



Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome

Andrey Rzhetsky^{1, 2,*} and Shawn M. Gomez¹

¹Columbia Genome Center and ²Department of Medical Informatics, Columbia University, New York, NY 10032, USA

Received on April 20, 2001; revised and accepted on July 13, 2001

ABSTRACT

Motivation: Current growth in the field of genomics has provided a number of exciting approaches to the modeling of evolutionary mechanisms within the genome. Separately, dynamical and statistical analyses of networks such as the World Wide Web and the social interactions existing between humans have shown that these networks can exhibit common fractal properties—including the property of being scale-free. This work attempts to bridge these two fields and demonstrate that the fractal properties of molecular networks are linked to the fractal properties of their underlying genomes.

Results: We suggest a stochastic model capable of describing the evolutionary growth of metabolic or signal-transduction networks. This model generates networks that share important statistical properties (so-called scale-free behavior) with real molecular networks. In particular, the frequency of vertices connected to exactly k other vertices follows a power-law distribution. The shape of this distribution remains invariant to changes in network scale: a small subgraph has the same distribution as the complete graph from which it is derived. Furthermore, the model correctly predicts that the frequencies of distinct DNA and protein domains also follow a power-law distribution. Finally, the model leads to a simple equation linking the total number of different DNA and protein domains in a genome with both the total number of genes and the overall network topology.

Availability: MatLab (MathWorks, Inc.) programs described in this manuscript are available on request from the authors.

Contact: ar345@columbia.edu

INTRODUCTION

Little is known about the mechanisms behind the origin and evolution of molecular networks, such as metabolic networks in bacteria and signal-transduction networks in

eukaryotes. An accepted approach for studying such complex systems often includes the formulation of a model (hypothesis), with eventual comparison of specific model predictions to physical reality. In this article, we describe a mathematical model that generates hypothetical molecular networks sharing important statistical similarities to the networks observed in nature. Furthermore, this model leads to a simple equation linking the network topology to the total number of ‘building blocks’ that the network comprises.

This paper has the following composition. First, we introduce terminology and derive a model of genesis for molecular networks. Next, we fit the model to experimental data, derive an estimator for the number of distinct protein and DNA motifs/domains required for maintaining network topology, and obtain corresponding estimates for two genomes. Finally, we discuss the applicability of our model to several types of scale-free networks.

REAL-LIFE NETWORKS ARE OCCASIONALLY SCALE-FREE

We use the term network to denote an oriented graph, defined by a set of vertices and a set of oriented edges. For example, a large collection of Hypertext Markup Language (HTML) documents can be viewed as an oriented graph. Each vertex of such a graph corresponds to a unique HTML document; incoming edges represent references to this document from other documents, and each hyperlink defined within this document corresponds to an outgoing edge. Another example of a network is a collection of journal articles representing the development of a scientific discipline over a given period of time. Each article can be mapped to a unique graph vertex; references to other articles mentioned within this article map to outgoing edges, and references to this article map to incoming edges.

In a molecular network, vertices of the graph correspond to molecules, such as proteins, genes, RNAs, lipids, ions, and the like. An oriented edge between two substances in

*To whom correspondence should be addressed.

such a network represents an interaction (such as binding or phosphorylation) between corresponding molecules, and the direction of the edge is defined by the temporal order of events.

The networks just mentioned exhibit so-called scale-free behavior (Albert *et al.*, 2000; Barabasi and Albert, 1999; Cohen *et al.*, 2000; Jeong *et al.*, 2000). In a nutshell, this means that a small part of each network has essentially the same statistical properties as any larger part; that is, the networks are self-similar. More specifically, for each of these networks, a plot of the observed frequency of vertices having exactly k outgoing (or incoming) edges as a function of k , in log–log coordinates, appears as a straight line with negative slope (Figure 1a). The value of the slope can differ for different networks; for each individual network, however, the slope is constant, and is the same whether it is for the whole network or for but a small part thereof.

For all scale-free networks, the frequency of vertices having exactly k outgoing (or incoming) edges obeys the following equation (Albert *et al.*, 2000; Barabasi and Albert, 1999; Cohen *et al.*, 2000; Jeong *et al.*, 2000):

$$p(k) = c \cdot k^{-\gamma}, \quad (1)$$

where c is a normalizing constant and parameter γ varies across networks, but usually has a value between 1 and 3.

We know that such scale-free properties must be part of any network that is derived from our model. We hope that, by incorporating only a few (perhaps minimal) biologically realistic mechanisms into our model, the desired scale-free properties will emerge naturally.

