if the corresponding two points are at distance smaller than ρ. (B) *Inverse Random Geometric Model.* Similar to the geometric model, except two nodes are connected when the corresponding points are at distance R or above (here $d=2$). (C) *Erdős-Rényi model.* Every edge is independently inserted at a probability p.

sistently contaminating proteins get eliminated. Thus, the quantity, quality and reliability of PPI networks have drastically improved (Yu*, et al.*, 2008). In addition, large-scale co-immunoprecipitation experiments allow the incorporation of labile protein interactions into PPI networks (Schulze and Mann, 2004).

To test the validity of our conclusions, we have repeated the described protocol by omitting the Y2H data, leaving more than 50% of the data mainly extracted from co-immunoprecipitation experiments, X-ray and SILAC-based TAP technologies. This subset shows essentially the same results.

We use the distribution of maximal bicliques in a graph as a yardstick against which to compare different generative models of graphs. We have investigated the PPI network of *Saccharomyces cerevisiae*. It transpires that the graphs generated by the *Duplication-Divergence* model are in much better agreement with the actual network than those generated by the *Barabási-Albert* model. Clearly both models can be expanded and refined in various ways, and we have restricted ourselves to the basic versions of either model. Still, we believe that our findings are rather indicative of the general picture. Indeed, other closely related models have also been shown to give rise to many large bicliques (Kumar*, et al.*).

In conclusion, we have suggested a new perspective on the question of PPI network modeling. The distribution of maximal bicliques, is not only an intuitive method to distinguish between models, but also that it is effective and decisive. Our results, based on the biclique perspective, conclusively show the ability of the DD model to capture a key essence of PPI networks.

ACKNOWLEDGEMENTS

RS was awarded a fellowship from the SCCB, the Sudarsky Center for Computational Biology.

Funding: This study is partially supported by Prospects (EU, Framework VII) and the Israel Science Foundation ISF 592/07.

REFERENCES

Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks, *Reviews of Modern Physics*, **74**, 47.

Alon, N.*, et al.* (2008) Biomolecular network motif counting and discovery by color coding, *Bioinformatics*, **24**, i241-i249.

Ashburner, M.*, et al.* (2000) Gene Ontology: tool for the unification of biology, *Nature Genetics*, **25**, 25-29.

Bader, G. and Hogue, C. (2002) Analyzing yeast protein–protein interaction data obtained from different sources, *Nature Biotechnology*, **20**, 991-997.

Barabási, A.-L. and Albert, R. (1999) Emergence of Scaling in Random Networks, *Science*, **286**, 509-512.

Bebek, G.*, et al.* (2006) The degree distribution of the generalized duplication model, *Theoretical Computer Science*, **369**, 239-249.

Bollobás, B. and Riordan, O.M. (2005) *Mathematical results on scale-free random graphs*. Wiley-VCH Verlag GmbH & Co. KGaA.

Bu, D.*, et al.* (2003) Topological structure analysis of the protein-protein interaction network in budding yeast, *Nucleic Acids Research*, **31**, 2443-2450.

Colak, R.*, et al.* (2009) Dense graphlet statistics of protein interaction and random networks, *Pac Symp Biocomput*, 178-189.

D'Souza, R.M.*, et al.* (2007) Emergence of tempered preferential attachment from optimization., *Proceedings of the National Academy of Sciences, USA*, **104**, 6112-6117.

Deeds, E.J.*, et al.* (2006) A simple physical model for scaling in protein–protein interaction networks, *Proceedings of the National Academy of Sciences, USA*, **103**.

Dorogovtsev, S.N. and Mendes, J.F.F. (2002) Evolution of networks, *Advances in Physics*, **51**, 1079-1187.

Erdös, P. and Rényi, A. (1959) On random graphs, I, *Publicationes Mathematicae (Debrecen)*, **6**, 290-297.

Gandhi, T.*, et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets, *Nature Genetics*, **38**, 285-293.

