

Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks

Shawn M. Gomez,* Shaw-Hwa Lo[†] and Andrey Rzhetsky*[‡]

*Columbia Genome Center, [†]Department of Statistics and [‡]Department of Medical Informatics, Columbia University, New York, New York 10032

Manuscript received January 23, 2001
Accepted for publication August 6, 2001

ABSTRACT

Regulatory networks provide control over complex cell behavior in all kingdoms of life. Here we describe a statistical model, based on representing proteins as collections of domains or motifs, which predicts unknown molecular interactions within these biological networks. Using known protein-protein interactions of *Saccharomyces cerevisiae* as training data, we were able to predict the links within this network with only 7% false-negative and 10% false-positive error rates. We also use Markov chain Monte Carlo simulation for the prediction of networks with maximum probability under our model. This model can be applied across species, where interaction data from one (or several) species can be used to infer interactions in another. In addition, the model is extensible and can be analogously applied to other molecular data (e.g., DNA sequences).

RECENT achievements in genome sequencing, coupled with advances in cellular biology, have raised hopes for an imminent leap forward in our understanding of the regulatory machinery of life. However, we have yet to make the transition from a linear one-dimensional sequence of genes to an integrated multidimensional model of metabolic and regulatory networks. Despite their importance, relatively little is understood, with a major complication being the general lack of data on the mechanism, rate, and even existence, of interactions between known genes and proteins. In fact, only recently have sufficient data sets become available to provide support for the analysis of such large-scale networks (UETZ *et al.* 2000; XENARIOS *et al.* 2000).

Molecular-interaction networks feature proteins, nucleic acids, and small molecules as primary players. Since genes are passive carriers of information, and because there are relatively few enzymatic or structural RNA molecules, the majority of important biological functions are carried out by proteins. Being linear sequences of amino acids at the level of primary structure, at the functional level, proteins can be broken down into segments that correspond to functional domains or conserved motifs. Like amino acids, these domains are discrete “letters,” combinations of which give rise to the diversity of protein form and function.

In this work, we assume that the existence of a network connection between proteins, which may or may not involve a physical interaction between them, is a func-

tion of the domain composition of each. For convenience of description, we treat nonprotein network nodes as single-domain proteins. If we move along a network pathway, a domain of an upstream protein may favor interaction with a domain of a downstream protein. In addition to a physical connection, the term “interaction” in our model can represent more general relationships between domains, e.g., information flow. Furthermore, we assume that once a given pair has proven effective, nature will tend to reuse it in other networks within the same organism, as well as in other organisms. Thus the model we describe here is based on quantifying, from data taken from known networks, the frequency with which a domain in one protein is observed immediately upstream or downstream of domains in another protein. We then use this information to infer the probability of unknown interactions. Below, we describe our model and how its parameters are estimated, verify its validity with cross-validation, and show sample applications to real biological networks.

MODEL DESCRIPTION

The model we now describe assigns probabilities to all possible networks formed from a fixed number of vertices. Note that, rather than trying to reproduce the actual genesis of regulatory networks in evolution, our model has the more modest purpose of providing each network with a probability value in such a way that networks having more features typical of real networks have higher probabilities.

We represent a network as an oriented graph, $G = \langle V, E \rangle$, where the vertices, V , correspond to proteins, and the edges, E , correspond to interactions between

Corresponding author: Andrey Rzhetsky, Columbia Genome Center and Department of Medical Informatics, Columbia University, Russ Berrie Pavilion, Rm. 121H, 1150 St. Nicholas Ave., Unit 109, New York, NY 10032. E-mail: ar345@columbia.edu

proteins. Each vertex of the network is composed of one or more domains or motifs, which are identified through comparison with existing databases of protein domains (*e.g.*, Pfam; BATEMAN *et al.* 2000). We use the frequency of separate occurrence of domains d_m and d_n in two connected vertices of a known network to infer probabilities of “attraction” $p(d_m, d_n)$ (*i.e.*, that an oriented edge will be formed) between these domains. As described in detail later, these probabilities are used to determine the probability of individual protein-protein interactions.

This model has two independent stochastic steps, and the probability of an individual network emerges as a product of the probabilities associated with these two stages. In the first stage, every pair of proteins i and j may be connected to each other with an “attraction” probability p_{ij} (we explain how to compute this probability later) or not connected with probability $(1 - p_{ij})$. We can imagine this process as being performed by a machine that, for every pair of vertices, tosses imaginary biased coins, each coin specific to a particular pair of proteins. If it is heads, an edge between the two vertices is formed; if it is tails, it does not form. The coin is biased by prior information about the domains in each of the vertices, leading to some edges having a probability >0.5 (attraction) and some edges <0.5 (repulsion). For a network with $|V|$ vertices, there are $2^{|V||V|}$ possible networks with oriented edges (*e.g.*, a network we describe later consists of 11 vertices and thus has 10^{36} possible configurations). We then define the probability of a single network with the particular edge set E as

$$P(E) = \prod_{e_{ij} \in E} p_{ij} \prod_{e_{kl} \notin E} [1 - p_{kl}]. \quad (1)$$

Using this process, we are able to assign a probability to any configuration of edges between a set of vertices. Networks containing many high-probability edges will have higher probabilities $P(E)$.

