Advances in Information Mining, ISSN: 0975–3265, Vol. 1, Issue 1, 2009, pp. 1-4

Distribution Analysis in Metabolic Networks: *Power Laws*

[1]Sangya Pundir and [1*]Somnath Tagore
[1]Department of Biotechnology & Bioinformatics, Dr D Y Patil University, Plot 50, Sec 15, CBD Belapur, Navi Mumbai 400614, India
*Corresponding Author Email: somnathtagore@yahoo.co.in

Abstract: Many new structural patterns have been discovered in diverse biological, social and information networks. One of them are metabolic networks, the most widely studied large scale networks in biology, known to have a power law degree distribution and the exponent γ is observed to be the same for all species. However, empirical evidence elucidating the nature of the process that gives rise to such structure is lacking this far. In this paper we review facts about power law distribution as relevant to metabolic networks. In particular we concentrate on the evolutionary and other implications of such a power law distribution.

Keywords: Power law, degree, rank, metabolic networks, graph theory, log-log scale.

1 INTRODUCTION

Various hypotheses regarding evolutionary mechanisms have been proposed and explored in mathematical models of evolving networks. For growing networks, a 'preferential attachment' of new nodes to higher degree nodes as well as a 'proportionate change' **[1]** mechanism whereby nodes with higher degree experience proportionately higher changes in degree has been proposed to account for the power law degree distribution of the network. The latter process can lead to robust exponents. However, it is not clear whether this hypothesis is applicable to the evolution of metabolic networks **[2]**. For one, the metabolic network is not a growing network; during the course of evolution the number of metabolites has remained in the range of a few hundred to about a thousand for all organisms **[3]**.

Furthermore, so far no concrete evidence has been presented for a preferential attachment or proportionate change process during its evolution **[4]**. Systems Biology looks at networks in an integrative manner and thus might hold the key to unravelling them. It is a new emerging field with more scope for innovation and development **[5]**. Applying statistical and mathematical approaches to the study of metabolic networks is a revolutionary concept **[6]**. Study of Systems biology is essential to simulate metabolic concepts using models.

A power law is an inverse mathematical relationship between two quantities. If one quantity is the frequency of an event, the relationship is a power-law distribution, and the frequencies **[7]** decrease very slowly as the size of the event increases. A power law is any polynomial relationship that exhibits the property of scale **[8]** invariance. If quantities follow power law then as their event size increases, their frequency of occurrence decreases. A linear relationship is produced when both logs are taken of a quantitative property x and its function f(x). A straight line obtained on the log-log plot shows that the quantities follow a power law and is often called the signature of the power law. The relation can be represented as $f(x) = ax^k$.

Power laws also describe other kinds of relationships, such as the metabolic rate of a species and its body mass, and the size of a city and the number of patents it produces and the diameter of the internet **[9]** etc. What this relationship means is that there is no typical size in the conventional sense **[10]**. Power laws are found throughout the natural and manmade worlds, and are an active area of scientific research.

2 TECHNICAL DEFINITION

A power law is any polynomial relationship that exhibits the property of scale invariance. The most common power laws relate two variables and have the form

$$f(x) = ax^k + o(x^k)$$

where a and k are constants, and $o(x^k)$ is an asymptotically small function of x. Here, k is typically called the scaling exponent, where the word "scaling" denotes the fact that a power-law function satisfies

$$f(cx) \propto f(x)$$

where c is a constant. Thus, a rescaling of the function's argument changes the constant of proportionality but preserves the shape of the function itself. This point becomes clearer if we take the logarithm of both sides:

$$\log(f(x)) = k \log x + \log a$$

Notice that this expression has the form of a linear relationship with slope k. Re-scaling the argument produces a linear shift of the function up or down but leaves both the basic form and the slope k unchanged. Power-law relations characterize a staggering number of naturally occurring phenomena, and this is one of the principal reasons why they have attracted such wide interest. For instance, inverse-square laws, such as gravitation and the Coulomb force, are power laws, as are many common mathematical formulae such as the quadratic law of area of the circle **[11]**. However much of the recent interest in power laws comes from the study of probability distributions: it's now known that the distributions of a wide variety of quantities seem to follow the power-law form, at least in their upper tail (large events) **[12]**. The behavior of these large events connects these quantities to the study of theory of large deviations, which considers the frequency of extremely rare events like stock market crashes and large natural disasters. It is primarily in the study of statistical distributions that the name "power law" is used; in other areas the power-law functional form is more often referred to simply as a polynomial form or polynomial function.

