Published in final edited form as:
Comput Biol Chem. 2004 October ; 28(4): 257–264.

Scale-free networks versus evolutionary drift

Teresa M. Przytycka[a,*] and **Yi-Kuo Yu**[a,b]
a *NCBI/NLM/NIH 8600 Rockville Pike, Bethesda, MD 20894, USA*

b *Department of Physics, Florida Atlantic University, Boca Raton, FL 33431, USA*

Abstract

Recent studies of properties of various biological networks revealed that many of them display scale-free characteristics. Since the theory of scale-free networks is applicable to evolving networks, one can hope that it provides not only a model of a biological network in its current state but also sheds some insight into the evolution of the network. In this work, we investigate the probability distributions and scaling properties underlying some models for biological networks and protein domain evolution. The analysis of evolutionary models for domain similarity networks indicates that models which include evolutionary drift are typically not scale free. Instead they adhere quite closely to the Yule distribution. This finding indicates that the direct applicability of scale-free models in understanding the evolution of biological network may not be as wide as it has been hoped for.

Keywords

Biological networks; Evolutionary drift; Domain evolution; Scale-free models; Yule distribution

1. Introduction

The functioning of a biological system largely depends on the mutual interactions among its constituent components such as proteins. It is a common practice to represent such a system by a network, within which objects are represented as nodes and relations are represented as edges linking related pairs of nodes. A biological network is broadly defined as any network (graph) where the nodes are identified with some biologically relevant entities and edges define a relation over these entities. For example, when one represents each protein as a node and draws an edge to connect two nodes if the two corresponding proteins interact with each other, one obtains a graph representation of the protein-protein interaction network (Jeong, Mason, Barabasi, & Oltavi, 2001). In a metabolic network, nodes usually correspond to metabolites and edges to reactions (Barabasi & Albert, 1999; Fell & Wagner, 1999; Jeong, Tombor, Albert, Oltavi, & Barabasi, 2000; Ravasz, Somera, Nonfru, Oltvai, & Barabasi, 2003; Jeong et al., 2001). Yet another biological network describes co-occurrence of protein domains within proteins. Here, the nodes correspond to domains and two domains are connected by an edge if they occur in the same protein (Wuchty, 2001). Aside from considering the interactions among different components, one may also consider the similarities among them. In this way, one may consider and investigate protein sequence similarity graphs (Yona, Linial, & Linial, 1999; Karev, Wolf, & Koonin, 2003) and protein structure similarity graphs (Dokholyan, DeLisi, Shakhnovich, & Shakhnovich, 2002). Here, the nodes correspond to protein domains and edges indicate a sequence/structure similarity. In this case it is natural to allow weighed edges, where the weight of a given edge corresponds to a numerical similarity measure. In fact, the main focus of this paper is on such weighted similarity graphs.

* Corresponding author. Tel.: +1 301 4021723; fax: +1 301 4804637. *E-mail address:* przytyck@ncbi.nlm.nih.gov (T.M. Przytycka)..

The growing acquisition rate of biological data has made it possible to address questions concerning various biological networks. At the same time, it is also natural to ask whether there exists a characteristic that is shared by different biological networks (Alm & Arkin, 2003). A number of studies suggest that the scale-free properties, observed in various evolving real world networks, is also the shared characteristic among biological networks (Barabasi & Albert, 1999; Gisiger, 2001; Wolf, Karev, & Koonin, 2002).

To describe better what one means by a scale-free network, let us first consider the probability function, $p(k)$, that records the probability for a randomly chosen node to have k edges connecting to it.[1] Formally, a network is considered scale-free provided that for any k_1, k_2 the ratio $p(k_1)/p(k_2)$ is invariant under the rescaling of k_1 and k_2. More precisely,

$$\frac{p(k_1)}{p(k_2)} = \frac{p(ak_1)}{p(ak_2)} = F\left(\frac{k_1}{k_2}\right) \quad (1)$$

where α a positive constant and F is called the scaling function. Apparently, upon the change of the scale, i.e. inflating (or deflating) k to αk the ratio $p(k_1)/p(k_2)$ remains the same, hence the term “scale free.”

As shown in Appendix A, the scale-free property (1) is satisfied if and only if the probability function $p(k)$ follows a power-law, i.e. $p(k) \propto k^{-\gamma}$. To better visualize the scale-free properties, one may graph ln $p(k)$ against ln k. The graph will show a straight line with slope $-\gamma$. Note that in order for the probability function $p(k)$ to be normalizable, i.e. $\sum_{k=1}^{\infty} p(k) = 1$, we need to have $\gamma > 1$, in the purely power-law distribution. A scale-free network considered by us will therefore have a small number of highly connected vertices (hubs) and large number of low degree vertices. Of course, this general description is not limited to the data analysis of a network system only. An example is the observation and modeling of scale-free phenomena in the size distribution of protein families (Qian, Luscombe, & Gerstein, 2001; Karev, Wolf, Rzhetsky, Berezovskaya, & Koonin, 2002).