EXISTING MODELS: MOST RANDOM GRAPHS ARE NOT SCALE-FREE

On one hand there are a number of existing models of growing random graphs that have some relevance to regulatory networks. The simplest stochastic model (Erdos and Rényi, 1960) starts with a set of unconnected vertices and then proceeds through all possible pairs of vertices making a new edge with a constant probability. This and a few other more complicated models (Jespersen and Blumen, 2000; Kuperman and Abramson, 2001; Mathias and Gopal, 2001; Newman *et al.*, 2000; Roy *et al.*, 2001) produce graphs with a bell-shaped rather than a power-law connectivity distribution. To obtain random graphs with scale-free properties, the existing models explicitly assume that the graph (network) is growing via addition of new vertices and new edges in such a way that the probability of a new vertex being connected to an ‘old’ vertex is proportional to connectivity (the number of edges incident) of the old vertex (Aiello *et al.*, 1999; Albert and Barabasi, 2000; Albert *et al.*, 2000; Barabasi and Albert, 1999; Krapivsky *et al.*, 2000; Lander *et al.*, 2001; Roy *et al.*, 2001).

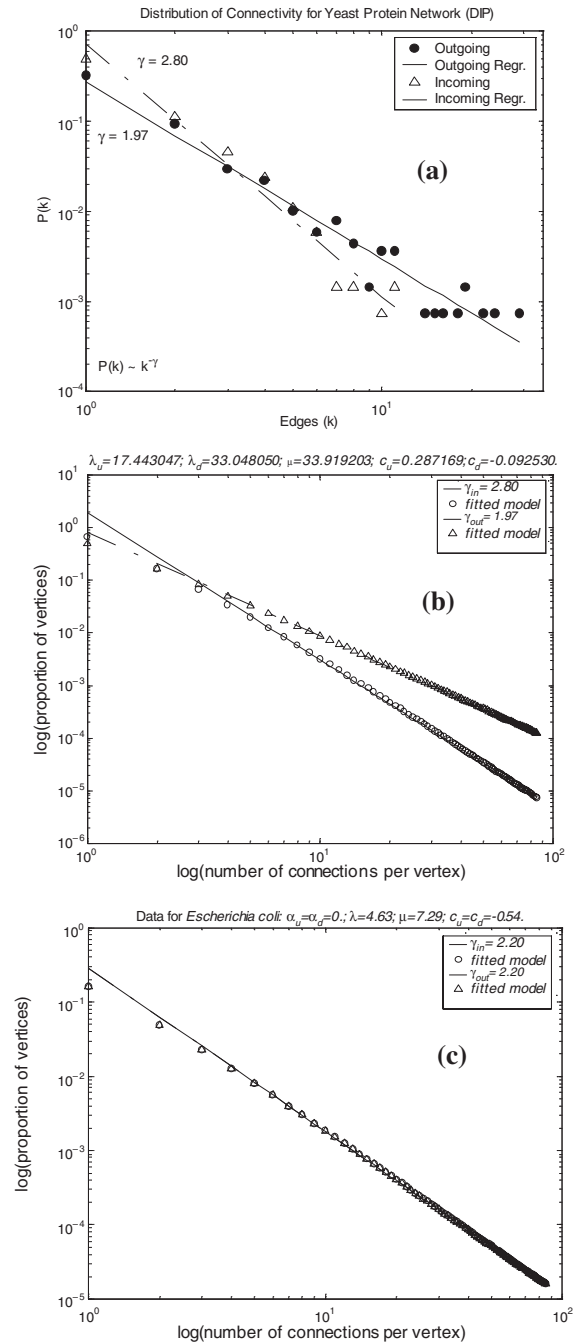


Fig. 1. Fitting of the model described in this article to yeast and *E.coli* data. (a) The frequency of network vertices connected to exactly k incoming edges (open triangles) or k outgoing edges (closed circles) as a function of the number of edges connected (k) in log–log coordinates. The distributions are well approximated with straight lines with slope -2.80 (incoming edges) and -1.97 (outgoing edges). As a noted added in proof, after this article was submitted for review, the scale-free properties of the yeast data were reported by Jeong *et al.* (2001). (b) Fit of the model to the straight lines corresponding to yeast data; parameter estimates are shown in Table 1. (c) Fitting of the network evolution model to *E.coli* data.

On the other hand, there are a number of studies aimed at modeling large-scale genomic changes in evolution (Aravind *et al.*, 2001; Sankoff and Goldstein, 1989; Slanina and Kotrla, 2000; Valdivia, 1999; Yanai *et al.*, 2000), which focus on a number of important problems other than topological properties of the corresponding molecular network.

The goal of this study is to formulate the simplest biologically plausible model of evolution for molecular networks that would generate scale-free graphs as a result of underlying biological assumptions rather than an explicit postulation of a graph-theoretic mechanism.