In the second stage, networks are sorted into a finite number of bins, each corresponding to a particular “network topology.” In this case, “network topology” is defined as the particular distribution of edges coming into and out of each vertex of the network. Note that for a given number of vertices, it is possible to have a large number of edge configurations that are characterized by the same topology, so each bin represents a collection of networks with identical topologies. The number of incoming edges, or *indegree*, of a vertex in an oriented graph is the number of oriented edges that end at this vertex. Similarly, the *outdegree* of a vertex is the number of oriented edges that start at this vertex. For a pair of proteins connected with a single oriented edge, the upstream protein has a single outgoing edge while the downstream protein possesses a single incoming edge. For each network we compute the number of vertices that have outdegree zero, n_0^{out} ; one, n_1^{out} ; two, n_2^{out} ; and so on to n_N^{out} (where the subscript indicates the

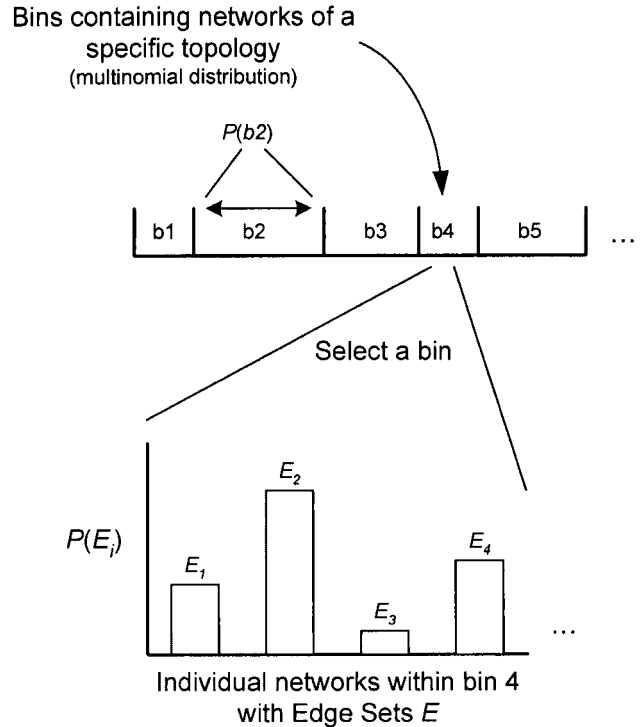


FIGURE 1.—Sampling of a network. The distribution of network topologies is modeled with a multinomial distribution. Individual bins contain a collection of networks with identical topologies. Networks within a given bin have probabilities defined on the basis of their edge composition. See text for further details.

number of edges, and N is the total number of vertices in the graph). Similarly, we compute the numbers of vertices with indegrees 0, 1, 2, . . . N . Now we put into one bin all networks with identical sets $\{n_x^{\text{in}}\}$ and $\{n_y^{\text{out}}\}$. For each bin we define a sampling probability $P(\{n_x^{\text{in}}\}, \{n_y^{\text{out}}\})$ that is computed as the product

$$P(\{n_x^{\text{in}}\}; \{\pi_x^{\text{in}}\}, |V|) \times P(\{n_y^{\text{out}}\}; \{\pi_y^{\text{out}}\}, |V|), \quad (2)$$

where

$$P(\{n_z\}; \{\pi_z\}, |V|) = \frac{|V|!}{n_0! \dots n_N!} \prod_{z=0}^N \pi_z^{n_z}. \quad (3)$$

The probability distributions π_x^{in} and π_y^{out} give the probability of a network having x incoming and y outgoing edges, respectively. These distributions are explained in greater detail later (see *Estimating parameters relevant to the topology of real networks*). Finally, the second step is finished with a random (multinomial) sampling of a bin with probability $P(\{n_x^{\text{in}}\}, \{n_y^{\text{out}}\})$ and then random (uniform) sampling of a network from within that bin (see Figure 1). Note that, at the topological level, we cannot distinguish between the situations where (a) protein 1 has n_i inputs and protein 2 has n_j inputs and (b) protein 1 has n_j inputs and protein 2 has n_i inputs. Rather, this distinction is made at the level of individual edges. Each individual edge has a probability associated with it and