Among the first biological networks shown to have the scale-free property were metabolic networks (Jeong et al., 2000; Ravasz et al., 2003). Why would a metabolic network evolve to be scale free? Jeong et al. (2000) pointed out that such organization makes the network robust against random error: if a randomly chosen node fails, it is unlikely to be a hub, and at the same time corrupting a low connectivity node will not disturb significantly the topology of the network. They also state that the removal of a random enzyme from the *Escherichia coli* metabolic network usually leaves the network functional. This observation suggests a correlation between the connectivity of the protein in the network and the essentiality of that protein in the metabolic pathway, and has been used to predict gene essentiality (Jeong et al., 2001; Jeong, Oltavi, & Barabasi, 2003).

Power law distribution has a long history in various disciplines including economics, social sciences, computer science and life sciences. For an attempt to provide a historical perspective see (Casselman, 2004; Mitzenmacher, 2003). Scale-free phenomena had already been robustly confirmed in other disciplines of natural science. For example, scale-free behavior occurs in the (second-order) phase transitions of physical systems such as the paramagnetic/ ferromagnetic phase transition and fractals. In these examples, the robustness of the scale-free property crucially relies on the fact that the system under consideration is of practically infinite size.

[1] In general, one can consider any set of objects, not necessarily nodes of a graph, and replace the number of neighbors with the quantity of interest in a system. For example, $p(k)$ can be the frequency of a word consisting of k characters in a dictionary.

Unfortunately, many biological networks have relatively small numbers of nodes, typically of order 10^3 or even less, and caution is necessary in analysis. For example, in order to observe the straight line characteristic for even as little as one unit on a log scale, say $k = 10, \ldots, 10^2$, we need to have a network with the number of vertices significantly larger than 10^2. Therefore, $p(k) \propto k^{-\gamma}$ can be assumed correct only for $k > k_{\min} > 1$ but significantly smaller than the total number of the nodes in the network, n. This then brings in an interesting connection to growing networks for which the total number of nodes keeps increasing. In such a context, the distribution $p(k) \propto k^{-\gamma}$ for a scale-free network will hold provided that $k_{\min} \leq k \leq k_{\max} < n(t)$ where $n(t)$ is the number of nodes in the network at time t. As the time t approaches infinity, both the total number of nodes and the $k_{\max}$ approach infinity, and the scale-free property becomes exact. For example, the world wide web (www) not only is a growing network but also follows this asymptotic scale-free property (Ebel, Mielsch, & Bornholdt, 2002; Faloutsos, Faloutsos, & Faloutsos, 1999).

The www example naturally inspires one to ask: is it possible to infer the evolutionary history of a biological network based on its current topology? The simplest evolution model proposed for scale-free networks is the preferential attachment model (Barabasi & Albert, 1999). At each step of this model, a newly generated node has a certain probability $P < 1$ to attach itself to other existing nodes. When in the attachment mode, the new node will choose to link to an existing node with probability proportional to that existing node's degree. Fell and Wagner (1999) reasoned that if a similar process was involved in the evolution of metabolic networks then highly connected vertices should correspond to metabolites that are phylogenetically the oldest. They substantiated this point by considering the evolutionary origin of most highly connected metabolites.

In a different study, Wuchty considered a protein domain network. The nodes of the network correspond to protein domains and two domains are connected by an edge if they occur together in at least one protein (Wuchty, 2001). This network was also found to display scale-free characteristics. However, no relation of the vertex connectivity to the age of the corresponding domain has been found. Furthermore, the characteristic exponents (the slopes of the distributions of $p(k)$ plotted in double logarithmic scale) vary from one kingdom to another (Wuchty, 2001). Thus, the simplest preferential attachment model of network evolution does not seem applicable here.

In pursuit of further understanding of the relation between biological networks and scale-free networks, researchers began to propose formal evolutionary models (Rzhetsky & Gomez, 2001; Qian et al., 2001; Karev et al., 2002) that explain the data. Obviously, such models are necessarily gross simplifications of evolutionary processes but nevertheless provide an important test of possible evolutionary mechanisms. The first formal scale-free model of the evolution of a biological network was proposed by Rzhetsky & Gomez (2001). Subsequently, Karev et al. (2002) proposed and analyzed carefully a theoretical model for protein domain evolution that attempts to explain the power law distribution of domain family sizes. Such theoretical models are no longer restricted by data size and can be used to predict the past and the future topology of the network. Remarkably, the evolution model proposed by Karev et al. is not scale free in general. However, when certain constraints on the evolutionary parameters are met, the distribution becomes asymptotically scale free for large values of k.

