BIOINFORMATICS

Vol. 27 ISMB 2011, pages i142–i148
doi:10.1093/bioinformatics/btr201

Generative probabilistic models for protein–protein interaction networks—the biclique perspective

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

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

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. As PPI networks are rich and complex, 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 observed property of the network 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.

Contact: regevs01@cs.huji.ac.il; michall@cc.huji.ac.il; nati@cs.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* 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) the collection of data which depict large networks in high-resolution detail, and (ii) 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 recently become 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* and the *Caenorhabditis elegans* (Stumpf *et al.*, 2008).

PPIs are usually represented by an undirected 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-Hyrbid (Y2H) Systems (Ito *et al.*, 2001; Krogan *et al.*, 2006) and co-immunoprecipitation (Gavin *et al.*, 2002; von Mering *et al.*, 2002). HTP 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 *Escherichia coli* and *Helicobacter pylori* were investigated as well, but they are far from being complete (Rain *et al.*, 2001). The PPI networks of 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 of 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), 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 biases (Han *et al.*, 2004).

It is not obvious how to infer direct pairwise interactions from HTP techniques that focus on protein complexes (i.e. co-immunoprecipitation). This is especially challenging when complexes of several proteins are considered. To document and

*To whom correspondence should be addressed.

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 Erdös–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 with 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, 2004; Reiko *et al.*, 2005; Stumpf, 2005; Stumpf *et al.*, 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 *BA* model) was introduced (Barabási and Albert, 1999). In this generative model, one starts with a seed graph, and new nodes are iteratively added 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 (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 *DD* model was first suggested in attempting to explain the structure of the world-wide-web graph (Kumar *et al.*, 2000). It was later adopted for the analysis of PPI networks (Bebek *et al.*, 2006; Pastor-Satorras *et al.*, 2003). Here, too, one starts with a seed graph that undergoes a growth process as follows. At every step, an existing node is randomly and uniformly selected and duplicated, i.e. a new node is created 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, each existing node is connected to the new node with a probability r (see Fig. 1B. also see Supplementary Material 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).

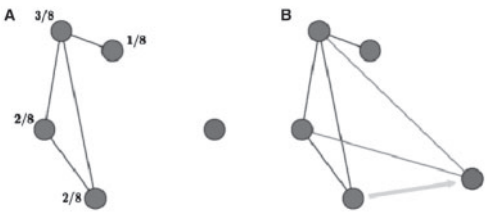

Fig. 1. Model illustrations. (**A**) BA model. A new node (blue) is added at each step. For each of its outgoing edges, the probability of connecting to a given existing node is proportional to the degree of that node. (**B**) DD model. A new node (blue) is added at each step by duplicating an older 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 *DD*- and *BA*-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 $(A,B \subseteq V, A \cap B = \phi)$, 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 when $A \subseteq A_2, B \subseteq B_2$ and $A_2 \times B_2 \subseteq E$, then necessarily $A = A_2, B = B_2$.

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 *BA* 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.*, 2000). 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 PPIs

Protein Interactions were extracted from database of interacting proteins (DIP) (Salwinski *et al.*, 2004) version December 30, 2009. The database contains experimental data covering 5033 proteins of the *S.cerevisiae* (baker's yeast) and 22 118 interaction edges, originating from 16 444 experiments. Note that ~90% of interactions are reported by the HTP experimental techniques of Y2H and co-immunoprecipitation. Additional, low-throughput methods 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 more than 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 $i \in \{1,\ldots,n\}, j \in \{1,\ldots,d\}$. 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: $\|x_i - x_j\| \leq \rho$.

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: $\|x_i - x_j\| \geq R$.

ER model: there are m_0 nodes, and each possible edge between two nodes exists uniformly at random with a probability P. See Figure 4 for a schematic representation of the seed graph models.

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 *ER 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

DD model: we follow the formalization of Bebek *et al.* (2006). To create a graph of n nodes, one begins with a seed graph of size m_0. Then, $n - m_0$ iterations are performed. In every iteration t, a node v is uniformly selected from the previous nodes, $[0,\ldots,t-1]$. For each of the nodes w in $N(v)$, the set of neighbors of v, an edge is created between w and the new node t with probability P, or deleted with probability $1-P$. Then, for each of the nodes u in $[0,\ldots,t-1]$, u and t are connected with probability r/t. Parallel edges are then merged.

