JB Minireview—Bioinformatics & Systems Biology

J. Biochem. **138**, 1–4 (2005)
DOI: 10.1093/jb/mvi094

Scale-Freeness and Biological Networks

Masanori Arita*

Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo; Institute for Advanced Biosciences, Keio University; and Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency

Received May 10, 2005; accepted May 11, 2005

The notion of scale-freeness and its prevalence in both natural and artificial networks have recently attracted much attention. The concept of scale-freeness is enthusiastically applied to almost any conceivable network, usually with affirmative conclusions. Well-known scale-free examples include the internet, electric lines among power plants, the co-starring of movie actors, the co-authorship of researchers, food webs, and neural, protein–protein interactional, genetic, and metabolic networks. The purpose of this review is to clarify the relationship between scale-freeness and power-law distribution, and to assess critically the previous related works, especially on biological networks. In addition, I will focus on the close relationship between power-law distribution and lognormal distribution to show that power-law distribution is not a special characteristic of natural selection.

Key words: lognormal, network, power-law, scale-free.

What is "scale-free"?

Despite the multitude of "scale-free" studies (*1–4*), the meaning of "scale-free network" has never been decided precisely (*5*). Literally, the expression refers to the topological invariance of a network structure, no matter how coarsely it is viewed (*6*). More specifically, a scale-free network must have highly connected nodes (*i.e.*, "hubs") at its center, and these hubs must include global hubs, irrespective of the scale of the observed network. For example, the web pages of internet stores are linked with pages of their regular customers, which can be considered (local) hubs. Among such hubs, giant stores like Amazon and Yahoo function as global hubs; they are linked with customer pages and local hubs. Intuitively, a scale-free network exhibits such a self-similar structure no matter whether it is viewed globally or locally. This topological perspective (*i.e.*, self-similarity) has often been overlooked in previous "scale-free" analyses, presumably because in their seminal work, the originators of the scale-free phenomenon, Barabási and Albert, used the word "scale-free" to refer to a network with a power-law degree distribution (*7*).

Power-law degree distribution

In a network, the number of links at each node is called its "degree." For example, we can say "Amazon and Yahoo are high-degree nodes on the internet." Let us consider a distribution of degrees for the entire network. A lattice shows a uniform degree distribution, as all nodes have the same degree. A traditional random network (the Erdös-Rényi model) exhibits a Poisson-like degree distribution (*8*). In essence, the current "scale-free" fever is attributable to the observation that many naturally arising networks show a power-law distribution rather than a Poisson, Gaussian, or uniform distribution.

The power-law distribution states that the probability p of an event is an inverse power of its value x, *i.e.*, $p \sim x^{-\gamma}$ (γ: constant; "~" denotes "proportional to"). In the mid 20th century, the distribution was made famous by the work of the linguist George K. Zipf, who reported that the probability p of the xth most frequently used word is inversely proportional to its rank x: $p \sim x^{-\gamma}$ ($\gamma = 1$) (*9*). Zipf's law is confirmed in many natural languages and also holds for city sizes, firm sizes, or firm incomes in different times and places (*10–13*). The reason for its convergence to $\gamma = 1$ is still under debate.

Interestingly, a power-law distribution has a scale-free character: from the definition, the relationship between $\log p$ and $\log x$ (*i.e.*, on a log-log plot) becomes linear, and its slope $-\gamma$ is independent of the scaling of the x-axis. For a proof, let us denote the scaling of x as $x = Ky$ (K: constant). The relationship between $\log p$ and $\log y$ remains linear with a slope $-\gamma$.

$$\log p = -\gamma \log x = -\gamma (\log y + \log K)$$

Note, however, that this scale-free character has nothing to do with network topology, because the power-law distribution itself does not deal with networks, as in Zipf's law. Only when we say "power-law *degree* distribution," do we start dealing with networks. When the degree distribution obeys the power law, from the definition, most nodes have very few links (low degrees), whereas a tiny fraction of nodes (*i.e.*, hubs) have very many links (large degrees). This property alone, however, does not imply the self-similar topology of the network, because we can arbitrarily swap the links without changing the overall degree distribution. In fact, it is possible to construct a self-dissimilar network whose degree distribution obeys the power law (*5*). Then why do many researchers assume that a network is scale-free (*i.e.*, self-similar) when its degree distribution obeys the power-law? At