In this paper, we investigate the probability distribution and scaling properties underlying networks of protein domain evolution. We start with analyzing the Big Bang model of Dokholyan et al. (2002). In this work, the authors observed a scale-free characteristic of the data and proposed a purely divergent evolutionary model to explain the data. To study in a broader context, we designed a different model, which allows for convergent and divergent evolution and also fits the data. Interestingly, the statistics of the models are quite different and

the inability to reject any of them is only a consequence of insufficient data. Subsequently, we address the question of the relevance of the scale-free property in modeling. We show that our model follows closely the scale-free distribution while, contrary to the claim, the Big Bang does not. In fact the data obtained by simulating the Big Bang model enjoys an excellent fit to the Yule distribution. This is an interesting observation, since the relevance of the Yule distribution in biological data has been observed before (Borodovsky & Gusein-Zade, 1989; Martindale & Konopka, 1996). Using a slightly re-designed Big Bang model we demonstrated that the main contributor to the Yule-like distribution is the random drift that is present in the Big Bang model but absent from our model.

Finally, to put in more a rigorous yet simple context some informal statements about scale-free graphs, we formally show the relation between scale-free property and power law in Appendix A, and we illustrate a relation between scale-free property and fractals in Appendix B. We also show how the hypothetical drift will affect fractal-like network.

2. Two evolutionary models and two questions

In a recent paper, Dokholyan et al. (2002) analyzed a protein domain fold similarity network. The nodes of this network are protein domains (as defined in CATH classification (Orengo, Michie, Jones, Jones, Swindellsand, & Thornton, 1997) and two domains are connected with an edge if they share significant structural similarity (defined as z-value ≥ 9); see reference (Dokholyan et al., 2002) for details and the justification of the choice of z-value). Just as other biological networks discussed above, this network displays scale-free characteristics. Namely, when $\ln p(k)$ is plotted against $\ln k$, the resulting plot can be fitted with a straight line with slope -1.6 for $k = 1, \ldots, k_{\max} = 70$ (compare Fig. 3 in reference (Dokholyan et al., 2002)). The authors proposed a divergent "Big Bang" (BB) model to explain the scale-free property of the network. The model is appealing in many ways, but as we argue below, it is unlikely to be scale free. We have designed an alternative model that allows for both divergent and convergent evolution. We refer to this model as the Hierarchical Preferential Attachment (HPA) model. This model also fits the data well and generates a network which follows scale-free characteristics more closely. With these two models at hand, we would like to address two important questions.

- First, what is the constraining power of the data on evolutionary models?
- Second, is a scale-free model indeed the best way of modeling biological data?

To address the first question, we compare the fitting quality to the biological data generated from the two evolutionary models. The BB model of Dokholyan et al. (2002) assumes that the protein universe evolved from one (or a small number of) domain(s) by divergent evolution (Shakhnovich, Dokholyan, DeLisi, & Shakhnovich, 2002). The evolution proceeds in time steps in each of which the following actions are taken:

1. choose a random node and duplicate it (as a model of gene duplication);
2. choose a random number x from the interval (0, 1) to represent the distance between the parent node and the new node. This corresponds to random mutation in the duplicate. If the distance is smaller than some threshold value w (the parameter in the simulation is set to 0.75), then the new node and its parent are considered to be similar and are connected by an edge, otherwise it is assumed that they diverged to the degree that their structural similarity is no longer recognizable.
3. the distance between structural neighbors of the parent node and the new node is set randomly in such a way that triangular inequality between such neighbors, parent node and the new node is satisfied.

4. the distance between all pairs of nodes is increased by a constant D (the parameter in the simulation is set to 10^{-4}) to model the fact that the domains keep diverging. If the distance between any pair exceeds the threshold w, it is assumed that structural similarity is lost and the corresponding edge is removed from the network.

The authors show that the model fits the data for small k. The model slightly undercounts the number of connections that are below threshold w. As demonstrated in Fig. 1, it is possible to have a configuration in which two domains that are not connected by an edge are both connected by small weight edges to a common domain, thus violating the triangular inequality.

Although it would be nice to avoid such cases, computational quantification of this effect would be quite time consuming. An alternative model that generates a $p(k)$ distribution consistent with the data is proposed for the purpose of demonstration. Our Hierarchical Preferential Attachment (HPA) model contains both divergent evolution and convergent evolution, as described below.

