Divergent Evolutionary Drift contradicts Power Law

Teresa Przytycka[1] *and Yi-Kuo Yu*[2]

[1]*National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA and* [2]*National Center for Biotechnology Information, NLM, NIH and 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 some insight into the evolution of the network. Here, we re-investigate the probability distributions and scaling properties underlying some models for biological networks and protein domain evolution and point out possible traps in applying a scale free framework to such data. In particular, we demonstrate that divergent evolutionary drift, which is plausible evolutionary mechanisms, is not compatible with scale free models.

Contact: przytyck@ncbi.nlm.nih.gov

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. In this fashion, one can describe protein interaction networks and metabolic networks (Jeong *et al.* (2001); Barabasi *et al.* (1999); Jeong *et al.* (2000); Ravasz *et al.* (2002); Fell *et al.* (2000); Jeong *et al.* (2001)) and the co-occurrence of protein domains within proteins (Apic *et al.* (2002); Wuchty (2001).) Aside from considering the mutual interactions among different components, one may also consider the similarities among them (Dokholyan *et al.* (2002)). The growing acquisition rate of biological data has made it possible to ask whether there exists a characteristic that is shared by different biological networks (Alm *et al.* (2003).) A number of studies suggest that the scale free properties is the shared characteristic (Barabasi *et al.* (1999); Gisiger *et al.* (2001); Wolf *et al.* (2002).)

To describe the concept of 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. 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. Or more precisely,

$$\frac{p(k_1)}{p(k_2)} = \frac{p(\alpha k_1)}{p(\alpha k_2)} = F\left(\frac{k_1}{k_2}\right) \qquad (1)$$

where α is a positive constant and F is called the scaling function. 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." It is not hard to see that when the scale free property (1) is satisfied, the probability function $p(k)$ must follow the power-law, i.e. $p(k) \sim k^{-\gamma}$. 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 $-\gamma$. In particular, a scale free network will have a small number of highly connected vertices (hubs) and large number of low degree vertices.

In pursuit of further understanding of the relation between biological networks and scale free networks, researchers began to propose formal evolutionary models (Rzhetsky *et al.* (2001); Qian *et al.* (2001); Karev *et al.* (2002)) that explain the data. Obviously, such models are necessarily gross simplifications of the evolutionary processes but nevertheless provide an important test of possible evolutionary mechanisms.

COMPARISON OF TWO EVOLUTIONARY MODELS

In a recent paper, Dokholyan *et al.* (2002) analyzed a protein domain fold similarity network. The nodes of this network are protein domains connected with an edge if they share significant structural similarity. 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 $k = 1, \ldots, k_{max} \sim 70$ (compare Figure 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, which also fits the data quite well but generates a network which follows scale free characteristics more closely.

The BB model of Dokholyan *et. al.* assumes that the protein universe evolved from one (or a small number of) domain(s) by divergent evolution (Dokholyan *et al.* (2002).) The evolution proceeds in time steps in each of which the following actions are taken:

1. Choose a random node and duplicate it.
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 smaller than some threshold value w, then the new node and its parent are considered to be similar and are connected by an edge, otherwise 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 to model evolutionary drift. If the distance between any pair exceeds the threshold w, and the edge is removed from the network.

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.
4. With probability p (the parameter in the simulation is set to 10^{-2}) the new node will be subject to convergent evolution:
   (a) If $x < w$, one neighbor of the duplicated node is picked with probability proportional to the degree of this neighbor. The distance between this 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.
   (b) 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. The new distance is set using ultrametric condition as in step 3.

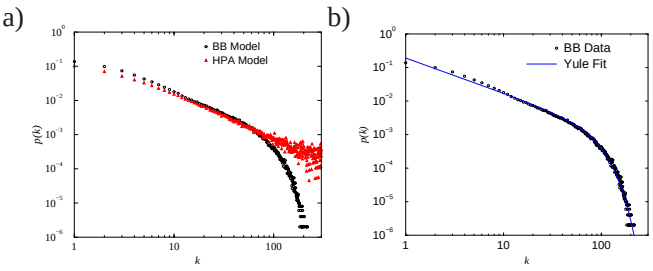

Fig. 1. a) The double log graphs for $p(k)$ BB and HPA model b) The fit of the Yule distribution to the BB Model ($p(k) = 0.19k^{-1.05}(0.99986)^{k^2}$.)

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 Figure 1a). Note that for k in the range $(1, 100)$, the two methods generate a very similar distribution. Thus the statistics collected from real data (where $k_{max} \sim 70$) do not suffice to prove either of the proposed models incorrect.

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 non-zero $p(k)$ value is at $k = 305$. With 5000 iterations repeated 1000 times we would expect to see a non-zero tail further on. The last non-zero value with the HPA model is at around $k = 1200$ which makes it a better candidate for a scale free model.