Scientific interest in power law relations stems partly from the ease with which certain general classes of mechanisms generate them. The demonstration of a power-law relation in some data can point to specific kinds of mechanisms that might underlie the natural phenomenon in question, and can indicate a deep connection with other, seemingly unrelated systems. The ubiquity of power-law

relations in physics is partly due to dimensional constraints, while in complex systems **[13]**, power laws are often thought to be signatures of hierarchy or of specific stochastic processes. A few notable examples of power laws are the Gutenberg-Richter law for earthquake sizes, Pareto's law of income distribution, structural self-similarity of fractals, and scaling **[14]** laws in biological systems. Research on the origins of power-law relations, and efforts to observe and validate them in the real world, is an active topic of research in many fields of science, including physics, computer science, linguistics, geophysics, sociology, economics and more.

3 EXAMPLE OF POWER LAW DISTRIBUTION

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a power law probability distribution that coincides with social, scientific, geophysical, actuarial, and many other types of observable phenomena. Outside the field of economics it is at times referred to as the Bradford distribution. Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. This idea is sometimes expressed more simply as the Pareto principle or the "80-20 rule" which says that 20% of the population controls 80% of the wealth. It can be seen from the probability density function (PDF) graph on the right, that the "probability" or fraction of the population that owns a small amount of wealth per person is rather high, and then decreases steadily as wealth increases **[15]**. This distribution is not limited to describing wealth or income distribution, but to many situations in which an equilibrium is found in the distribution of the "small" to the "large".

A great many power-law distributions have been conjectured in recent years. For instance, power laws are thought to characterize the behavior of the upper tails for the popularity of websites, number of species per genus, the popularity of given names, the size of financial returns, and many others. However, much debate remains as to which of these tails are actually power-law distributed and which are not **[16]**. For instance, it is commonly accepted now that the famous Gutenberg-Richter Law decays more rapidly than a pure power-law tail because of a finite exponential cutoff in the upper tail.

4 PREDICTION TECHNIQUES

Power law distribution is detected based on a property of the metabolic networks such as the degree of the metabolites **[17]**. The degree of metabolites is determined by the connectivity of the metabolites **[18]**. It is the number of edges passing through the node which is in other words the number of connections of the metabolite. The next element required is a function of the property such as the rank of the metabolites in this case it is the cumulative frequency distribution function of the degree of the metabolites **[19]**. The metabolite(s) with the highest degree is ranked first and so on.

Thus the ascending order of ranks represents the descending order of metabolite degree. Once the rank and degree have been determined, the power law graph is plotted. The graph is plotted on the log-log scale with the quantitative property i.e the metabolite degree on the X axis and its function i.e the metabolite rank on the Y axis. If an inverse straight line is obtained, then power law distribution is being followed. Assuming that the degree of a metabolite can be described by a random variable D, plotting data estimates the counter-cumulative probability function P(log D > k) **[20]**.

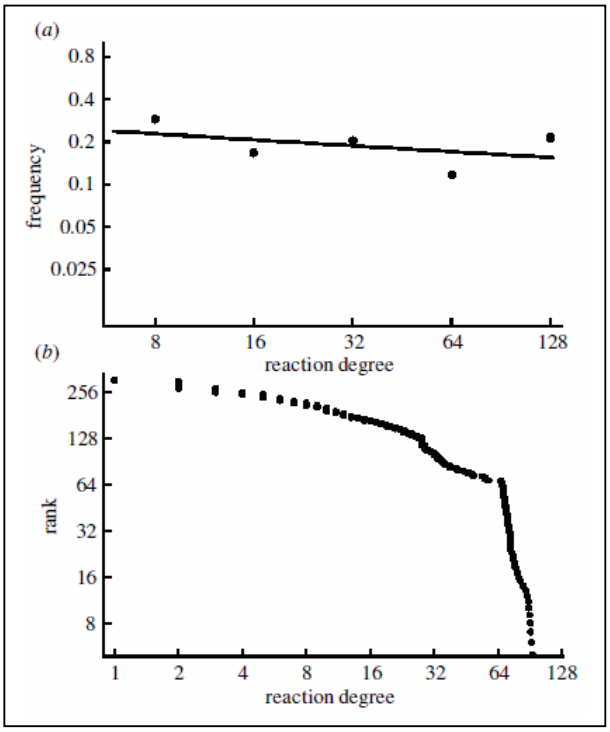

Figure 1: Power law distribution graphs with various possible property selections for metabolites

5 APPLICATIONS

Mathematically, the behaviour of a metabolic network can be captured as a system of ordinary differential equations in the metabolite concentrations **[21]**. Power law plotting shows the frequency of occurrence of entities in a given pathway with respect to their occurrence **[22]**. Thus we can determine which metabolites play a major role in the network and need to be targeted in order to alter the network. This suggests that the highly connected metabolites linking the individual pathways into a connected network are responsible for the great variance in degree **[23]**. Their high connectivity provides the `glue' of the network and is also responsible for the short path length **[24, 25]**. This approach is illustrated by a partial reconstruction of a model for the "breakthrough organism," the last organism to use RNA as the sole genetically encoded biological catalyst **[26]**.