1. Choose a random node and duplicate.
2. Choose a random number, x, from the interval (0, 1) to represent the distance between the parent node and the new node. If the distance is less than some threshold value w (the parameter in the simulation is set to 0.7) the new domain will belong to the same connected component (same family) as the parent domain.
3. If $x < w$ then the distances between the neighbors of the parent node and the new node are set according to an ultrametric condition. Specifically, if the distance between the parent node and its neighbor is y then the distance between that neighbor and the new node is set to max (x, y).
4. With probability p (the parameter in the simulation is set to 10^{-2}) the new node will be subject to convergent evolution:
5. If $x < w$, one neighbor of the duplicated node is picked with probability proportional to the degree of this neighbor (like local preferential attachment).
6. The distance between the neighbor and the new node is drawn at random from set of distances allowed by triangular inequality. If this random value is smaller than the distance computed in the previous step, we change the distance to the smaller value and restore (locally) triangular inequality if violated.
7. If $x \geq w$, the duplicate will attach itself to an existing node in the network with probability proportional to the degree of that node.[2] The new distance is set using ultrametric condition as in step 3.

Note that both models have two parameters: threshold w and "escape factor" D for the BB model, and threshold w and "convergence factor" p for the HPA model. In the choice of w for the HPA we followed the method proposed for BB model (Dokholyan et al., 2002). Similarly to the choice of $D = 10^{-4}$ in the BB model, we choose the second parameter p of the HPA model to be to 10^{-2} a value chosen to fit the data and not fully justified.

3. The limitation of finite data

We simulated both models for 5000 steps, repeated the simulation 1000 times and then took the average. The results of the simulations from the two models are presented in Fig. 2.

[2]Note that from the perspective of the connectivity of the resulting graph this action is indistinguishable from creating a node de novo with preferential attachment of the new node to existing nodes.

Note that for k in the range (1–100), the distribution functions generated by the two models are very similar. Thus the statistics collected from real data (where $k_{max} \sim 70$) do not suffice to prove either of the proposed models incorrect. An interesting related question is whether either of the two models described above is actually scale free. A formal proof of the scale-free property of a model can be quite hard and perhaps gets harder as the model gets more complicated. Looking at a finite data set cannot substitute for formal proof; however, sufficiently long simulation can shed some light on whether the network is likely to be scale free. Observing the difference between the distributions of $p(k)$ for large k in both models, it is reasonable to speculate that in the large k limit the two distributions are different. The BB method is less likely to be scale free. First, the data fit a straight line only over a short range of relatively small k. Secondly, the largest nonzero $p(k)$ value is at $k = 305$. With 5000 iterations repeated 1000 times we would expect to see a non-zero tail further on. (This is the so-called "heavy tail" property of the power law distribution). The last non-zero value with the HPA model is at around $k = 1200$ which makes it a better candidate for modeling a scale-free behavior. However, we need to remember that these are indicators based on finite data.

4. Divergent drift versus scale-free phenomena

Although the authors of the BB model made the effort to model the presumably scale-free behavior of the system, the resulting model is very unlikely to generate a scale-free network.

Yet, this does not exclude the model from being correct. In fact, the BB model is intuitively appealing and seems to fits the biological data equally well, if not slightly better than HPA model. Furthermore, there is no convincing argument that evolution of protein domain similarity networks needs to be scale free. In fact it has been argued before (Borodovsky & Gusein-Zade, 1989; Martindale & Konopka, 1996) that a Yule distribution provides a better fit than a power law to at least some biological data. While the relatively small amount of biological data makes it hard to make a strong case one way or the other, theoretical models are ideal for testing such claims.

We found an excellent fit of the Yule distribution (Yule, 1928) to the BB model (see Fig. 3).

In general, the Yule distribution is represented by

$$F(R) = aR^{-\gamma}b^{R}$$

where F is the frequency of occurrence (equivalent to our $p(k)$) and R the rank from most frequent to least frequent. Our fitting of BB model results into $p(k) = 0.19k^{-1.05}(0.99986)^{k^2}$ indicating $R = k^2$, $b = 0.99986$, $a = 0.19/2 = 0.095$, and $\gamma = (1 + 1.05)/2 = 1.025$. The excellent fit to the Yule distribution, as shown in Fig. 2 is unlikely to be a coincidence and thus it is worthwhile to examine which property of the BB model is most likely to contribute to such behavior. One possibility is that it results from finiteness of the data. However, we claim that the prime contributor to the Yule-like distribution is the random drift that is present in the BB model but absent from our HPA model. In the Appendix B, we show that the scale-free network corresponding to Sierpinski's triangle loses the scale-free property upon the introduction of a divergent drift.

To elucidate the effect of the evolutionary drift, let us consider a slightly modified Big Bang model. Namely, rather than using triangular inequality we set the distance of a newly created node to its parent's neighbors using ultrametric condition. This avoids the distance inconsistency discussed before. At each time step, each existing node duplicates with a probability p. As in original BB model we have an escape factor D. (For efficiency of the simulation the edge weights are set to be integers in interval (0, 1000) and parameter D is replaced with equivalent parameter d defining the probability of adding 1 to an edge weight

after each step.) (Table 1). Simulations were performed for different values of d with the renaming parameters $w = 0.75$ and $p = 0.001$ fixed. In consecutive trials, values of d were set to 0, 0.0625, 0.125, 0.25, and 0.5. The results of the simulation and the fits to the Yule distribution are presented in Fig. 4. The quality of the fits is striking. For simulation that does not include any drift ($d = 0$), the fit is a straight line. Defining ε by $b = 1 - \varepsilon$, we see that the distribution functions $p(k)$ at larger ε values, representing larger drifts, tend to diverge more from the power law distribution but are always well fitted by the Yule distributions.