DIVERGENT DRIFT DESTROYS SCALE-FREE PHENOMENA

It has been argued before (Borodovsky *et al.* (1989); Martindale*et al.* (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 to the BB

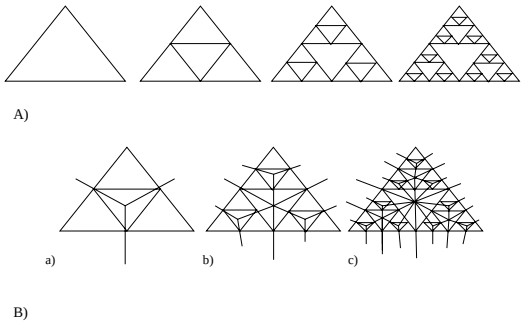

Fig. 2. (A) Iterative construction of Sierpinskis triangle. (B) Evolution of the network corresponding to the Sierpinskis triangle construction.

model (see Figure 1 b). It is worthwhile to examine which property of the BB model is most likely to contribute to such behavior. We claim that the prime contributor to the Yule-like distribution is the divergent evolutionary drift that is present in the BB model but absent from our HPA model.

To have a closer look at the effect of addition of such divergent drift on otherwise scale free construction consider a scale free network which is a direct translation of a fractal known as Sierpinski's triangle. Namely 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 Figure 2 (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 2 vertices of degree 3×2^s; 3^i vertices of degree $3 \times 2^{s-i}$ where $i = 1, \ldots s$ and thus $\sim \frac{1}{2}3^{s+1}$ vertices total. Therefore,

$$\frac{p(k_1 = 3 \times 2^{s-i})}{p(k_2 = 3 \times 2^{s-j})} = \frac{3^i}{3^j} = \left[\frac{k_1}{k_2}\right]^{-\ln 3/\ln 2} . \quad (2)$$

Thus construction is scale free with exponent $\gamma = \ln 3/\ln 2$.

Now assume that edges introduced at a given step are lost after some number of, say 10, steps 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 $\sim \frac{1}{2}3^{11} \sim 10^5$ nodes for the drift to make an impact and several orders of magnitude more nodes to actually observe this statistically.

Thus indeed, the divergent drift is not compatible with scale free phenomena. Therefore, although it is fashionable to draw a straight line through the log-log plot whenever data admits it, one needs to exercise caution in interpreting such graphing as a scale free property of a biological system.

REFERENCES

Alm,E. and Arkin,A.-P. (2003) Biological Networks. *Current Opinion in Structural Biology*, **13**, 193–202.

Apic,Gordana, Hubber,Wolfgang and Teichmann,Sarah (2001) Multi-Domain proteins and domain pairs: comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomic*, **4**, 67–78.

Barabasi,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Borodovsky,M.Y. and Gusein-Zade,S.M. (1989) A general rule for ranged series of codonfrequencies in different genomes. *J. Biomol Struct Dyn.*, **6**, 1001–1012.

Dokholyan,N.V., DeLisi,C., Shakhnovich,B. and Shakhnovich,E.I. (2002) Expanding Protein Universeand its Origin from the Biological Big Bang. *Proc. Natl. Acad. Sci.*, **99**, 14132–14136.

Fell,A.D. and Wagner,A. (2000) The small world of metabolism. *Nature Biotechnology*, **18**, 1121–1122.

Gisiger,T. (2001) Scale Invariance in Biology: Coincidenceor Footprint of a Universal Mechanism?. *Biol. Rev.*, **76**, 161–209.

Jeong,H., Tombor,B., Albert,R., Oltavi,Z.N. and Barabasi,A.-L. (2000) The Large-Scale Organization of Metabolic Networks. *Nature*, **5**, 651–651.

Jeong,H., Mason,B.S., Barabasi,A.-L. and Oltavi,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jeong,H., Oltavi,Z.N. and Barabasi,A.-L. (2003) Prediction of Protein Essentiality Based on Genomic Data.. *ComPelxUS*, **1**, 19–28.

Karev,G., Wolf,Y.I., Rzhetsky,A.Y., Berezovskaya,S.F and Koonin,E.V. (2002) Birth and Death of ProteinDomains: A Simple Model of Evolution Explains Power Law Behavior.. *BMC Evolutionary Biology*, **2**, 18.

Martindale,C. and Konopka,A.K. (1996) Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers Chem.*, **20**, 35–38.

Qian,J., Luscombe,N.M. and Gerstein,M. (2001) Protein Family and Fold Occurrences in Genomes:Power-lawBehavior and Evolutionary Model. *J. Mol. Biol.*, **313**, 673–681.

Ravasz,E., Somera,A.L, Nonfru,D., Oltavi,Z.N. and Barabasi,A.-L. (2002) Hierarchical Organization of Modularity in Metabolic Networks. *Science*, **297**, 1551–1555.

Rzhetsky,A. and Gomez,S. (2001) Birth of scale-free molecular networks and the number of distinctDNA proteins domains ore genome. *Bioinformatics*, **17**, 988–996.

Wolf,Y.I, Karev,G. and Koonin,E.V. (2002) Scale-free networks in biology: new insights into the fundamentals ofevolution?. *BioEssays*, **24**, 105–109.

Wuchty,S. (2001) Scale-Free Behavior in Protein Domain Networks. *Mol. Biol. Evol.*, **18**, 1694–1702.