Many of these models make a key prediction: Highly connected nodes are old nodes, nodes having been added very early in a network's history thus reflecting evolution **[27]** of metabolic pathways. Following the power law distribution also means that as

the rate of the reaction of metabolites would increase, its occurrence would decrease **[28, 29]**. Conventionally, non-linear curve-fitting algorithms have been used for modeling **[30]** but power law modeling is a compact way of representing pathways and reduces data redundancy and redundant edges. Metabolic networks have a history. They have not been assembled in their present state at once. They have grown, perhaps over billion years, as organisms increased their metabolic and biosynthetic abilities **[31]**. Having to take into account this history raises many questions **[32]**.

One of the questions to be asked here is how a network arrives at a power law degree distribution if it grows. The perhaps simplest mathematical model capable of growing power-law distributed networks involves only two simple rules **[33]**. First and unsurprisingly, it adds nodes to a graph. Second, it connects this node to previously existing nodes according to a second rule, where already highly connected nodes are more likely to receive a new connection than nodes of lesser connectivity. Over many node additions, a power-law degree distribution emerges. A great variety of variations to this model have been proposed **[34]**. They differ greatly in detail but retain in some way or another, the rule that new connections preferably involve highly connected nodes **[35]**. But more importantly, many of these models make a key prediction:

Highly connected nodes are old nodes, nodes having been added very early in a network's history. Since traditional graph theory on regular graph or random graph cannot explain the high variability of degree sequence **[36]**, the discovery of the power-law degree distribution has stimulated a great deal of work in the construction of the so-called "scale-free" networks **[37]**, aiming to match the power-law distribution and other large scale statistical properties, as well as to provide a universal theory to understand all complex networks **[38]**. Thus we can see the implication of detecting a power law distribution in determining the evolutionary **[39]** path of the metabolic network.

6 CURRENT STRATEGIES

The best way to detect power law distribution is to plot the properties on a log-log graph and see whether a straight line is obtained **[40-44]**. A property based on connectivity such as the degree must be plotted on the x axis and a property obtained by its cumulative distribution function **[45]** such as the rank is plotted on the y axis. Many selections of properties are possible such as reaction degree and rank etc **[46]**.

The selection of log-log scale compresses the graph and makes it possible to plot a large number of points by emphasizing the tail region **[47-50]**. In a way it is like looking at the bigger picture in the graph. The initial points with less difference between their values on the axis can be seen to be farther apart while the points with greater difference between their values are placed together towards the end of the axis. A similar effect can be achieved by taking log to the base ten of all the values and plotting them on a regular graph **[51-52]**.

7 CONCLUSION

Power law is a robust and useful analysis method for metabolic networks. A power law graph can be easily plotted to detect whether the given metabolites follow a power law. If the metabolites do follow power law, estimation about the more essential and less essential metabolites can be made. Power law is a compact presentation of a large amount of metabolic data. The linear correlation described above is an overall classification of the evolutionary process. A deeper understanding would require going into the mechanisms by which such a correlation comes about, as well as into the departures from the statistical pattern. Metabolic network evolution ultimately rests on mechanisms of enzyme structure evolution, which in turn involves the molecular evolution of genes that code for the enzymes. The latter is governed both by random processes as well as the forces of selection.

The following random processes that are biologically plausible can in principle give rise to a proportionate change in metabolic networks: A metabolite with high degree gene corresponding to one of these enzymes mutates in a manner that disturbs the binding site of this metabolite on the enzyme, the corresponding reaction could be lost. The more enzymes the metabolite binds to, the proportionately higher is the probability of losing its reactions through random mutations. On the other hand if the gene duplicates and diverges, that can introduce a new enzyme to which the metabolite binds and hence a new reaction for it to participate in. Large degree metabolites have a larger pool of interacting enzymes whose genes can duplicate, and hence if genes duplicate randomly, the number of new reactions a given metabolite participates in is also expected to be positively correlated with its degree.