BA model: we follow the formalization of Bollobás and Riordan (2005). To create a graph of n nodes, one begins with a seed graph of size m_0. Then, $n - m_0$ iterations are performed. In every iteration t, m edges are added sequentially, stemming from node t. At each edge addition, a probability of $deg(v)/C$ is assigned for all vertices v in $[0,\ldots,t-1]$, 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 includes all edges previously introduced. Parallel edges are merged.

2.5 Comparison between distributions

For two biclique distributions p and q, and sizes $2 \leq n \leq m$, their l_2-distance is defined as:

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

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.

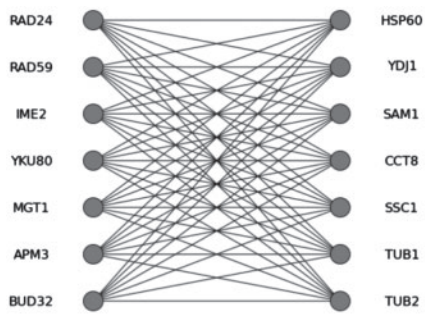

Fig. 2. An example of a biclique from the yeast PPI graph. The biclique of size 7, 7 is shown. Proteins are indicated by their gene names. Proteins on the left set are related to DNA repair and to events that take place at mitosis and meiosis. Proteins on the right set participate in folding and chaperone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

2.7 Enumeration of bicliques and implementation

Biclique exhaustion and enumeration was performed with the Linear time Closed itemset Miner, Maximal Bicliques (LCM-MBC) algorithm and the Frequent Pattern, Maximal Bicliques (FP-MBC) software (Li *et al.*, 2007). The algorithm is based on gradually expanding maximal bicliques, with efficient backtracking and pruning. 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 BioAssociation protocol (Jenssen *et al.*, 2001). Sets of proteins were tested as a group in an enrichment protocol for gene ontology (GO) annotation enrichment using DAVID (Huang da *et al.*, 2009). Only statistically significant annotations are reported ($P < 0.05$).

3 RESULTS

3.1 Bicliques in the PPI networks are of biological relevance

We analyzed the PPI network of *S.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, Figure 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 checked for GO annotation keyword enrichment in each set (Ashburner *et al.*, 2000)

Statistical analysis of the left set, based on literature gene associations (Jenssen *et al.*, 2001), 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, proteins related to stressdependent folding, peptide transport to mitochondria and complex formation are included. All these proteins are highly

abundant and they 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)

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 and telomere maintenance) and sets of genes involved in protein maintenance (e.g. unfolded protein binding, stress response, protein refolding and chaperones) appear in two sides in the analyzed bicliques (Supplementary Table S1). Such crosstalk also prevails in the analysis of the gene expression programs of *S.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 *S.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 ($2 \leq n \leq m$), 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 and B and Supplementary 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 emanating from the seed graph, we ran both models with a minimal seed consisting of a complete graph on three 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 PPIs 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 l_2 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 $49\,691 \pm 40$ edges (mean $\pm$ SD). 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 (2157)]. The l_2 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 $43\,251 \pm 13\,379$ 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 (2265)]. The l_2 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.

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 $\sim$100 nodes. Three different methods were used for the generation of seed graphs: (i) 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 connected vertices are positively correlated (Fig. 4A). (ii) 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 $47\,286 \pm 9092$ edges (mean $\pm$ SD). The l_2 distance to the real graph is 12.35 ± 3.08