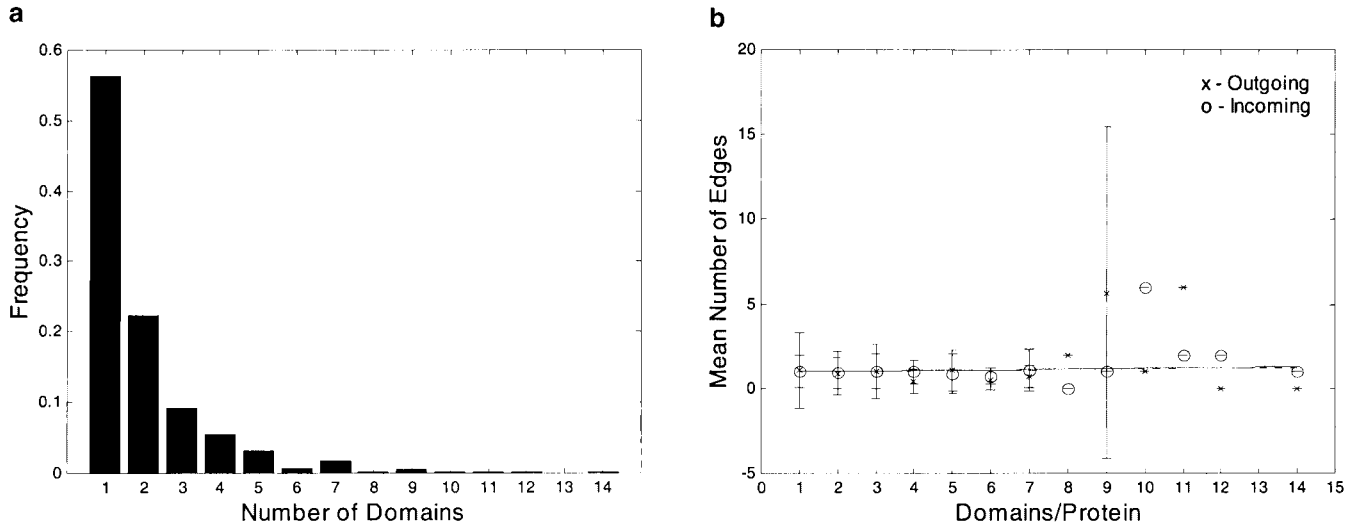


FIGURE 2.—The number of domains per protein does not determine network connectivity. Data are from the yeast network interaction data described in APPLICATION TO REAL NETWORKS. (a) The frequency of proteins with a given number of domains. (b) The number of edges outgoing (x) or incoming (o) to a protein is independent of the number of domains. Error bars represent 1 standard deviation. Regression lines are shown with slopes of 0.024 (intercept = 0.96) and 0.021 (intercept = 0.97) for outgoing and incoming edges, respectively. The large deviation for the number of edges outgoing from proteins with nine domains is due to the fact that only three data points comprise this set. All other points with eight or more domains consist of a single sample (and thus have an undefined variance).

thus so does the complete set of edges that make up a given network $P(E)$. A network topology is automatically defined when given this same set of edges E . The probability of this particular topology, however, is determined separately and may or may not be favorable (biologically realistic). We do not distinguish between (a) and (b) above because *topologically* they are identical. They are not identical, however, at the level of individual edges. Each of these edges in (a) and (b) will have a different probability associated with it, with presumably one version of (a) or (b) being the correct and thus most favorable one.

It is not difficult to verify that the product of the former two stochastic steps would give the probability of sampling any given network:

$$P(E) \times P(\{n_x^{\text{in}}\}; \{\pi_x^{\text{in}}\}, |V|) \times P(\{n_y^{\text{out}}\}; \{\pi_y^{\text{out}}\}, |V|). \quad (4)$$

Networks with both favorable sets of edges E and favorable topologies will have the highest probabilities.

Although the second stage of our network-generating process may appear to the reader as artificial and even unnecessary, it is not. As we elaborate below, real biological networks have a very characteristic topology that distinguishes them from the vast majority of arbitrary random networks. Therefore, in a situation where information about protein domain interactions is far from being complete, a restriction on acceptable network topology is used to improve the prediction ability of the algorithm that we develop here.

Computing probabilities of protein interactions: In this model we consider proteins as “bags of domains,” where each individual pair of domains, d_m and d_n , has a probability

of getting attracted, $p(d_m, d_n)$. If $p(d_m, d_n) > 0.5$, the domains “attract” each other, while for $p(d_m, d_n) < 0.5$, we can say that domains “repel” each other. Considering a pair of multidomain proteins i and j , where v_i and v_j are the sets of domains for each protein (a domain of each type occurs in v_i no more than once, even if the i th protein has multiple domains of the same kind), we assume that the probability of attraction (edge probability) between these proteins is given in terms of domain attraction probabilities as

$$p_{ij} = \sum_{d_m \in v_i} \sum_{d_n \in v_j} \frac{p(d_m, d_n)}{|v_i| |v_j|}. \quad (5)$$

This definition of edge probability is reasonable insofar as the number of edges going into or out of a vertex is not correlated with the number of distinct domains in either of the interacting proteins. We verified that this assumption indeed holds with real data (see Figure 2).

Estimating probabilities of attraction between domains: Facts on interactions between proteins published in the research literature have strikingly different reliabilities. This is in part due to the fact that it is uncommon to publish the negative results of an experiment. As a result, the presence of an interaction between proteins is usually backed by multiple experiments while the absence of interaction may correspond to a failed experiment or just the absence of experiments at all (the only exclusion from this observation is exhaustive two-hybrid screening, where all results, both positive and negative, are reported). Therefore we decided to estimate the probabilities of “attraction” between two domains in such a way that the absence of a connection