BUILDING BLOCKS OF NETWORK FUNCTIONALITY: DNA MOTIFS AND PROTEIN DOMAINS

To model the growth of molecular networks, we need to specify an implementation for the encoding of oriented edges. In HTML documents, the outgoing edges are encoded as a list of unique addresses of ‘downstream’ documents. In molecular networks each edge is implemented as a pair of mutually specific molecular structures. For example, in the case of an edge corresponding to phosphorylation of protein B by protein A, the beginning of an edge in protein A is encoded as a kinase domain, whereas the ending of the same edge in protein B is encoded as a specific protein domain recognized by the kinase domain of protein A. In another example, the beginning of an edge is encoded as a DNA-binding domain of a transcription factor, whereas the ending of the same edge at a gene regulated by the transcription factor is encoded by a DNA motif specifically recognized by the DNA-binding domain.

Therefore, we assume that each vertex of a molecular network has either both upstream and downstream domains, or just one of the domains. Furthermore, we use the oversimplified assumption that each individual molecule has no more than two domains. Finally, we assume that upstream and downstream domains undergo independent duplication and that after duplication, each new domain copy randomly picks a domain from the available pool of domains of the opposite type (upstream or downstream) to form a two-domain protein. If there is no domain of the opposite type available, a one-domain protein is formed (see Figure 2).

How far does this assumption diverge from reality? If, in fact, multiple upstream and downstream domains are commonly used, the total number of domains per gene or protein should correlate with the number of incoming and outgoing edges. We have found that there is no detectable correlation between the number of domains and the number of connections per protein (data not shown) and assume that one upstream and one downstream domain per gene/protein is reasonable for first-pass modeling (Gomez *et al.*, 2001).

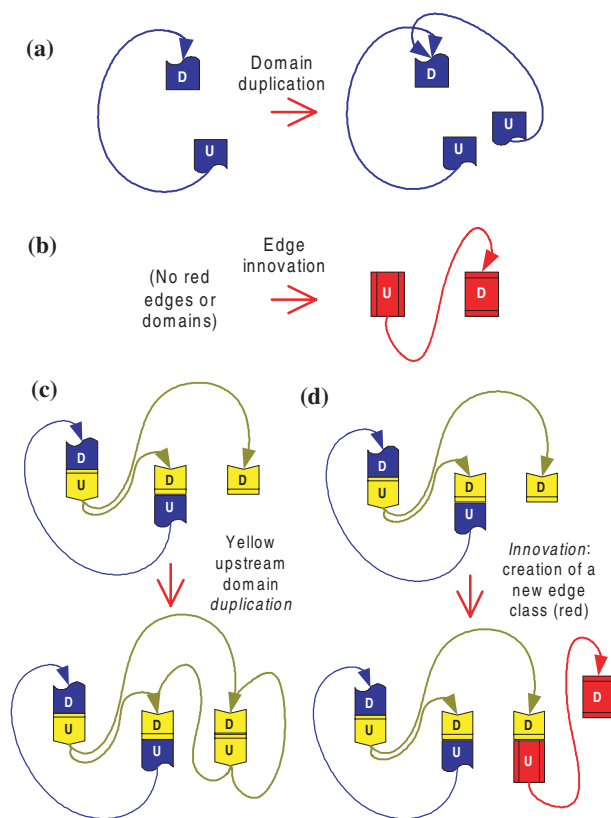


Fig. 2. (a, b) Evolutionary events affecting network growth: *duplication*, (a), and *innovation* (edge birth), (b). Note that duplication operates on domains, while innovation operates on edges. In the figure all upstream domains are marked with letter ‘U’ while all downstream domains are marked with letter ‘D’. (c, d) Examples of domain duplication, (c), and edge innovation, (d), in the same simple three-protein network. As long as the duplicated domain preserves all connections of the parental copy, connectivity of the network grows with each domain duplication. For the network at the top of figures (c) and (d) we have the following network parameter values: $D(t) = 2$; $d_{1,u} = 2$ (blue and yellow); $d_{2,u} = 0$; $d_{1,d} = 1$ (blue); $d_{2,d} = 1$ (yellow); $d_{3,d} = 0$. (c) For the network at the bottom: $D(t) = 2$; $d_{1,u} = 1$ (blue); $d_{2,u} = 2$ (yellow); $d_{3,u} = 0$; $d_{1,d} = 1$ (blue); $d_{2,d} = 1$ (yellow); $d_{3,d} = 0$. (d) For the network at the bottom: $D(t) = 3$; $d_{1,u} = 3$ (red, yellow, blue); $d_{2,u} = 0$; $d_{1,d} = 2$ (blue and red); $d_{2,d} = 1$ (yellow); $d_{3,d} = 0$.

We can actually describe both HTML and article-reference networks in terms of upstream and downstream domains: an upstream domain corresponds to a document unique identifier, and a downstream domain is an HTML link or article reference. However, there is an important difference between a molecular network and the two other network types: the latter include a strictly enforced requirement that each vertex’s upstream domain (identifier) be unique. This requirement clearly does not hold for molecular networks: the same transcription factor often

activates more than one gene, so several distinct genes may have the same upstream domain. The downstream domains do not have to be unique in any of the three types of networks. For example, many different documents can reference the same document and therefore can share the same ‘downstream domain’.