*For correspondence: Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwanoha 5-1-5 CB05, Kashiwa, 277-8561. Tel: +81-4-7136-3988, Fax: +81-4-7136-4074, E-mail: arita@k.u-tokyo.ac.jp

least a part of the reason comes from the assumption that networks are constructed more or less randomly. When links are randomly shuffled without changing the power-law degree distribution, the resulting network is most likely to exhibit the self-similar property (*5*). Since many research reports in biology use convenient computer simulations where network models are constructed with a random-number generator (*14*, *15*), their simulation-results almost always exhibit self-similar structures.

Why the power law?

Even if what biologists observe is not scale-freeness but power-law distributions only, the initial question remains unanswered: why is the power law so prevalent in naturally occurring networks?

Stating the conclusion first, we have not yet discovered the answer. The natural abundance of the power-law distribution has been investigated for over a century in wide areas of research. In 1881, the astronomer/mathematician Simon Newcomb found that the probability that a number has the leading digit of d is $\log_{10}(1 + 1/d)$ (d = 1,2,…,9); this results if the probability of seeing the number n is P(number = n) ~ $n^{-\gamma}$ ($\gamma \to +1$) (this is known as the law of Frank Benford, the physicist who independently discovered and verified the law in 1938) (*16*, *17*). In 1897, the economist Vilfredo Pareto reported that the number of people with an income equal to x is proportional to its inverse power: P(income = x) ~ $x^{-\gamma}$ (*18*). The current infatuation with the scale-freeness was started by the physicists Albert-László Barabási and Réka Albert, who applied the power-law story to the field of network study (*7*). The mathematical foundation of the Barabási-Albert model for the emergence of scale-free networks (in fact the power-law distribution) is based on the preferential attachment ("rich gets richer") principle, which was originally proposed by the statistician George U. Yule in 1924 and was later elaborated by the cognitive scientist Herbert A. Simon (*19*, *20*). Thus, the story is like folklore that has been retold for over a century.

Lognormal distribution

Theoretically, power-law distribution is closely related with lognormal distribution (*21*). A random variable X is lognormally distributed if $\log(X)$ is normally distributed (*i.e.*, Gaussian distribution) (*22*, *23*). This distribution arises from a random, multiplicative process, i.e., a multiplicative version of a random walk by coin-flipping (*24*). A more elaborate version is also well-known as the Black-Scholes model in economics (*25*). A lognormal distribution may look like a power-law distribution in that its log-log plot can become linear for a wide range of x. The logarithm of the density function of a lognormal distribution is

$$\log f(x) = -\log x - \log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2$$

where μ is the mean and σ the standard deviation. If σ is sufficiently large, the third term will be almost constant for small x, and hence the log-log plot becomes linear (*21*). The central limit theorem states that the sum of many independent, identically distributed random variables shapes a Gaussian distribution. Likewise, the product of many independent, identically distributed, positive random variables shapes a lognormal distribution (*24*).

This observation implies the natural prevalence of lognormal distributions around us. Almost all physical and chemical laws are ruled by multiplication, not by addition (consider any physical law such as $E = MC^2$). What mathematics tells us is that many, independent multiplicative processes will produce, in their limits, lognormal distributions. Indeed, in parallel with the Gaussian distribution, the lognormal distribution has long been used as an appropriate model for skewed distributions (*22*). The law of proportional effect, which produces a lognormal distribution, was first proposed by the astronomer Jacobus C. Kapteyn in 1903 and established as early as 1931 by the economist Robert Gibrat (*26*, *27*). Thus, while the lognormal distribution should have become as familiar as the Gaussian distribution, it disappeared from the standard curriculum of statistics, perhaps because of the misconception that addition is easier than multiplication.

Assuming that lognormal distributions are prevalent in nature, it is easy to understand the prevalence of power-law distributions. Power-law distributions may arise from lognormal distributions upon small tweaks, for example, when the data-sampling time is not uniform (*28*), or when a lower boundary is put into effect during a random walk (*29*). The well-known preferential attachment is regarded as a variant of the latter model (*30*). In general, our observation of natural phenomena is rarely unbiased. Therefore, it is not surprising that biases in data acquisition make original multiplicative processes look power-law–like (*28*).