Power laws provide robustness against perturbations. Upon removal of randomly chosen nodes, the mean distance between network nodes that can still be reached from each other (via a path of edges) increases only very little. This distance is also known as the network diameter. Also, graphs with power-law degree distributions fragment less easily into large disconnected sub-networks upon random node removal.

REFERENCES

[1] Fell, DA, Wagner, A. (2000) The small world of metabolism, *Nat Biotechnol*, **18(11)**, 11212.

[2] Fersht, A. (1985) Enzyme Structure and Mechanism, 2nd edit, W H Freeman & Co.

[3] Albert, R., Jeong, H., Barabasi, AL. (1999) Internet: diameter of the world-wide web, *Nature*, **401**, 130-131.

[4] Bak, P. (1990) Self-organized criticality, *Physica A*, **163**, 403-409.

[5] Barabasi, AL, Albert, R. (1999) Emergence of scaling in random networks, *Science*, **286**, 509-512.

[6] Benner, SA, Ellington, AD, Tauer, A. (1989) Modern metabolism as a palimpsest of the RNA world, *Proc. Natl Acad. Sci. USA*, **86**, 7054-7058.

[7] Cascante, M., Melendez-Hevia, E., Kholodenko, BN, Sicilia, J, Kacser, H. (1995) Control analysis of transit-time for free and enzyme-bound metabolites physiological and evolutionary significance of metabolic response-times, *Biochem. J*, **308**, 895-899.

[8] Chiva, E., Tarroux, P. (1995) Evolution of biological regulation networks under complex environmental constraints, *Biol. Cybernet*, **73**, 323-333.

[9] Babu, M. (2004) Structure and evolution of transcriptional regulatory networks, *Current Opinion in Structural Biology*, **14**, 283–291.

[10] Bar-Joseph, Z. (2003) Computational discovery of gene modules and regulatory networks, *Nature Biotechnology*, **21(11)**, 1337–1342.