5. Conclusions and further research

Evolutionary drift is a fundamental evolutionary mechanism. Including it in modeling of evolution is both natural and desirable. However, a model that includes such drift is not scale free unless the effect of the drift is neutralized by other less profound evolutionary mechanisms such as the “convergent” evolutions.

This provides an important reason for which a biological network may not be scale free. Therefore, although it is fashionable to draw a straight line through the log–log plot whenever the data admits it, one needs to exercise caution in interpreting such graphing as strong support for the scale-free property of a biological system. The deviation from the straight line, often attributed to the finiteness of the data, may be due to other important evolutionary mechanisms.

Although it has been argued that the power law connectivity makes a network robust against random deletion (Jeong et al., 2001), it is obvious that the same argument can also be used to support the robustness of networks whose $p(k)$ decays faster than power-law, e.g. the Yule distribution. In other words, existence of small number of highly connected hubs and a large number of weakly connected nodes does not imply that the network is scale free.

Our simulations illustrated a non-trivial relation between the parameter measuring the rate of drift and the parameters of Yule distribution. It would be interesting if one could derive such a relation using formal mathematical arguments. For instance, such attempts in expansion-modification system have been done (Li, 1991;Mansilla & Cocho, 2000). However, the mathematical derivation is not the main pursuit of this study and is deferred for future investigation. Instead, we provide below an intuitive understanding why the modified BB model will have Yule-like distributions.

It is interesting to relate the results of our simulations to the original work of Yule (1924). Yule did not consider the statistical properties of biological networks but instead the distribution of species within genera. There exists, however, a close relation between the evolutionary model Yule provided and the (modified) BB model without drift. Namely, in the Yule model any species has a certain probability of giving birth to a new species within the same genus. Furthermore, with some probability a species in any genus can also mutate enough to create a new single-species genus. This can be naturally expressed in terms of the modified BB model with $d = 0$. Note that if a similarity relation is transitive then the similarity graph consists of a set of fully connected subgraphs (cliques) In fact, this is exactly how a network created by the modified BB model without drift looks. To cast the Yule model in the network language, observe that species belonging to each genus can be naturally represented as cliques. A mutation that creates a new species within the same genus corresponds to adding a node and connecting it to all other species in the given genus. A mutation leading to creating a single-species genus corresponds to creating an isolated vertex. This is exactly the underlying idea of the (modified) BB model for $d = 0$. Thus it is not surprising that limiting behavior of both the modified BB model without drift and the Yule model is scale free. In the same paper, Yule studied the finite-time behavior of his model. In this non-equilibrium state the distribution of species within genera is better described by (what we now call) the Yule distribution. Thus,

one could hypothesize that the distribution like the one observed in Fig. 3 results from simulating a system not yet in equilibrium. But it is easy to check (by increasing simulation time) that this is not the case.

The equilibrium behavior of the (modified) BB model with drift mimics that of the non-equilibrium behavior of the Yule model. In the BB model with drift $d > 0$, the connected sub-clusters of the network are no longer cliques and the link between network statistics and population statistics becomes less obvious. In this case, an interesting parameter to look at is the clustering coefficient. The clustering coefficient of a node with k neighbors is $2c/k(k-1)$ with c being the number of edges among the k neighbors. For any vertex in a network consisting of a set of cliques, the clustering coefficient is 1. If the network consists of a number of approximately *equal density* clusters (the clustering coefficient is close to a constant), the network statistics are expected to be similar to the population statistics studied by Yule. This is no longer true if the clustering coefficient changes significantly with k. We have tested (unpublished data) the average clustering coefficient for networks created in our modified BB simulation. The clustering coefficient decreases very slowly with k and can be reasonably approximated as a constant (~1).

To be able to pinpoint the effect of the drift on the model, it was necessary to keep our models as simple as possible. Consequently, the models presented here can accommodate many natural modifications/extensions. For example, one may wish to make the divergent factor to be cluster-specific to mimic different evolutionary rates of protein families. Even within a given cluster, one may even consider a (cluster-) size dependent divergent rate to discourage too many copies of highly similar proteins. Furthermore, the combination of ultrametric distances and slow evolutionary drift makes the resulting network more *cliquish* (having higher clustering coefficient) then it is observed in real data. A more realistic model should allow for loosing connections more liberally. Doing so without running into inconsistencies similar to ones demonstrated in Fig. 1 will require a careful design.