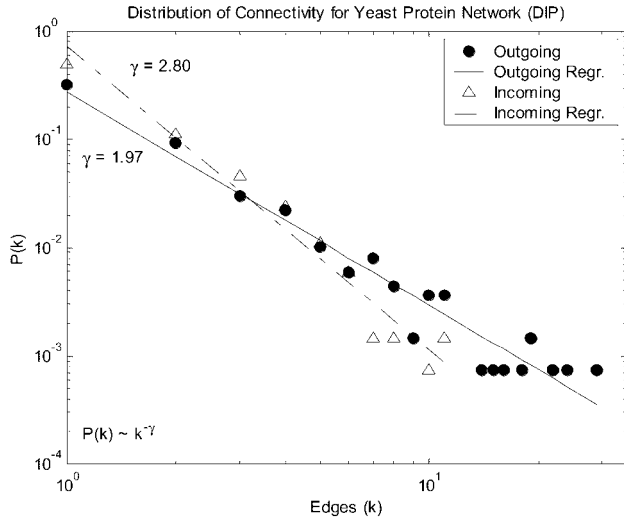


FIGURE 3.—Probability distribution of vertices, with k incoming (open triangles) or outgoing (solid circles) edges. Edge distributions were calculated from the DIP database; they consisted of 1366 vertices with 1479 edges.

is treated as the absence of data, while the counts of known connections are used to estimate the probabilities. That is, for domains d_m and d_n , we compute the attraction probability as

$$p(d_m, d_n) = \frac{1}{2} \left(1 + \frac{k_{mn}}{k_m k_n + \Psi} \right), \quad (6)$$

where Ψ is a positive real-valued pseudocount, k_{mn} is the number of edges in the training set that contain at least one domain d_m at the vertex of edge origin and at least one domain d_n at the vertex of edge destination, k_m is the number of distinct vertices that contain at least one domain d_m , and k_n is the number of distinct vertices that contain at least one domain d_n . For this work we chose $\Psi = 1$, which can be increased (or decreased) if one wants to require the accumulation of greater amounts of data before the prior becomes significantly altered. As an example, if there are two upstream proteins with domain m and two downstream proteins each with domain n , a perfect correspondence between protein domains and the existence of an edge would lead to $k_{mn} = 4$ (all possible edges exist), $k_m = 2$, $k_n = 2$, and $p(d_m, d_n) = 0.9$ (assuming that $\Psi = 1$).

As a result, this formulation assigns probabilities $>50\%$ to edges that have known connections and probabilities *equal to 50%* to edges that have no known connections. In the absence of experimental observations ($k_m = k_n = k_{mn} = 0$), the probability of an edge forming between any two domains is exactly 50%, which in turn leads to a probability of 50% for forming an edge between any two proteins, regardless of the number of domains in each of them. In the absence of data, *all networks* can be assigned a nonzero probability. Note that, while the model allows for it, the current methodology (specifi-

cally, Equation 6) does not generate probabilities of <0.5 for domain-domain interactions. While we could have expanded the probability range from 0 to 1, the compressed scale of 0.5–1 does not affect the results, and a future version of this model, combined with the collection of appropriate data (*e.g.*, negative experimental results, appropriate two-hybrid data, etc.), will use the range of 0–0.5 for modeling “repulsive” effects between domains.

To summarize, this model assigns a probability to every possible network with $|V|$ vertices. This probability is based on both local and global network properties. At the local level, the probability of a vertex having an interaction with another vertex is dependent on the domain composition of each. If, as previously determined by training data, the set of domains in one protein is likely to be attracted to that of another protein, the probability of an edge existing between the two vertices increases to a value >0.5 . If no information is available about the likelihood of interaction for the set of domains contained in both upstream and downstream vertices, the probability of an edge forming between the two is taken to be 0.5. At a global scale, the probability of the network (based solely on local properties) is modified on the basis of how well it represents real biological networks. Networks with topologies (distribution of incoming and outgoing edges per vertex) that are more biologically realistic are given higher likelihoods. The probability of any given network is the product of both the local and global probabilities.

Estimating parameters relevant to the topology of real networks: When we estimated the parameters $\{\pi_x^{\text{in}}\}$ and $\{\pi_y^{\text{out}}\}$ from the database of interacting proteins (DIP) dataset, we found that the estimated values for both sets followed a power-law distribution (Figure 3). This means that, in logarithmic coordinates, the relationship between the number of connections per vertex and the proportion of vertices with that many connections is *linear* for both incoming and outgoing edges. This power-law distribution is a property of scale-free systems. These systems have properties or behaviors that are invariant across changes in scale. A demonstration of this phenomenon is shown in Figure 4, where it is not possible to determine the scale of the object (in this case, cauliflower) without a reference object. This property has also been seen in networks (BARABASI and ALBERT 1999; ALBERT *et al.* 2000), including social and nonbiological networks, where the probability of a vertex having k incoming or outgoing edges is given by

$$\pi_k = ck^{-\gamma}, \quad (7)$$

but with the values of γ and c being different for incoming and outgoing edges. For the outgoing edges of our network, a linear fit at logarithmic scale gave estimates of $c = 0.30$ and $\gamma = 1.97$, whereas incoming edges were distributed with $c = 0.56$ and $\gamma = 2.80$. For this work,