Now we will try to clarify the notion of ‘domain’ as it is used in this article. There are at least two frequently used definitions of a protein domain. Structural biologists define domains based on visual analysis of three-dimensional protein structures by human experts. Computational biologists define domains mainly through similarities in the primary sequences of several proteins. Despite considerable ideological differences between these definitions, both definitions are used in practice and the currently popular databases, such as InterPro (Apweiler *et al.*, 2001) and Pfam (Bateman *et al.*, 2000) incorporate domains defined in both ways.

For the purpose of our modeling, a domain is defined as a functional unit that provides a specific interaction between two molecules. Note that, in the case of downstream domains, our definition lumps together protein and DNA domains so as to include network edges that correspond to transcription activation/inhibition events. In our analysis of the real data presented in this article, we cannot help but use InterPro and Pfam databases to identify protein domains, but the reader is advised to keep in mind that this approach is not formally sufficient in satisfying the model assumptions.

CLASSES OF EDGES AND TWO TYPES OF EVENTS ASSOCIATED WITH NETWORK EVOLUTION

To simplify our job in describing the model we define *classes of edges*. As long as neither upstream nor downstream domains, across all vertices in the network, are required to be unique, a molecular network can use multiple copies of the same domain to encode multiple edges of the same kind. To distinguish between types of edges and different edges of the same type, we say that all edges of a network, formed by the same set of two domains, belong to the same class. In our model the number of edges in each class is allowed to fluctuate randomly, but it never falls below 1, which corresponds to the assumption that natural selection discourages evolutionary loss of a function.

The major simplification in the analytical treatment of the model comes from the idea that, having classes of edges defined, we will monitor only the numbers of upstream and downstream domains in each class, without paying attention to how these domains are combined into genes or proteins. As is common in modeling stochastic processes, we assume that network evolution is governed by a homogeneous continuous-time Markov

process, where individual changes in the network arrive spontaneously at constant rates.

Now we are ready to define the major evolutionary events that affect network growth. The first type of event is the duplication of genes and proteins (see Figure 2a). We assume that the probability of observing the duplication of a single domain during a small time interval Δt is equal to $\lambda \Delta t$, where λ is the instantaneous rate of duplication. Upstream and downstream domains may be duplicated with separate rates, with rate λ_u for upstream domains and λ_d for downstream domains. The second type of event is the birth of new classes of edges (see Figure 2b). We assume that the rate of growth of the number of edge classes over a short time interval, Δt , is proportional to the product of the current number of edge classes, $D(t)$, and the innovation rate, μ :

$$D(t + \Delta t) = D(t) + \mu D(t) \Delta t, \quad (2)$$

which, as is easily verified, leads to an exponential growth of $D(t)$ over time.

To conclude our model definition, we introduce two more variables: $d_{i,u}(t)$ and $d_{j,d}(t)$. Here $d_{i,u}(t)$ indicates the number of classes at time t that have exactly i copies of the same upstream domain (u subscript stands for *upstream*). The second variable, $d_{j,d}(t)$, is defined in the same way, but refers to downstream domains. From this definition it becomes clear that both $d_{j,d}(t)$ and $d_{i,u}(t)$ must sum to the total number of edge classes:

$$\sum_i d_{i,u}(t) = \sum_j d_{j,d}(t) = D(t). \quad (3)$$

ANALYZING THE MODEL

Now, equipped with all necessary information, we derive the difference equations that describe the system parameters at time $(t + \Delta t)$ as a function of the system parameters at time t :

$$\left\{ \begin{array}{l} d_{1,u}(t + \Delta t) = d_{1,u}(t) + \mu D(t) \Delta t \\ \quad \quad \quad - \lambda_u d_{1,u}(t) \Delta t + o(\Delta t), \\ \dots \\ d_{i,u}(t + \Delta t) = d_{i,u}(t) + \lambda_u d_{i-1,u}(t) (i-1) \Delta t \\ \quad \quad \quad - \lambda_u d_{i,u}(t) i \Delta t + o(\Delta t), \\ \quad \quad \quad \text{where } 2 \leq i \leq N-1, \\ \dots \\ d_{N,u}(t + \Delta t) = d_{N,u}(t) + \lambda_u d_{N-1,u}(t) \\ \quad \quad \quad \times (N-1) \Delta t + o(\Delta t). \end{array} \right. \quad (4)$$

In a nutshell, the system describes the birth of new types of edges, which arrive into our hypothetical world always in a single copy, and, if they are lucky, their subsequent ‘diffusion’ from the one-copy state to a many-copies state. Since one of the assumptions of a Poisson (and therefore a Markovian) process is that the probability of observing

more than one random event over a small time interval, Δt , is negligible (or, more precisely, is of order $o(\Delta t^2)$), all terms of each difference equation reflect dynamics involving only one random event at a time.