Acknowledgements

We thank Timothy Doerr, Eugene Koonin, David Landsman, Anna Panchenko,Yuri Wolf, and in particular John Wootton for their comments.

References

Alm E, Arkin AP. Biological networks. Current Opinion in Structural Biology 2003;13:193–202. [PubMed: 12727512]

Barabasi AL, Albert R. Emergence of scaling in random networks. Science 1999;286:509–512. [PubMed: 10521342]

Borodovsky MY, Gusein-Zade SM. A general rule for ranged series of codon frequencies in different genomes. J Biomol Struct Dyn 1989;6:1001–1012. [PubMed: 2556159]

Casselman B. Networks. Notices of AMS 2004;51:293–394.

Dokholyan NV, DeLisi C, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proceedings of National Academy of Science 2002;99:14132–14136.

Ebel H, Mielsch LI, Bornholdt S. Scale-free topology of e-mail networks. Physical Review 2002;313:673–681.

Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationship of the internet topology. Comp Comm Rev 1999;29:251–262.

Fell DA, Wagner A. Emergence of scaling in random networks. Science 1999;286:509–512. [PubMed: 10521342]

Gisiger T. Scale Invariance in Biology: Coincidence or Footprint of a Universal Mechanism? Biology Review 2001;76:161–209.

Jeong H, Tombor B, Albert R, Oltavi ZN, Barabasi AL. Large-scale organization of metabolic networks. Nature 2000;5:651–654. [PubMed: 11034217]

Jeong H, Mason SP, Barabasi AL, Oltavi ZN. Lethality and centrality in protein networks. Nature 2001;411:41–42. [PubMed: 11333967]

Jeong H, Oltavi ZN, Barabasi AL. Prediction of protein essentiality based on genomic data. ComPelxUS 2003;1:19–28.

Karev G, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. Evolutionary Biology 2002;2:18. [PubMed: 12379152]

Karev G, Wolf YI, Koonin EV. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics 2003;19:1889–1900. [PubMed: 14555621]

Li W. Expansion-modification systems: a model for spatial 1/*f* spectra. Physical Review A 1991;43:5240–5260. [PubMed: 9904836]

Martindale C, Konopka AK. Oligonucleotide frequencies in DNA follow a Yule distribution. Computers and Chemistry 1996;20:28–35.

Mansilla R, Cocho G. Multiscaling in expansion-modification systems: an explanation for long range correlation in DNA. Complex Systems 2000;12:207–240.

Mitzenmacher M. A brief history of generative models of power law and lognormal distributions. Internet Mathematics 2003;1:226–251.

Orengo CA, Michie AD, Jones S, Jones DT, Swindellsand MB, Thornton JM. CATH—A hierarchic classification of protein domain structures. Structure 1997;5:1093–1108. [PubMed: 9309224]

Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrences in genomes: power-law behavior and evolutionary model. Journal of Molecular Biology 2001;313:673–681. [PubMed: 11697896]

Ravasz E, Somera AL, Nonfru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. Science 2003;297:1551–1555. [PubMed: 12202830]

Rzhetsky A, Gomez S. Birth of scale-free molecular networks and the number of distinct DNA proteins domains ore genome. Bioinformatics 2001;17:988–996. [PubMed: 11673244]

Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. Functional fingerprints of folds: evidence for correlated structure-function evolution. Molecular Biology 2002;326:1–9.

Wolf YI, Karev G, Koonin EV. Scale-free networks in biology: new insights into the fundamentals of evolution? BioEssays 2002;24:105–109. [PubMed: 11835273]

Wuchty S. Scale-free behavior in protein domain networks. Molecular Biology E 2001;18:1694–1702.

Yona G, Linial N, Linial M. ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins 1999;37:360–378. [PubMed: 10591097]

Yule G. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F. R. S. Philosophical Transactions of the Royal Society of London, Series B 1924;213:21–87.

Appendix A

Here, we show that the scale-free definition implies a power law and vice verse. To see that the power law distribution satisfies the scale-free definition (1), one may do a direct substitution and see that

$$\frac{p(k_1)}{p(k_2)} = \frac{k_1^{-\gamma}}{k_2^{-\gamma}} = \frac{(ak_1)^{-\gamma}}{(ak_2)^{-\gamma}} = \frac{p(ak_1)}{p(ak_2)} = \left(\frac{k_1}{k_2}\right)^{-\gamma} = F\left(\frac{k_1}{k_2}\right)$$

To show the deduction from the other direction, we note that from (1)

$$F\left(\frac{k_1}{k_2}\right) = \frac{p(k_1)}{p(k_2)}\frac{p(k_2)}{p(k_3)} \quad (2)$$

we have

$$F(x_1)F(x_2) = F(x_1 x_2) \tag{3}$$

This equation is the key that leads to the power law distribution. For convenience, we shall prove such a case when the variable x is continuous.[3]