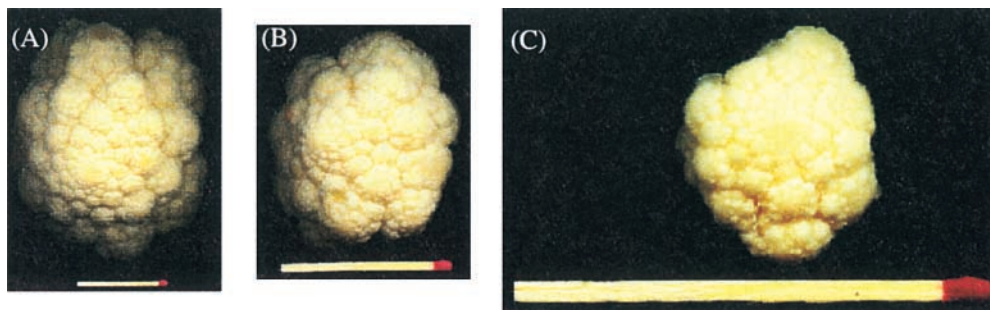


FIGURE 4.—Scale-free (self-similarity) properties of a common cauliflower plant: it is virtually impossible to determine whether one is looking at a photograph of a complete vegetable or its part, unless an additional scale-dependent object (a match) is added. (A) Complete vegetable; (B) a small segment of the same vegetable; (C) small part of the segment

shown in B. The same match was used in all three photographs to provide a sense of scale for an otherwise scale-free structure. The idea originated from the book written by PEITGEN *et al.* (1992). We purchased the vegetable in Sloan's supermarket in Manhattan's Upper West Side; the photographs were obtained by direct scanning of the objects with a Compaq S⁴ 100 scanner.

we used the power law for π_k only for nonzero k ; the values of π_0^{in} and π_0^{out} were estimated as

$$\pi_0^{\text{in}} = 1 - \sum_{k=1}^{\infty} \pi_k^{\text{in}}, \quad \pi_0^{\text{out}} = 1 - \sum_{k=1}^{\infty} \pi_k^{\text{out}}. \quad (8)$$

Examples of the influence of network topology on the likelihood of a given network are shown in Figure 5.

APPLICATION TO REAL NETWORKS

To test the efficacy of this model, we needed networks with large numbers of known protein-protein interactions. A complication in this process arises in that, even if a large number of interactions are known, not all of them have a defined domain composition. For this work, we used *Saccharomyces cerevisiae* protein-protein interactions taken from the DIP (<http://dip.doe-mbi.ucla.edu/>; XENARIOS *et al.* 2000). We determined the domains involved in each interaction by analyzing protein sequences with hmmpfam (BATEMAN *et al.* 2000), a publicly available software tool that referenced 2015 domains at the time of this analysis. We analyzed a total of 642 protein-protein interactions (all with at least 1 domain) and then used them to determine the domain-domain interaction probabilities. Data (in this case a list of undirected protein-protein interactions) used for studying the effect of vertex removal on network edge distributions were taken from the Fields Lab home page (<http://depts.washington.edu/sfields/>).

The yeast protein network is scale free: We know that the observed power-law behavior for the distribution of edge types within the network implies a scale-free system. To provide another means of verification, we determined the value γ for a large network (1823 vertices). We then ran a bootstrap procedure for 200 iterations, where 30 vertices were randomly removed from the network and the value of γ and 95% confidence intervals were determined for each. After this was completed, 60 vertices were removed and the process was repeated. This was repeated until the final 200 iterations with 113 total vertices in the network. The effect of vertex removal on γ is shown in Figure 6 (mean of γ and 95% confidence intervals displayed) and shows that this

network is remarkably scale invariant. This implies that knowledge of the topology of a small part of a network should provide a reliable means of estimating the complete network's topology.

Cross-validation: We used cross-validation to determine the effectiveness of the model in predicting the overall network configuration. Cross-validation is a general technique for evaluating the efficiency (and hence, validity) of statistical algorithms. It typically involves dividing a dataset into two disjoint subsets, one of which is used for training, with the other being used for model validation. We used the jackknife version of cross-validation, where the training set consists of the complete network minus a single specified edge. We then compared the likelihood of the complete graph (the model validation set of data) to that of the complete graph minus one edge. If the likelihood of the full network was greater than the likelihood of the reduced network, we considered the edge to be positively predicted. This step was performed iteratively until all edges were considered. Analysis of the test network indicated that the model predicted 93% of the known 642 edges used in the test; the remaining 7% represent *false negatives*. We similarly estimated the rate of *false positives* as $\sim 10\%$ by starting with the full known network and attempting to add a single edge between unconnected vertices. Note that this measure of false positives assumes that all edges not included in the true network should not exist. Currently, we cannot determine which, if any, of the false-positive edges correspond to true—but currently unknown—connections. While we should be able to achieve even greater accuracy by including more data in the training set, these results suggest that the model is valid and is capable of making reasonably accurate predictions.