Let us focus on the right-hand side of the first equation. Recall that $d_{1,u}(t)$ is the number of edge classes that have exactly one copy of the upstream domain. The increase in $d_{1,u}(t)$ therefore occurs due to the birth of new classes. The birth of new classes occurs at rate μ and is proportional to the current total number of classes, $D(t)$. The decrease in the number of single-copy classes can occur through one of the classes acquiring an extra domain copy and moving to a two-copy group via domain duplication mechanism.

The second equation, for $d_{i,u}$ ($i > 1$), describes gains and losses suffered by the population of i -domain classes. The gains come from the collection of $(i - 1)$ -domain classes, as some of them acquire an additional domain copy due to domain duplication. The decrease in the number of i -copy classes occurs due to domain gain by duplication. Note that the rate of duplication in each difference equation depends not only on the number of classes, but also on the number of domain copies in each class.

The total number of edge classes in the whole system grows only through an increase in the number of single-copy classes. This fact is rather intuitive, because it is unlikely that mutation and selection would produce more than one copy of a new edge at once. To make the system computationally manageable, we restricted the maximum number of domain copies to N . However, for practical computations, N can be made reasonably large, so that system behavior will be essentially the same as with an infinite N . (In our data-fitting exercises we used $N = 150$.)

It is convenient to change variables from the numbers of classes in each category, $d_{i,u}(t)$, to the proportions of classes, $p_{i,u}(t)$, using the following definition.

$$p_{i,u}(t) = \frac{d_{i,u}(t)}{D(t)}, \tag{5}$$

where proportions of i -copy populations of upstream and downstream domains independently sum to 1. Rearranging the left-hand side of each equation to obtain $[p_{i,u}(t + \Delta t) - p_{i,u}(t)]/\Delta t$ and evaluating the limit of both sides of each equation as Δt goes to zero, we obtain the following system of differential equations. (For brevity we shorten the notation of $p_{i,u}(t)$ to $p_{i,u}$ since the time value is the

same for all functions in the system.)

$$\begin{cases} \dot{p}_{1,u} = \mu(1 - p_{1,u}) - \lambda_u p_{1,u}, \\ \dots \\ \dot{p}_{i,u} = \lambda_u(i - 1)p_{i-1,u} - \lambda_u i p_{i,u} - \mu p_{i,u}, \\ \dots \\ \dot{p}_{N,u} = \lambda_u(N - 1)p_{N-1,u} - \mu p_{N,u}. \end{cases} \tag{6}$$

where $2 \leq i \leq N - 1$,

The intensity matrix corresponding to this system of ordinary linear differential equations is shown in equation (7). (As is required of an intensity matrix, each row sum is equal to 0, and the absolute value of each diagonal element is equal to the sum of the off-diagonal elements in the same row.)

$$Q_u = \begin{pmatrix} -\lambda_u & \lambda_u & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \mu & -\mu - 2\lambda_u & 2\lambda_u & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \mu & 0 & -\mu - 3\lambda_u & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mu & 0 & 0 & \dots & -\mu - i\lambda_u & i\lambda_u & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mu & 0 & 0 & \dots & 0 & 0 & 0 & \dots & -\mu - (N-1)\lambda_u & (N-1)\lambda_u \\ \mu & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & -\mu \end{pmatrix}. \tag{7}$$

The solution for this system is given in terms of the intensity matrix as

$$\pi(t) = \pi(0)\mathbf{P}_u(t) = \pi(0)e^{Q_u t}, \tag{8}$$

where

$$\pi(t) = [p_{1,u}(t), p_{2,u}(t), \dots, p_{N,u}(t)],$$

and

$$\pi(0) = [1, 0, 0, \dots, 0].$$

This equation is easy to evaluate numerically—for example, in MatLab—for a specified set of parameter values.

The complete derivation routine for the upstream domains can be exactly recapitulated for downstream domains: the only change required in equations (4)–(8) is the substitution of all occurrences of subscript u with subscript d .

Note that under our model we can have vertices with indegree or outdegree equal to zero. This could happen if the rates of duplication for upstream and downstream domains are not equal: then some of the domains of the type present in excess may not be able to find a free domain of another kind to form a two-domain protein. The scale-free properties of networks clearly would not apply to the zero-degree vertices and we shall exclude these vertices from further consideration.