Take $x_1 = x$ and $x_2 = x + \delta$ with δ an arbitrary number. We then have from (3)

$$F(x)[F(x+\delta) - 1] = F(x+\delta x) - F(x) \tag{4}$$

which, upon dividing both side by δ and utilizing $F(1) = 1$, leads to

$$F(x)\left[\frac{F(1+\delta) - F(1)}{\delta}\right] = x\left[\frac{F(x+\delta x) - F(x)}{\delta x}\right]$$

If one defines the constant $(dF(x)/dx)_{x=1}$ as $-\gamma$ one has in the $\delta \to 0$ limit the following differential equation

$$x\frac{dF}{dx} = -\gamma F(x) \tag{5}$$

whose solution is $F(x) = \beta x^{-\gamma}$. Since by definition, $F(1) = 1$, we see that the parameter $\beta = 1$. Putting this back into the condition (1), we have

$$\frac{p(k_1)}{p(k_2)} = \frac{p(ak_1)}{p(ak_2)} = \left(\frac{k_1}{k_2}\right)^{-\gamma} \tag{6}$$

Write the most general form of $p(k)$ in the form $q(k)k^{-\gamma}$. We find that $q(k_1) = q(k_2)$ for arbitrary k_1 and k_2 indicating that $q(k)$ must be a constant. We therefore have established that $p(k)$ must follow the power law when the scale-free condition is met.

Appendix B

We use Sierpinki's triangle construction to illustrate the relation of scale-free networks and fractals and the effect of evolutionary drift on this construction.

Sierpinski's triangle is a fractal constructed iteratively as illustrated in Fig. 5(A). The translation from Sierpinski's triangle construction to a network generation is quite straightforward. We let the vertices of the network correspond to the triangles of the carpet and make two vertices connected if the boundaries of the corresponding triangles intersect (see Fig. 5(B)). For uniformity, we add an extra vertex that corresponds to the outside of the external triangle (not drawn in the figure). It is easy to check that in the network obtained after s steps there are two vertices of degree 3×2^s; 3^i vertices of degree $3 \times 2^{s-i}$ where $i = 1, \ldots s$ and thus $\sim 1/2 3^{s+1}$ vertices total. Therefore,

$$\frac{p(k_1 = 3 \times 2^{s-i})}{p(k_2 = 3 \times 2^{s-j})} = \frac{3^i}{3^j} = \left(\frac{k_1}{k_2}\right)^{-\ln 3/\ln 2} \tag{7}$$

By definition (1) and Eq. (6), we find that the construction is scale free with $\gamma = \ln 3/\ln 2$. Now, consistently with the idea of divergent drift, assume that edges introduced at a given step are

[3]However, a different line of derivation can also be devised to prove the case of discrete x.

lost after some number of steps, say 10, due to a divergent drift. Then the maximum degree of a node in the network is bounded by $3 \times 2^{10} \sim 3000$ thus the scale-free property is lost. However, we will need to generate $\sim 1/22^{11} \sim 10^5$ nodes for the drift to make an impact and several orders of magnitude more nodes for the impact to be statistically significant.

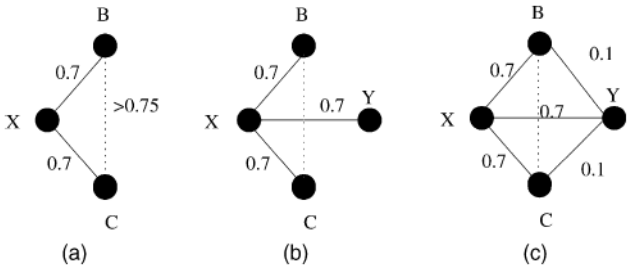

Fig. 1.

The possibility of edge assignment inconsistency in the BB model: (a) configuration before duplicating node *X*. The distance between B and C is larger than $w = 0.75$; (b) the configuration after creating *Y* (a duplicate of *X*); and (c) a possible configuration from randomly choosing weights for the edges connecting *Y* with neighbors. Observe that distances between points B, Y, and C are inconsistent with triangular inequality.

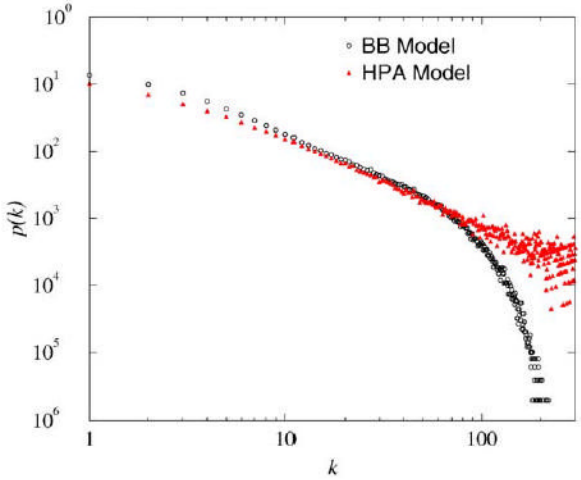

Fig. 2.
The double log graphs for $p(k)$ for the Hierarchical Preferential Attachment Model and the Big Bang model.