Markov chain Monte Carlo: For nearly all species, many interactions within a biological pathway are currently unknown. As our model allows us to compute the probability for any possible arrangement of edges that connect a set of vertices, we implemented a Markov chain Monte Carlo (MCMC) simulation approach (HASTINGS 1970; GILKS *et al.* 1996), which allowed us to compute posterior probabilities for all edges while effec-

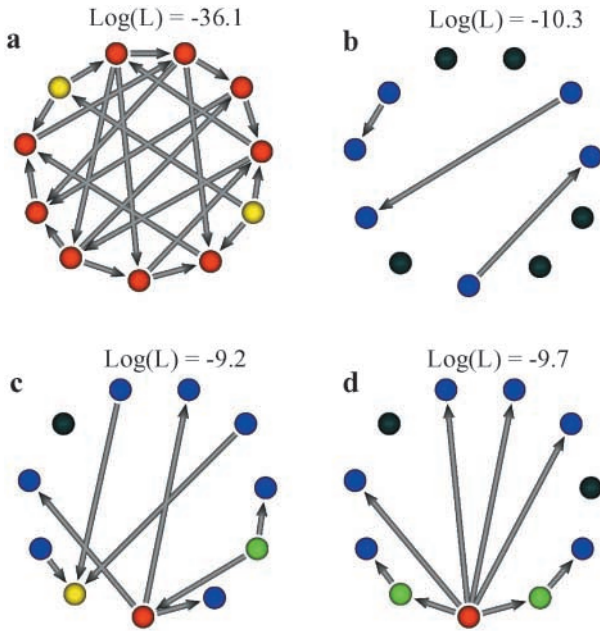


FIGURE 5.—Log of network likelihood based only on network topology. a and b represent, respectively, a highly (overly) connected and minimally connected network. c depicts a more realistic (optimum) configuration. d shows a network with the same number of edges as c but with a less-favorable arrangement of incoming and outgoing edges. Networks with less-negative log scores are more likely. Networks were created with CUTENET (KOIKE and RZHETSKY 2000).

tively sampling from the astronomically large number of possible networks.

We implemented a reversible-jump methodology (GREEN 1995) typical for Bayesian model selection, treating different networks as alternative statistical models. We chose a uniform prior distribution over all networks, because, without additional information, we have no reason to prefer one network over another. Starting with an arbitrary network, the algorithm either adds or removes, with equal probability, a defined number of edges. Edges to be added or deleted are respectively sampled from the pool of edges that are included or excluded from the current network, with the probability of selecting any given edge dependent on only the number of edges from which to choose. Adding or removing edges in this manner, the system jumps from network X to a new network Y . The proposed new state Y is sampled from the *proposal distribution*, $q(b|a)$. The new network Y is then accepted with probability

$$\alpha(X, Y) = \min \left\{ 1, \frac{L(Y)q(X|Y)}{L(X)q(Y|X)} \right\}, \quad (9)$$

where $L(\cdot)$ is the likelihood of the given network. If the proposed new state is accepted, then network Y becomes the current network; otherwise, the old network X remains as the current model. The stochastic process moves through the space of possible networks, on average keeping each edge in an on or off state in proportion

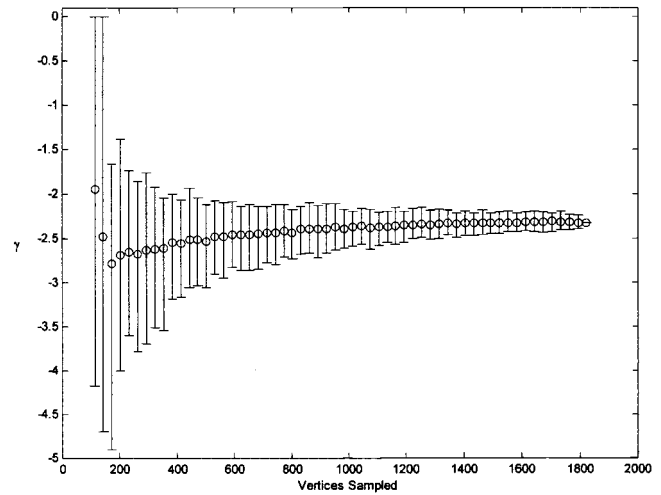


FIGURE 6.—Distribution of edges per vertex is scale invariant. The mean of γ (~ -2.3), along with 95% confidence intervals, is shown. See text for further details.

to the posterior probabilities of this edge being present or absent in the correct network.

As a small-scale example, we selected a group of 11 yeast proteins known to interact with at least one other member of the group and attempted to predict these edges (Figure 7). The probabilities of a given edge, based on domain-domain interactions alone, are shown in Figure 7a. Note that all edges except (7, 1) (x -axis, y -axis) are found in the original data. The posterior probability estimated through simulation is shown in Figure 7b, and all known edges except (10, 1) are predicted reliably. This result is not merely a sampled version of Figure 7a; rather, it incorporates the constraints imposed by the edge distributions on the topology of the network. Thus edges (7, 1) and (10, 1) are not supported with high confidence due to their low domain-domain interaction probabilities and to the influence of the edge distributions. The effect of topology constraints can also be seen where regions of low probability [*e.g.*, the vicinity of (4, 8)] are associated with proteins that already have a high-probability edge; addition of a second edge is unlikely. The nonsymmetrical pattern is due to differences between the outgoing and incoming edge distributions. Although they are easily differentiated from unlikely edges, all likely edges have relatively low posterior probabilities.

For very small systems, a significant amount of information can be gained simply from looking at the edge probabilities between a given set of vertices, with very little additional information coming from topology information. However, use of the MCMC method described here should be particularly valuable for the prediction of large networks, where large amounts of protein interaction data with complicated domain architectures (such as those of higher organisms) and a computationally intensive number of network topologies are the norm.