Now we are coming across an unexpected property of our model: the proportions of the k -copy classes of

upstream and downstream domains ($p_{k,u}$ and $p_{k,d}$) and the frequency of vertices with exactly k incoming (or outgoing) edges ($E_{k,in}$ and $E_{k,out}$) turn out to be equal.

$$\begin{aligned} E_{k,in}(t) &= p_{k,u}(t), \\ E_{k,out}(t) &= p_{k,d}(t). \end{aligned} \quad (9)$$

This property of our model becomes obvious from analysis of examples of networks shown in Figure 2: the number of incoming edges for each downstream domain is exactly equal to the number of corresponding upstream domains, and, *vice versa*, the number of outgoing edges for each upstream domain is equal to the number of corresponding downstream domains. Therefore, the proportion of vertices with exactly k incoming edges is equal to the proportion of upstream domain classes with exactly k domain copies.

Now we can estimate parameters of the model by fitting simultaneously the proportions of k -connected vertices expected under our model to the corresponding proportions observed in a real network:

$$\sum_{k=1}^N (O_{k,in} - E_{k,in})^2 + \sum_{k=1}^N (O_{k,out} - E_{k,out})^2 \rightarrow \min. \quad (10)$$

FITTING THE MODEL TO REAL DATA

We estimated parameter γ (see equation (1)) separately for the connectivity distributions of incoming and outgoing edges in yeast pathways (we used the Database of Interacting Proteins (DIP); Xenarios *et al.*, 2000, 2001; see Figure 1a). We fitted our model to the estimated power-law plots estimated for two species, yeast and *Escherichia coli* (see Figures 1b and c). (We implemented the fitting in a MatLab program; this same program was used to produce all plots shown in this article.)

The primary purpose of fitting the model to the real data was to verify that the model can indeed produce scale-free distributions (that is, curves approaching straight lines in logarithmic coordinates) of frequencies of vertices incident to exactly k edges (see Figures 1a–c). To our delight, the model indeed turned out to be capable of reproducing the scale-free distributions. Furthermore, from equation (9), it follows that the distribution of frequencies of domains per genome also should be scale-free, and our analysis of real data confirmed this prediction (see Figure 3). We will further exploit this observation later in this article to predict the total number of distinct domains per genome (see next section).

In the model fitting, we considered data for incoming and outgoing edges simultaneously, obtaining independent duplication rate estimates for the incoming and outgoing edges while estimating the unique innovation rate. In

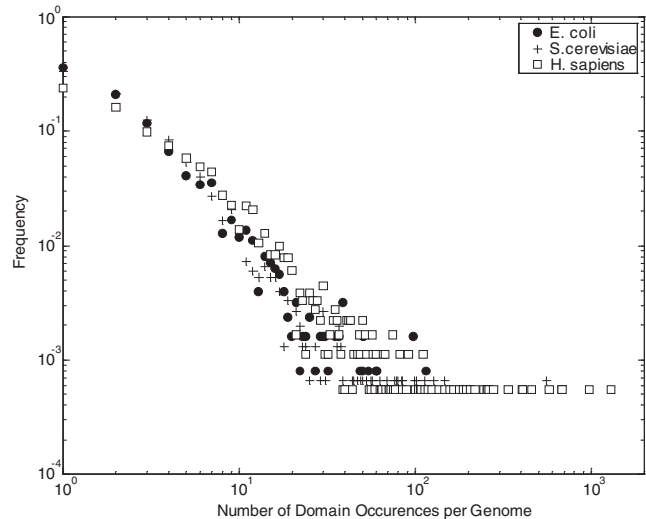


Fig. 3. Frequency of domain appearances in *E.coli* (closed circles), *Saccharomyces cerevisiae* (crosses), and *Homo sapiens* (open rectangles). Domains were identified with HMMER 2.1.1 (<http://hmmmer.wustl.edu>). For *E.coli* and *S.cerevisiae*, all known proteins were searched; for *H.sapiens*, domains from approximately 8000 proteins were used in the generation of this plot.

Table 1. Parameter values estimated through fitting the model to experimental values

| | $\lambda_{u}t$ | $\lambda_{d}t$ | μt |
|---------------|----------------|----------------|---------|
| Yeast | 17.44 | 33.05 | 33.92 |
| <i>E.coli</i> | 4.63 | 4.63 | 7.29 |

estimations of the yeast data (see Table 1), the duplication rate of the downstream domains was almost twice as large as the duplication rate of the upstream domains. In the case of the *E.coli* regulatory network, the slopes of the connectivity graphs for incoming and outgoing edges are equal. As a consequence, fitting of the model to this data results in identical parameter estimates for the upstream and the downstream domains (see Table 1).

HOW MANY DISTINCT DOMAINS ARE IN EXISTENCE?