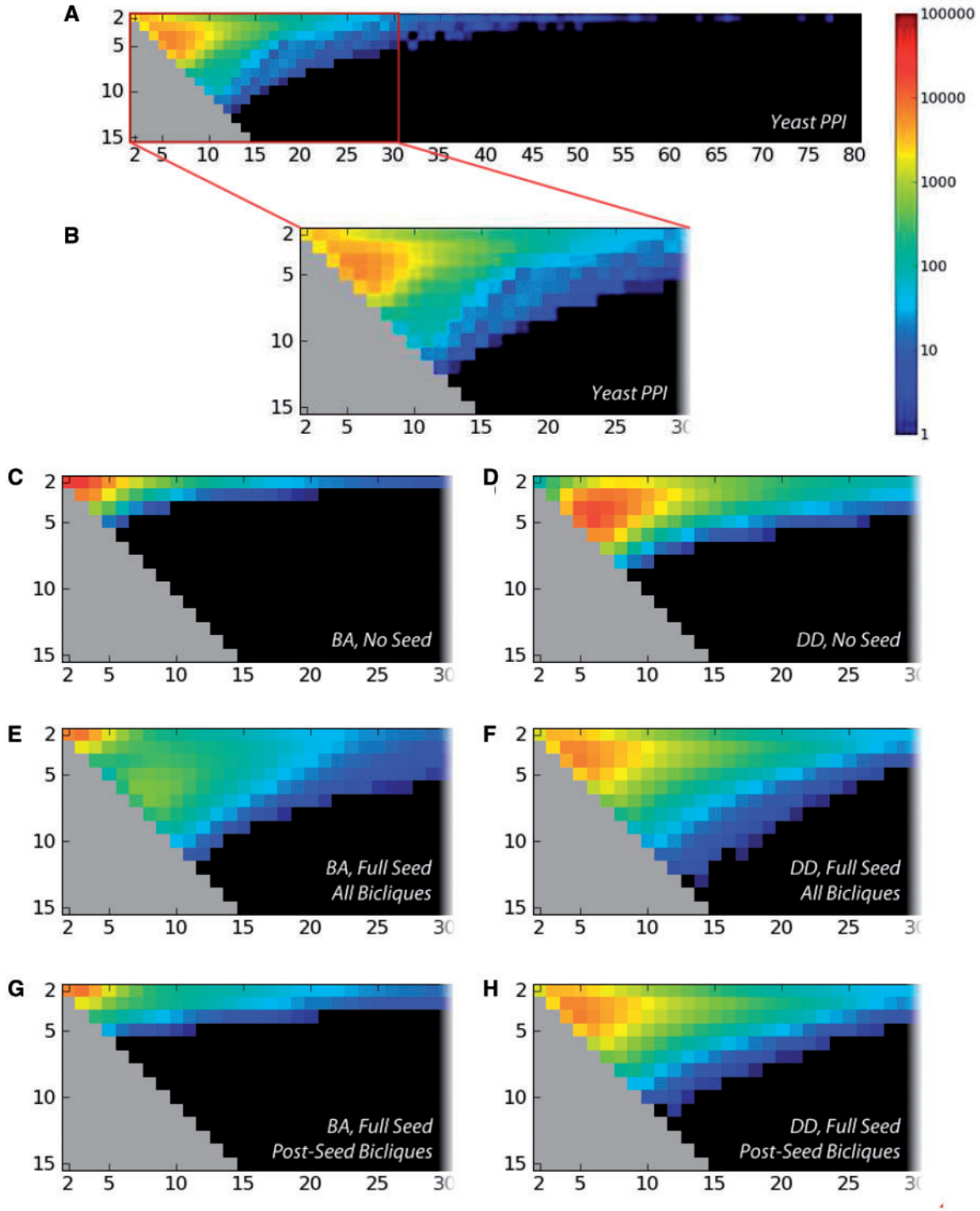

Fig. 3. Biclique distributions. A biclique is uniquely defined by the sizes of its two disjoint node sets. For each possible pair of sizes n $m(2 \leq n \leq m)$, the number of maximal bicliques of size (n,m) that exist in the graph is shown at a log-scale. The gray area corresponds to $n > m$, which is null by definition. This histogram is shown for: (**A**, **B**) The real PPI graph of *Scerevisiae*; (**C**) the best fit of the BA model, with a degenerate seed graph; (**D**) the best fit of the 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 Supplementary Fig. S1).

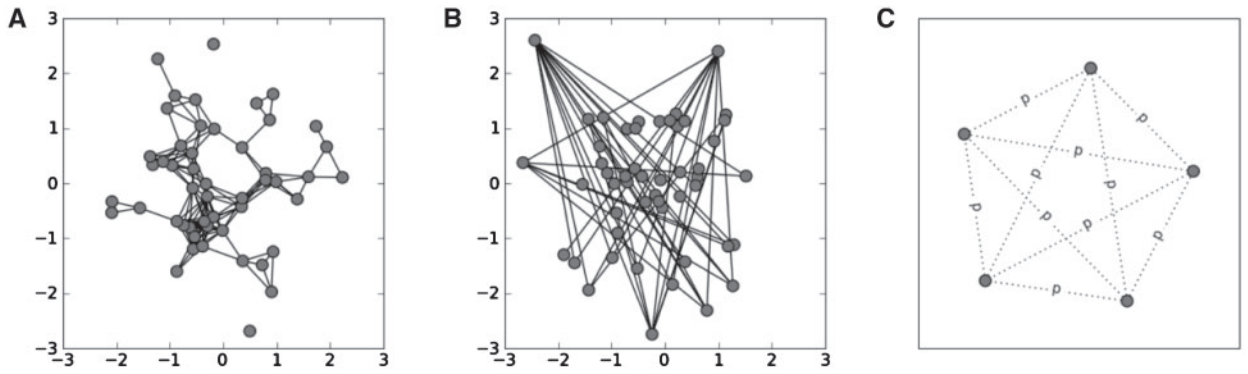

Fig. 4. Seed model illustrations. (**A**) *Random geometric model*. Random points are sampled from a standardized normal distribution on R^d (here $d=2$). Each node corresponds to a point, and two nodes are connected in the graph if the corresponding two points are at distance smaller than ρ. (**B**)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

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)	l_2 dist., best BA model (SD)	l_2 dist., 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; l_2 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. l_2 dist: l_2 distance to yeast PPI; BC: bicliques; SD: standard deviation.