Application to biological networks

Considering the surprisingly diverse appearance of both power-law and lognormal distributions, it is more pragmatic to choose a distribution that produces useful results depending on the interpretation of networks. It is futile to focus on one distribution as a universal, natural principle. The marked difference between the two distributions is that power-law distributions contain many more large elements than lognormal distributions, a property is often referred to as "heavy-tailed." Consequently, the variance is infinite in power-law and finite in lognormal distributions. As long as we deal with finite data, infinite variance is unattainable. At the same time, real-world data usually contain more large elements than expected from lognormal distributions (*22*, *28*, *31*), and we are torn between the available choices. An easier way out is lognormal distributions, simply because they become Gaussian distributions after their logarithm is taken. For example, almost all microarray analyses use statistical tests after taking the logarithm of raw gene-expression data (*32*), because lognormal distributions are implicitly assumed for gene expressions (note that almost all statistical tests assume Gaussian as their background distributions). If we were to use power-law distributions instead, we would need to invent necessary statistical tools accordingly.

Topology is not enough

So far, our discussion has dealt with the degree distribution. In reality, a more important factor is biochemical function of the network, including dynamics; this also

has been overlooked in previous "scale-free" analyses. It is well known in biology that protein-protein, genetic, and metabolic networks are highly constrained by different biological mechanisms; we cannot straightforwardly conclude the biological importance of hubs only from the network topology found by comprehensive interaction data or shared amino acid sequences (*33*, *34*).

For example, metabolic networks exhibit different network properties depending on the interpretation of topology. Their scale-freeness (in fact, power-law degree distribution) may be observed if "information transfer" or "transmission degree of perturbation" between metabolites is under focus (*35–38*). If biochemical, structural conversion is considered, which is what biologists usually do, metabolic pathways are much more constrained than the underlying network structure and not scale-free (*39–41*). Since the functioning pathways are the subset of connected routes in the underlying network, the hubs also vary, depending on how we view the networks. Pyruvate, acetyl CoA, and ATP become hubs if the degree of structural changes is under focus. On the other hand, water, phosphates, and NAD become hubs if the number of their occurrences is compared. Such discrimination requires a detailed analysis of molecular structures (*40*, *42*), and similar efforts would be required to assess the functioning network of proteins or genes.

Conclusion

If the biological meaning of networks is overlooked, discussion on the degree distribution of biological networks is futile. Since the emergence of power-law distribution is not special, care must be taken in assessing the biological implication of the "scale-free" conclusions. It is arguable whether hubs are biologically important or evolutionarily ancient. This short review could not cover other interesting properties of scale-free networks such as "small world," "robustness under random failure" or "preservation under random rewiring." For further reading, we recommend the article by Mitzenmacher (*21*), which describes the research history of power-law and lognormal distributions, and the review by Li *et al.* (*5*), which critically overviews the current understanding on scale-freeness and proposes its mathematically rigorous definition.

This research is partly supported by MEXT, Grant-in-Aid for Scientific Research on Priority Areas "Genome information science." The author thanks Dr. Reiko Tanaka for careful reading of the manuscript and useful suggestions, and Ursula Petralia for editing this manuscript.