Protein and DNA domain/motif databases have recently achieved dramatic improvements, but how far are they from providing a complete catalog of domains? Our recent analysis of Baker's yeast open reading frames with Pfam, as well as the proteome analysis of the European Bioinformatics Institute using the InterPro domain collections (Apweiler *et al.*, 2001; Bateman *et al.*, 2000), indicated that about 40% of all yeast reading

frames are not covered by any known domains. How many domains/motifs are still at large, unrecognized and uncounted? The conclusion that we reached in the previous section, that the domain/motif frequency follows a power-law distribution, implies that there should be many motifs/domains that occur with only one copy per genome. Furthermore, our model can help us to estimate the total number of distinct domains in any given genome, provided that we know the slope of the power-law curve for the distribution of network vertex connectivity.

The algebraic road to estimating the total number of distinct domains is laid through an expression for estimating the total number of distinct edge classes.

Parameters γ_u and γ_d can be directly estimated from real molecular network data (according to equation (9), γ_u and γ_d should have the same values as parameters γ_{in} and γ_{out} , respectively) even if these data are incomplete (because of the scale-free property, these parameters can be correctly estimated from any part of the network). Parameters c_u and c_d cannot be estimated from partial network data, but can be expressed in terms of γ_u and γ_d , respectively, if the total number of network vertices, V , in the network is known. Indeed, since a network with V vertices cannot have a vertex incident to more than V edges (a network graph is allowed to have two edges with the opposite direction per every pair of vertices, but not multiple edges with the same direction), the total number of domains of the same type is also bound by V , and we have the following identities.

$$p_{i,u} = c_{in} i^{-\gamma_{in}}; \quad p_{j,d} = c_{out} j^{-\gamma_{out}}; \\ c_{in} \sum_{i=1}^V i^{-\gamma_{in}} = c_{out} \sum_{j=1}^V j^{-\gamma_{out}} = 1. \quad (11)$$

Therefore, parameter c_u is expressed in terms of network size (V) and as

$$c_{in} = \left(\sum_{i=1}^V i^{-\gamma_{in}} \right)^{-1}. \quad (12)$$

Furthermore, summing expression (11) over all types of vertices in the network with V vertices, we obtain a relationship between the total number of distinct edge classes, $D(t)$, and the number of network vertices, V :

$$V = \sum_{i=1}^V p_{i,u} \cdot i \cdot D(t) = c_{in} \sum_{i=1}^V i^{-\gamma_{in}} \cdot i \cdot D(t). \quad (13)$$

We can then combine together equations (12) and (13), and obtain a simple equation connecting the total number of different edge classes with the total number of network vertices and parameter γ_{in} .

$$D(t) = V \frac{\sum_{i=1}^V i^{-\gamma_{in}}}{\sum_{j=1}^V j^{-\gamma_{in}+1}}. \quad (14)$$

Clearly, we can obtain an analogous estimate by replacing γ_{in} with γ_{out} . Note that the total number of genes per genome, G , is not equal to the number of network vertices, V . This inequality occurs because, in the network, each gene corresponds to an mRNA (or, occasionally, to more than one mRNA) and most genes correspond to proteins. Each RNA and protein molecule plays a role as an independent network vertex. In addition, there are network vertices that are not explicitly encoded by genes—such as nucleotides, lipids, sugars and ions. Therefore, the total number of network vertices is at least three times as large as the number of genes.

$$D(t) \geq 3 \cdot G \frac{\sum_{i=1}^V i^{-\gamma_{in}}}{\sum_{j=1}^V j^{-\gamma_{in}+1}}. \quad (15)$$

With the total number of genes $G = 5000$ (*E.coli*) and $\gamma = 2.2$ (Jeong *et al.*, 2000), we obtain $D > 4600$ (remember that D is the number of *pairs* of distinct domains). Similarly, assuming that for Baker's yeast, $G = 6500$, $\gamma_{in} = 2.80$, and $\gamma_{out} = 1.97$ (see Figure 1a), we obtain $D > 12900$.

DISCUSSION

With the possible exception of some bacterial metabolic networks, currently available data on molecular networks are significantly incomplete (Bono *et al.*, 1998; Goto *et al.*, 2000; Karp *et al.*, 2000; Ogata *et al.*, 1999, 2000; Overbeek *et al.*, 2000; Selkov *et al.*, 2000; Shi and Shimizu, 1998). Furthermore, the sample drawn from currently known molecular interactions is likely to be non-random. (For example, regulatory targets of the majority of transcription factors are unknown, and a large number of known open reading frames discovered in completely sequence genomes currently have no assigned function or interaction.) This implies that the parameter estimates presented in this article might turn out to be biased; as new knowledge becomes available, estimates of the total number of protein domains, and of DNA and RNA motifs, are likely to be re-established with a higher degree of accuracy. Nonetheless, the fact that the frequency distribution of protein and DNA domains follows a power law distribution is likely to be generally true for all prokaryotic and eukaryotic genomes.