[11] Barabasi, L., Albert, R. (1999) Emergence of scaling in random networks, *Science*, **286**, 509–512.
[12] Barabasi, L., Oltvai, Z. (2004) Network biology: understanding the cell's functional organization, *Nature Reviews - Genetics*, **5**, 101–113.
[13] Wagner, A., Fell, DA (2001) The small world inside large metabolic networks, *Proc Biol Sci*., **268(1478)**, 1803-1810.
[14] Barthelemy, M. et al. (2005) Dynamical patterns of epidemic outbreaks in complex heterogeneous networks, *Journal of Theoretical Biology*, **235**, 275–288.
[15] Becker, N. et al. (2005) Controlling emerging infectious diseases like SARS, *Mathematical Biosciences*, **193**, 205–221.
[16] Veflingstad, SR, Almeida, J., Voit, EO (2004) Priming nonlinear searches for pathway identification, *Theor Biol Med Model*, **1**, 8.
[17] Voit, EO, Almeida, J. (2004) Decoupling dynamical systems for pathway identification from metabolic profiles, *Bioinformatics*, **20**, 1670–1681.
[18] Di Camillo, B., Toffolo, G., Cobelli, C. (2009) A gene network simulator to assess reverse engineering algorithms, *Ann N Y Acad Sci*., **1158,** 125-142.
[19] Seki, K., Mostafa, J. (2009) Discovering implicit associations among critical biological entities, *Int J Data Min Bioinform*, **3(2)**, 105-23.
[20] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., Tomita, M. (2003) Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatics*, **19**, 643–650.
[21] Kimura S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama S., Kuramitsu, S., Konagaya, A. (2005) Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm, *Bioinformatics*, **21**, 1154–1163.
[22] Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., Eguchi, Y. (2001) Development of a system for the inference of large scale genetic networks, *Pac Symp Biocomput*., 446-458.
[23] Diaz-Sierra, R., Fairen, V. (2001) Simplified method for the computation of parameters of power-law rate equations from time-series, *Math Biosci*, **171** ,1–19.
[24] Sorribas, A., Lozano, JB, Fairen, V. (1998) Deriving chemical and biochemical model networks from experimental measurements, *Recent Res Devel in Physical Chem*, **2**, 553–573.
[25] Diaz-Sierra, R., Lozano, JB, Fairen, V. (1999) Deduction of chemical mechanisms from the linear response around steady state, *J Phys Chem A*, **103**, 337–343.
[26] Savageau, MA (1976) Biochemical systems analysis: A study of function and design in molecular biology, Addison-Wesley Pub. Co.
[27] De Atauri, P., Orrell, D., Ramsey, S., Bolouri, H. (2005) Is the regulation of galactose 1-phosphate tuned against gene expression noise? *Biochem J*, **387**, 77–84.
[28] Schmidt, H., Cho, KH, Jacobsen, EW (2005) Identification of small scale biochemical networks based on general type system perturbations, *Febs J*, **272**, 2141–2151.
[29] Zhao, J., Shimizu, K. (2003) Metabolic flux analysis of Escherichia coli K12 grown on 13C-labeled acetate and glucose using GC-MS and powerful flux calculation method, *J Biotechnol*, **101**, 101–117.
[30] Fischer, E., Sauer, U. (2003) Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS, *Eur J Biochem*, **270**, 880–891.
[31] Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, RN, Willmitzer, L. (2000) Metabolite profiling for plant functional genomics, *Nat Biotechnol*, **18**, 1157–1161.
[32] Price, DJ. (1965) Networks of science papers, *Science*, **149**, 510–515.
[33] Samorodnitsky, G., Taqqu, MS (1994) Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance, Chapman and Hall, New York - London.
[34] Serrano, MA, Boguna, M. (2005) Tuning clustering in random networks with arbitrary degree distributions, *Phys Rev E Stat Nonlin Soft Matter Phys*., **72(3 Pt 2)**, 036133.
[35] Serrano, MA, Boguna, M. (2006) Percolation and epidemic thresholds in clustered networks, *Phys Rev Lett.*, **97(8)**, 088701.
[36] Simon, HA (1955) On a class of skew distribution functions, *Biometrika*, **42**, 425–440.
[37] Spring, N., Mahajan, R., Wetherall, D., Anderson, T. (2004) Measuring ISP Topologies with Rocketfuel, *IEEE/ACM Transactions on Networking*, **12**, 2–16.
[38] De Martino, D., Dall'asta, L., Bianconi, G., Marsili, M. (2009) Congestion phenomena on complex networks, *Phys Rev E Stat Nonlin Soft Matter Phys*., **79(1 Pt 2)**, 015101.
[39] Scotti, S., Mauri, M., Barbieri, R., Jawad, B., Cerutti, S., Mainardi, L., Brown, EN, Villamira, MA. (2006) Automatic quantitative evaluation of emotions in E-learning applications, *Conf Proc IEEE Eng Med Biol Soc.*, **1**, 1359-1362.
[40] Tanaka, R. (2005) Scale-rich metabolic networks, *Phys Rev Lett.*, **94(16)**, 168101.
[41] Van Lint, JH, Wilson, RM (1992) A Course in Combinatorics, Cambridge University Press.
[42] Viger, F., Latapy, M. (2005) Efficient and simple generation of random simple connected graphs with prescribed degree sequence, *Proceedings of the 11th Conference of Computing & Combinatoric*, 154.
[43] Wang, J., Li, L., Low, S., Doyle, JC (2005) Cross-layer optimization in TCP/IP networks, *IEEE/ACM Transactions on Networking*, 13.
[44] Watts, DJ, Strogatz, SH (1998) Collective dynamics of "small world" networks, *Nature*, 393.
[45] Waxman, BM (1988) Routing of multipoint connections, *IEEE Journal of Selected Areas in Communications*, 6.
[46] Gómez-Gardeñes, J, Latora, V. (2008) Entropy rate of diffusion processes on complex networks, *Phys Rev E Stat Nonlin Soft Matter Phys*., **78(6 Pt 2)**, 065102..
[47] Priesemann, V., Munk, MH, Wibral, M. (2009) Subsampling effects in neuronal avalanche distributions recorded in vivo, *BMC Neurosci*., **10**, 40.
[48] Willinger, W., Govindan, R., Jamin, S., Paxson, V., Shenker, S. (2002) Scaling phenomena in the internet: Critically examining criticality, *Proceedings of the National Academy of Sciences*, **99**, 2573–2580.
[49] Yook, SH, Jeong, H., Barabasi, AL (2002) Modeling the internet's large-scale topology, *Proceedings of the National Academy of Sciences*, **99**, 13382–13386.
[50] Yule, G. (1925) A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, *Philosophical Transactions of the Royal Society of London (Series B)*, **213**, 21–87.
[51] Zegura, E., Calvert, KL, Donahoo, MJ (1997) A quantitative comparison of graphbased models for internet topology, *IEEE/ACM Transactions on Networking*, **5(6)**.
[52] Sanchirico, A., Fiorentino, M. (2008) Scale-free networks as entropy competition, *Phys Rev E Stat Nonlin Soft Matter Phys.*, **78(4 Pt 2)**, 046114.