Gasch, A.P.*, et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Mol Biol Cell*, **11**, 4241-4257.

Gavin, A.*, et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415**, 141-147.

Guldener, U.*, et al.* (2006) MPact: the MIPS protein interaction resource on yeast, *Nucleic Acids Research*, **34**, D436.

Han, J.*, et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network, *Nature*, **430**, 88-93.

Hormozdiari, F.*, et al.* (2007) Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution., *PLoS computational biology*, **3**, e118.

Huang da, W.*, et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*, **4**, 44-57.

Ito, T.*, et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569.

Jenssen, T.K.*, et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression, *Nat Genet*, **28**, 21-28.

Khanin, R. and Wit, E. (2006) How scale-free are biological networks, *Journal of Computational Biology*, **13**, 810-818.

Krogan, N.*, et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature*, **440**, 637-643.

Kultz, D. (2005) Molecular and evolutionary basis of the cellular stress response, *Annu Rev Physiol*, **67**, 225-257.

Kumar, R.*, et al.* (2000) Stochastic models for the Web graph, *Proceedings of the 42st Annual IEEE Symposium on the Foundations of Computer Science*, 57-65.

Li, H.*, et al.* (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale, *Bioinformatics*, **22**, 989-996.

Li, J.*, et al.* (2007) Maximal Biclique Subgraphs and Closed Pattern Pairs of the Adjacency Matrix: A One-to-One Correspondence and Mining Algorithms, *IEEE Transactions on Knowledge and Data Engineering*, **19**, 1625-1637.

Lima-Mendez, G. and van Helden, J. (2009) The powerful law of the power law and other myths in network biology, *Molecular Biosystems*, **5**, 1482-1493.

Lovász, L. and Szegedy, B. (2006) Limits of dense graph sequences, *J. Comb. Theory Ser. B*, **96**, 933-957.

Middendorf, M.*, et al.* (2004) Discriminative topological features reveal biological network mechanisms., *BMC bioinformatics*, **5**, 181.

Mitzenmacher, M. (2004) A Brief History of Generative Models for Power Law and Lognormal Distributions, *Internet Mathematics*, **1**, 226-251.

Newman, M.E.J. (2003) The structure and function of complex networks, *SIAM Review*, **45**, 167-256.

Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag, Heidelberg, Germany.

Pastor-Satorras, R.*, et al.* (2003) Evolving protein interaction networks through gene duplication., *Journal of theoretical biology*, **222**, 199-210.

Pržulj, N. (2004) Modeling interactome: scale-free or geometric, *Bioinformatics*, **20**, 3508-3515.

Rain, J.C.*, et al.* (2001) The protein-protein interaction map of Helicobacter pylori, *Nature*, **409**, 211-215.

Reguly, T.*, et al.* (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae, *Journal of Biology*, **5**, 11.

Reiko, T.*, et al.* (2005) Some protein interaction data do not exhibit power law statistics, *FEBS letters*, **579**, 5140-5144.

Salwinski, L.*, et al.* (2004) The Database of Interacting Proteins: 2004 update, *Nucleic Acids Research*, **32**, D449-D451.

Schulze, W.X. and Mann, M. (2004) A novel proteomic screen for peptide-protein interactions, *J Biol Chem*, **279**, 10756-10764.

Sharan, R.*, et al.* (2007) Network-based prediction of protein function, *Mol Syst Biol*, **3**, 88.

Stark, C.*, et al.* (2006) BioGRID: a general repository for interaction datasets, *Nucleic Acids Research*, **34**, D535.

Stumpf, M.*, et al.* (2005) Statistical Model Selection Methods Applied to Biological Networks. In Priami, C., Merelli, E., Gonzalez, P. and Omicini, A. (eds), *Transactions on Computational Systems Biology 3*. Springer, 65-77.

Stumpf, M.*, et al.* (2008) Estimating the size of the human interactome, *Proceedings of the National Academy of Sciences*, **105**, 6959.