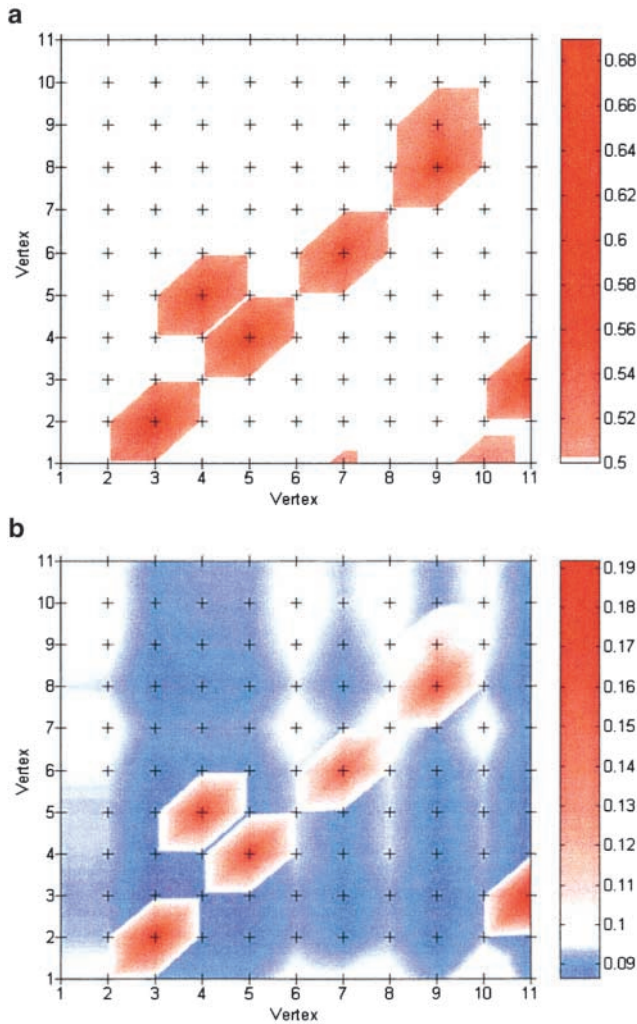


FIGURE 7.—MCMC simulation of a small network. Vertices are as follows: 1, transcription factor BAS1 (gi|101447); 2, oxoglutarate dehydrogenase precursor (gi|1070439); 3, dihydrolipoamide S-succinyltransferase precursor (gi|2144399); 4, cell division control protein CDC43 (gi|2144611); 5, protein farnesyltransferase chain RAM2 (gi|266880); 6, pre-mRNA splicing factor PRP21 (gi|280467); 7, hypothetical protein YBL067C (gi|626480); 8, omnipotent suppressor protein SUP45 (gi|626763); 9, suppressor 2 protein (gi|72877); 10, transcription factor GRF10 (gi|82888); 11, dihydrolipoamide dehydrogenase precursor (gi|82983). (a) Edge probabilities for the network based on domain-domain attraction probabilities alone. (b) Posterior probabilities of all edges of the network after 10^9 iterations of MCMC simulation. Red and blue colors represent probability above and below, respectively, the mean of all edge probabilities (in white). Note that only values at vertex intersections (+) have meaning; areas in between are interpolated and merely help to show gradients. Edges known to exist from the original data are [(3, 2), (4, 5), (5, 4), (7, 6), (9, 8), (9, 9), (10, 1), and (11, 3)]. The percentage of rejected edges in MCMC computation was 82%.

As a further example of the application of domain-domain interaction information, we selected 10 proteins known to function in the human apoptosis pathway from the KEGG database (GOTO *et al.* 1997). As is obvious from Figure 8, few edges were supported by yeast

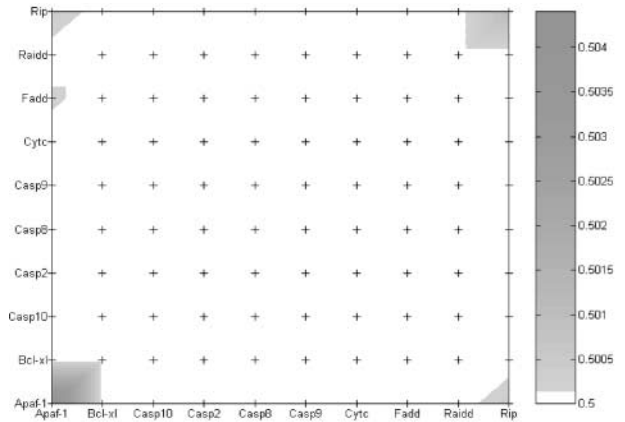


FIGURE 8.—Prediction of interactions among 10 proteins that are involved in the human apoptosis pathway. Only probabilities based on domain-domain interactions alone are shown.

training data; however, the most strongly predicted interaction was of Apaf-1 interacting with itself. A search of the signal-transduction literature revealed that Apaf-1 does, in fact, self-associate (HU *et al.* 1998; BENEDICT *et al.* 2000). We were not hitherto aware of this association, and it was not described within KEGG. We believe these results to be quite encouraging. While we were not able to predict the known network, this example is remarkable given the small amount of domain-domain interaction data available for training; it demonstrates the potential application of this method to predicting interactions across species. Accumulation of interaction from more complicated organisms should greatly enhance these predictions.