As previously discussed, studies of biomolecular networks (e.g. metabolic networks) have shown that the distribution of edges follows a scale-free distribution. This work, however, suggests that the origin of this property lies within the genome's domain 'architecture'. Domains, which are complementary and capable of forming edges, appear within the genome with scale-free frequencies and thus provide the impetus for scale-free edge distributions.

To our knowledge, we have suggested the first Markovian model that describes the birth of a molecular net-

work, spontaneously exhibiting scale-free network properties, without ever explicitly imposing requirements as to what the network topology should look like. In the future, we plan to extend this model to allow for multiple downstream and upstream domains per gene or protein, and for a larger set of parameters to describe the differences in physical properties of different kinds of domains.

Our model is applicable (with a minor modification) to portraying many other phenomena, such as the growth of the World Wide Web (WWW) or of a network of scientific references. The major difference between these two networks and a molecular network is that the downstream domains (article and HTML documents identifiers) are required to be unique. To enforce this property within our model, we must set to zero the rate of duplication for all downstream domains, so that all downstream domains eternally stay in a single copy, while the upstream domains (references to other articles or HTML documents) are free to multiply without a restriction.

ACKNOWLEDGEMENTS

We are very grateful to Lyn Dupre, Jeff Thorne, Jim Russo, and two anonymous reviewers for numerous helpful comments on the earlier versions of this article.

REFERENCES

- Aiello, W., Chung, F. and Lu, L. (1999) A random graph model for massive graphs. In *Thirty Second ACM Symposium on the Theory of Computing*. pp. 171–180.
- Albert, R. and Barabasi, A.L. (2000) Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, **85**, 5234–5237.
- Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.
- Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M. and Servant, F. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
- Aravind, L., Dixit, V.M. and Koonin, E.V. (2001) Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science*, **291**, 1279–1284.
- Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.
- Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
- Bono, H., Ogata, H., Goto, S. and Kanehisa, M. (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res.*, **8**, 203–210.
- Cohen, R., Erez, K., ben-Avraham, D. and Havlin, S. (2000) Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, **85**, 4626–4628.
- Erdos, P. and Rényi, A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, **7**, 17–61.
- Gomez, S.M., Lo, S.H. and Rzhetsky, A. (2001) Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics*, accepted.
- Goto, S., Nishioka, T. and Kanehisa, M. (2000) LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.*, **28**, 380–382.
- Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
- Jespersen, S. and Blumen, A. (2000) Small-world networks: links with long-tailed distributions. *Phys. Rev. E. Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, **62**, 6270–6274.
- Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M. and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.*, **28**, 56–59.
- Krapivsky, P.L., Redner, S. and Leyvraz, F. (2000) Connectivity of growing random networks. *Phys. Rev. Lett.*, **85**, 4629–4632.
- Kuperman, M. and Abramson, G. (2001) Small world effect in an epidemiological model. *Phys. Rev. Lett.*, **86**, 2909–2912.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrum, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Showkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Mathias, N. and Gopal, V. (2001) Small worlds: how and why. *Phys. Rev. E. Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics.*, **63**, 021117.
- Newman, M.E., Moore, C. and Watts, D.J. (2000) Mean-field solution of the small-world network model. *Phys. Rev. Lett.*, **84**, 3201–3204.
- Ogata, H., Fujibuchi, W., Goto, S. and Kanehisa, M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, **28**, 4021–4028.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
- Overbeek, R., Larsen, N., Pusch, G.D., D'Souza, M., Selkov, E.Jr,

- Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.
- Roy,S., Asavathiratham,C., Lesieutre,B.C. and Verghese,G.C. (2001) Network models: growth, dynamics, and failure. In *Thirty Fourth Hawaii International Conference on System Science*. IEEE, Hawaii.
- Sankoff,D. and Goldstein,M. (1989) Probabilistic models of genome shuffling. *Bull. Math. Biol.*, **51**, 117–124.
- Selkov,E., Overbeek,R., Kogan,Y., Chu,L., Vonstein,V., Holmes,D., Silver,S., Haselkorn,R. and Fonstein,M. (2000) Functional analysis of gapped microbial genomes: amino acid metabolism of *Thiobacillus ferrooxidans*. *Proc. Natl Acad. Sci. USA*, **97**, 3509–3514.
- Shi,H. and Shimizu,K. (1998) On-line metabolic pathway analysis based on metabolic signal flow diagram. *Biotechnol. Bioeng.*, **58**, 139–148.
- Slanina,F. and Kotrla,M. (2000) Random networks created by biological evolution. *Phys. Rev. E. Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, **62**, 6170–6177.
- Valdivia,R.H. (1999) Regulatory network analysis. *Trends. Microbiol.*, **7**, 398–399.
- Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: the Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.
- Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the Database of Interacting Proteins. *Nucleic Acids Res.*, **28**, 289–291.
- Yanai,I., Camacho,C.J. and DeLisi,C. (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys. Rev. Lett.*, **85**, 2641–2644.