Stumpf, M.P. (2005) Subnets of scale-free networks are not scale free: the sampling properties of networks, *Proceedings of the National Academy of Sciences, USA*, **102**, 4221-4224.

Uetz, P.*, et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, **403**, 623-627.

Ulitsky, I. and Shamir, R. (2007) Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks, *Mol Syst Biol*, **3**, 104.

Vermeulen, M.*, et al.* (2008) High confidence determination of specific protein-protein interactions using quantitative mass spectrometry, *Curr Opin Biotechnol*, **19**, 331-337.

von Mering, C.*, et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Xenarios, I. and Eisenberg, D. (2001) Protein interaction databases, *Current Opinion in Biotechnology*, **12**, 334-339.

Yu, H.*, et al.* (2008) High-quality binary protein interaction map of the yeast interactome network, *Science*, **322**, 104.

Appendix (Supplemental Figure S1)

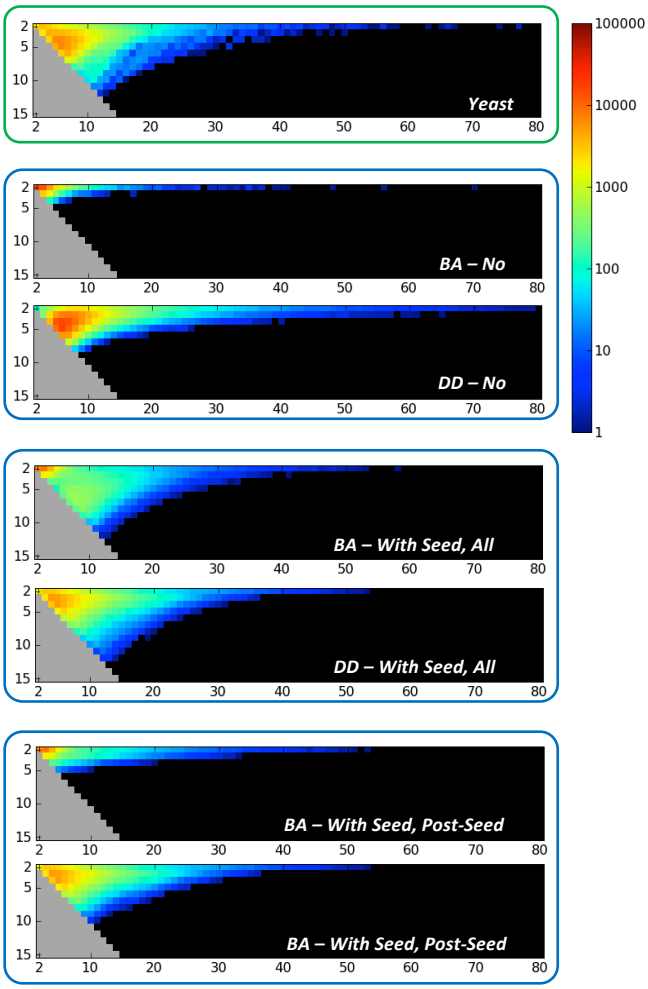

Fig. S1: *Biclique distributions.* A biclique is uniquely defined by the sizes of its two disjoint node sets. For each possible pair of sizes n,m ($2 \le n \le m$), the number maximal bicliques of size (n,m) exist in the graph is shown at a log-scale. The gray area corresponds to $n>m$, which is null by definition. This histogram is shown for: A) The real Protein-Protein Interaction graph of *Saccharomyces Cerevisi*ae; B) The best fit of the *Barabási-Albert (BA)* model, with a degenerate seed graph; C) The best fit of the *Duplication-Divergence* (*DD*) model, with a degenerate seed graph; D) The best fit of the *BA* model, with a full seed graph; E) The best fit of the *DD* model, with a full seed graph; F-G) The same as D-E, except that bicliques that are fully contained in the seed graph are omitted from the count. The Figure is a full histogram for Fig. 3 up to (15,80).