REFERENCES

1. Strogatz, S.H. (2001) Exploring complex network. *Nature* **410**, 268–276
2. Barabási, A.-L. and Bonabeau, E. (2003) Scale-free networks. *Sci. Am.* **288**, 60–69
3. Barabási, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113
4. Sharom, J.R., Bellows, D.S., and Tyers, M. (2004) From large networks to small molecules. *Curr. Opin. Chem. Biol.* **8**, 81–90
5. Li, L., Alderson, D., Tanaka, R., Doyle, J.C., and Willinger, W. (2005) Towards a theory of scale-free graphs: definition, properties, and implications. *arXiv:cond-mat/0501169*
6. Song, C., Havlin, S., and Makse, H.A. (2005) Self-similarity of complex networks. *Nature* **433**, 392–395
7. Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* **286**, 509–512
8. Erdös, P. and Rényi, A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61
9. Zipf, G.K. (1949) *Human Behavior and the Principle of Least Effort*, Addison-Wesley
10. Carroll, G.R. (1982) National city size distributions: what do we know after 67 years of research? *Prog. Hum. Geog.* **6**, 1–43
11. Gabaix, X. (1999) Zipf's law for cities: an explanation. *Q. J. Economics* **114**, 739–767
12. Axtell, R.L. (2001) Zipf distribution of U.S. firm sizes. *Science* **293**, 1818–1820
13. Li, W. (2005) Information on Zipf's law. Available at http://www.nslij-genetics.org/wli/zipf/
14. van Noort, V., Snel, B., and Huynen M.A. (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* **5**, 280–284
15. Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I., and Koonin, E.V. (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* **21**, 2058–2070
16. Newcomb, S. (1881) Note on the frequency of the use of digits in natural numbers. *Amer. J. Math.* **4**, 39–40
17. Benford, F. (1938) The law of anomalous numbers. *Proc. Am. Phil. Soc.* **78**, 551–572
18. Pareto, V. (1896, 1897) Cours d'economie politique. Reprinted as a volume of Oeuvres Complètes (Droz, Geneva, 1896, 1965). Pareto, V. Cours d'Economique Politique (Macmillan, Paris, 1897)
19. Yule, G. (1924) A mathematical theory of evolution, based on the conclusion of Dr. J.C. Willis. *F.R.S. Phil. Trans. R. Soc. Lond Ser. B* **213**, 21–87
20. Simon, H.A. (1955) On a class of skew distribution functions. *Biometrika* **42**, 425–440
21. Mitzenmacher, M. (2003) A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**, 226–251
22. Aitchison, J. and Brown, J.A.C. (1957) *The Lognormal Distribution*, Cambridge University Press, Cambridge, UK
23. Crow, E.L. and Shimizu, K. (eds.) (1988) *Lognormal Distributions: Theory and Applications*, Marcel Dekker, New York
24. Limpert, E., Stahel, W.A., and Abbt, M. (2001) Lognormal distributions across the sciences: keys and clues. *Bioscience* **51**, 341–352
25. Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities. *J. Political Economics* **81**, 637–654
26. Kapteyn, J.C. (1903) Skew frequency curves in biology and statistics in *Astronomical Laboratory*, Noordhoff, Groningen
27. Gibrat, R. (1931) *Les Inegalites Economiques*, Libraire du Recueil Sirey, Paris
28. Reed, W.J. and Hughes, B.D. (2002) From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature. *Phys. Rev. E* **66**, 067103
29. Kesten, H. (1973) Random difference equations and renewal theory for products of random matrices. *Acta Mathematica* **CXXXI**, 207–248
30. Champernowne, D. (1953) A model of income distribution. *Economic J.* **63**, 318–351
31. Montroll, E.W. and Shlesinger, M.F. (1982) On 1/f noise and other distributions with long tails. *Proc. Natl Acad. Sci. USA* **79**, 3380–3383
32. Baldi, P. and Hatfield G.W. (2002) *DNA Microarrays and Gene Expression*, Cambridge Univ Press, Cambridge, UK
33. Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–42
34. Rao, F. and Caflisch, A. (2004) The protein folding network. *J. Mol. Biol.* **342**, 299–306

35. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000) The large-scale organization of metabolic networks. *Nature* **407**, 651–654
36. Fell, D.A. and Wagner, A. (2000) The small world of metabolism. *Nat. Biotechnol.* **18**, 1121–1122
37. Wagner, A. and Fell, D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. London Ser. B* **268**, 1803–1810
38. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555
39. Ma, H.-W. and Zeng, A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277
40. Arita, M. (2004) The metabolic world of Escherichia coli is not small. *Proc. Natl Acad. Sci. USA* **101**, 1543–1547
41. Tanaka, R. (2005) Scale-rich metabolic networks *Phys. Rev. Lett.* **94**, 168101
42. Rahman, S.A., Advani, P., Schunk, R., Schrader, R., and Schomburg, D. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics* **21**, 1189–1193