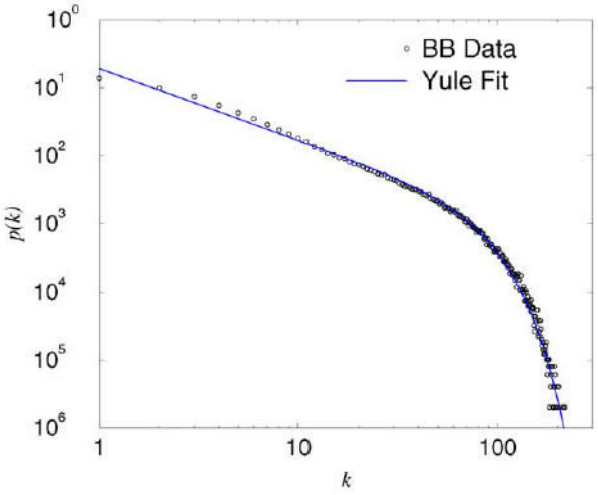

Fig. 3.

The fit of the Yule distribution to the BB model. The distribution function $p(k)$ is fitted by $p(k) = 0.19k^{-1.05}(0.99986)^{k^2}$.

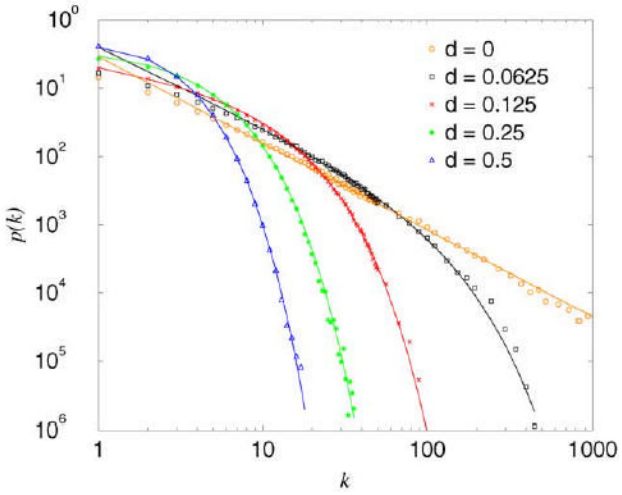

Fig. 4.

Simulations of increasing evolutionary drift in the modified BB model and their approximate fit with a family of Yule distributions (parameters tabulated in Table 1). For $k > 50$ the data points are binned with increasing bin size.

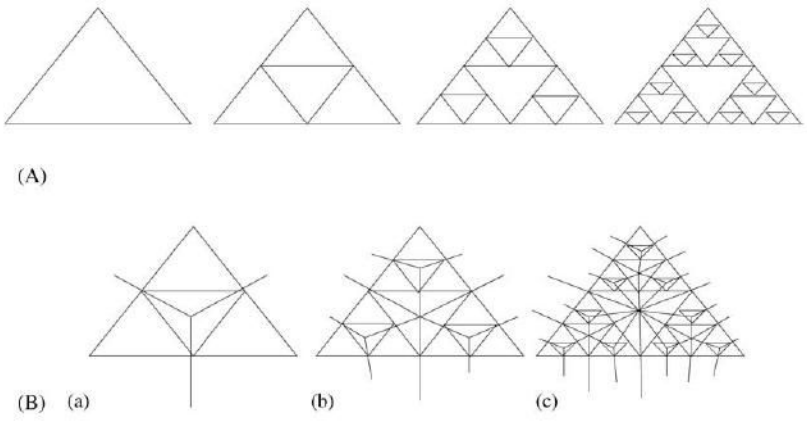

Fig. 5.

(A) Iterative construction of Sierpinski's triangle. (B) Evolution of the network corresponding to the Sierpinski's triangle construction. (a) $s = 1$; 2 vertices of degree 3 (one on the outer face, not drawn); (b) $s = 2$; 2 vertices of degree 3×2; 3 vertices of degree 3; and (c) $s = 3$; 2 vertices of degree 3×2^2; 3^2 vertices of degree 3×2; 3^3 vertices of degree 3.

Table 1

The relation between the parameter *d* in the model and parameters of the Yule family graphed in Fig. 4

d	0	0.0625	0.125	0.25	0.5
a	0.29	0.25	0.22	0.42	0.9
γ	−1.27	−1.0	−0.4	0.075	0.6
ε	0.0	0.00125	0.1	0.3	0.56