DISCUSSION

Based on the simple concepts of domain composition and network topology, this model allows us to characterize and predict both known and unknown protein interactions within a given species and potentially across species. Markov chain Monte Carlo techniques described earlier provide a computationally feasible way to calculate the posterior probability of a network, given data as

$$P(\text{network}_i | \text{data}) = \frac{P(\text{data} | \text{network}_i)P(\text{network}_i)}{\sum_{\text{all networks}} P(\text{data} | \text{network}_j)P(\text{network}_j)}$$

While we assumed a uniform prior distribution over all possible networks, the model does not require this. Furthermore, we can add additional information (in the form of priors) into the calculation as it becomes available.

In the study of regulatory pathways, this model could significantly reduce the number of required experiments by identifying a few most likely hypotheses. Such experimental analysis is itself an empirical way of validating the model, and we can likewise design experiments for this validation. Improvements could include additional interaction data and the introduction of more

domains for assignment to protein segments. We are currently enhancing the model by allowing the introduction of repulsion effects, which are implemented by allowing probabilities of <0.5 for domain-domain interactions. This information can be gathered from experiments (past and future) as well as from experts in the field. Also, the creation of pseudodomains for characterizing nonprotein substances and small molecules would allow their analysis within the network.

Despite the lack of data on various molecular parameters (*e.g.*, rate constants), modeling at this level of detail may provide significant benefits. For example, VON DASSOW *et al.* (2000) recently described a nonlinear differential-equation model for the simulation of the segment polarity network within *Drosophila*. Surprisingly, they found that the performance of this network was not dependent on the value of specific kinetic parameters but rather achieved stability through the topology of the network itself.

Of special interest is the finding that the connectivity of vertices appears to follow a power-law distribution, exhibiting scale-free behavior. Such behavior implies that the points where a newly added protein is connected to the network will occur preferentially with proteins having greater numbers of preestablished connections (*i.e.*, a “rich get richer” phenomenon). This phenomenon has been observed within metabolic networks, and, most recently, studies by Jeong and colleagues also demonstrated the scale-free nature of the protein-protein interaction network within yeast described here (JEONG *et al.* 2000, 2001). The presence of a large number of connections may indicate a fundamentally more important, or more versatile, protein function, a possible real-world example being the protein p53. It would be particularly intriguing to consider what types of evolutionary mechanisms would give rise to particular network topologies. With the continuing accumulation of cellular and molecular data, modeling approaches such as ours should provide a more compre-

hensive picture of such molecular networks and their role in biological function.

LITERATURE CITED

- ALBERT, R., H. JEONG and A. L. BARABASI, 2000 Error and attack tolerance of complex networks. *Nature* **406**: 378–382.
- BARABASI, A. L., and R. ALBERT, 1999 Emergence of scaling in random networks. *Science* **286**: 509–512.
- BATEMAN, A., E. BIRNEY, R. DURBIN, S. R. EDDY, K. L. HOWE *et al.*, 2000 The Pfam protein families database. *Nucleic Acids Res.* **28**: 263–266.
- BENEDICT, M. A., Y. HU, N. INOHARA and G. NUNEZ, 2000 Expression and functional analysis of Apaf-1 isoforms. Extra Wd-40 repeat is required for cytochrome c binding and regulated activation of procaspase-9. *J. Biol. Chem.* **275**: 8461–8468.
- GILKS, W. R., S. RICHARDSON and D. J. SPIEGELHALTER (Editors), 1996 *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, New York.
- GOTO, S., H. BONO, H. OGATA, W. FUJIBUCHI, T. NISHIOKA *et al.*, 1997 Organizing and computing metabolic pathway data in terms of binary relations. *Pac. Symp. Biocomput* **2**: 175–186.
- GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. **82**: 711–732.
- HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.
- HU, Y., L. DING, D. M. SPENCER and G. NUNEZ, 1998 WD-40 repeat region regulates Apaf-1 self-association and procaspase-9 activation. *J. Biol. Chem.* **273**: 33489–33494.
- JEONG, H., B. TOMBOR, R. ALBERT, Z. N. OLTVAI and A. L. BARABASI, 2000 The large-scale organization of metabolic networks. *Nature* **407**: 651–654.
- JEONG, H., S. P. MASON, A. L. BARABASI and Z. N. OLTVAI, 2001 Lethality and centrality in protein networks. *Nature* **411**: 41–42.
- KOIKE, T., and A. RZHETSKY, 2000 A graphic editor for analyzing signal-transduction pathways. *Gene* **259**: 235–244.
- PEITGEN, H.-O., H. JURGENS and D. SAUPE, 1992 *Chaos and Fractals: New Frontiers of Science*. Springer-Verlag, New York.
- UETZ, P., L. GIOT, G. CAGNEY, T. A. MANSFIELD, R. S. JUDSON *et al.*, 2000 A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623–627.
- VON DASSOW, G., E. MEIR, E. M. MUNRO and G. M. ODELL, 2000 The segment polarity network is a robust developmental module. *Nature* **406**: 188–192.
- XENARIOS, I., D. W. RICE, L. SALWINSKI, M. K. BARON, E. M. MARCOTTE *et al.*, 2000 DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**: 289–291.

Communicating editor: N. TAKAHATA