(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 in the 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, giving 19463.76 ± 1307.77 edges, close to the number of edges in the real PPI network. Here, the l_2 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.

4 DISCUSSION

The proper modeling of PPI networks is a major challenge from two perspectives: 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 originate 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 false positive rate (Bader and Hogue, 2002; Uetz *et al.*, 2000). However, with advances in experimental design (Vermeulen *et al.*, 2008; Yu *et al.*, 2008), consistently contaminating proteins are more easily 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 >50% of the data, mainly extracted from co-immunoprecipitation experiments, and from X-ray and stable isotope labeling by amino acids in cell culture (SILAC)-based tandem affinity purification (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 *S.cerevisiae*. It transpires that the graphs generated by the *DD* model are in much better agreement with the actual network than those generated by the *BA* model. Clearly, both models can be expanded and refined in various ways, and we have restricted ourselves to the basic versions of either model. In addition, other models may also be suggested. Indeed, other closely related models have also been shown to give rise to many large bicliques (Kumar *et al.*, 2000). Still, we believe that our findings are rather indicative of the general picture.

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 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.

Funding: Prospects (EU, Framework VII, partially); Israel Science Foundation (ISF 592/07); Fellowship from the SCCB, the Sudarsky Center for Computational Biology (to R.S.).

Conflict of Interest: none declared.

REFERENCES

Albert,R. and Barabási,A.-L. (2002) Statistical mechanics of complex networks. *Rev. Modern Phys.*, **74**, 47–97.

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. *Nat. Genet.*, **25**, 25–29.

Bader,G. and Hogue,C. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **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. *Theor. Comput. Sci.*, **369**, 239–249.

Bollobás,B. and Riordan,O.M. (2005) Mathematical results on scale-free random graphs. In Bornholdt,S. and Schuster,H.G. (eds) *Handbook of Graphs and Networks: From the Genome to the Internet*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG.

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

Colak,R. *et al.* (2009) Dense graphlet statistics of protein interaction and random networks. In *14th Pac. Symp. Biocomput.*, **14**, 178–189.

D'Souza,R.M. *et al.* (2007) Emergence of tempered preferential attachment from optimization. *Proc. Natl Acad. Sci. USA*, **104**, 6112–6117.

Deeds,E.J. *et al.* (2006) A simple physical model for scaling in protein–protein interaction networks. *Proc. Natl Acad. Sci. USA*, **103**, 311–316.

Dorogovtsev,S.N. and Mendes,J.F.F. (2002) Evolution of networks. *Adv. Phys.*, **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. *Nat. Genet.*, **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 Res.*, **34**, D436–D441.

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 Comput. Biol.*, **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. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

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. *J. Comput. Biol.*, **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. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pp. 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 Trans. Knowl. Data Eng.*, **19**, 1625–1637.

Lima-Mendez,G. and van Helden,J. (2009) The powerful law of the power law and other myths in network biology. *Mol. Biosyst.*, **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 Math.*, **1**, 226–251.

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

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, Heidelberg, Germany.

Pastor-Satorras,R. *et al.* (2003) Evolving protein interaction networks through gene duplication. *J. Theor. Biol.*, **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. *J. Biol.*, **5**, 11.

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

Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **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 Res.*, **34**, D535–D539.

Stumpf,M., *et al.* (2005) Statistical model selection methods applied to biological networks. In Priami,C. *et al.* (eds) *Transactions on Computational Systems Biology 3*. Springer, Weinheim, FRG, pp. 65–77.

Stumpf,M. *et al.* (2008) Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA*, **105**, 6959–6964.

Stumpf,M.P. (2005) Subnets of scale-free networks are not scale free: the sampling properties of networks. *Proc. Natl Acad. Sci. 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. *Curr. Opin. Biotechnol.*, **12**, 334–339.

Yu,H. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.

Yu,H. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network, *Science*, **322**, 104–110.