BIOINFORMATICS DISCOVERY NOTE Vol. 24 no. 6 2008, pages 741–743
doi:10.1093/bioinformatics/btn041

Systems biology

Biological pathway kinetic rate constants are scale-invariant

Scott Grandison and Richard J. Morris*

Department of Computational & Systems Biology, John Innes Centre, Norwich Research Park, Colney Lane, NR4 7UH Norwich, UK

Received on October 3, 2007; revised on January 24, 2008; accepted on January 25, 2008

Advance Access publication January 30, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Scale-free networks have had a profound impact in Biology. Network theory is now used routinely to visualize, navigate through, and help understand gene networks, protein–protein interactions, regulatory networks and metabolic pathways. Here we analyse the numerical rather than topological properties of biological networks and focus on the study of kinetic rate constants within pathways.

Results: We have analysed all current entries in the BioModels database and show that the kinetic rate parameters follow Benford's law closely. The cumulative histogram plot reveals an underlying power-law. This implies that these data are scale-invariant, thus placing biological network topology and their chemistry on an equivalent 'scale-free' power-law foundation.

Contact: Richard.Morris@bbsrc.ac.uk

1 INTRODUCTION

An analysis of the world-wide-web by Barabási and Albert (1999) led to the, perhaps surprising, finding that connectivity in such networks is not uniformly distributed but showed the emergence of 'hubs'. This elegant piece of work and subsequent efforts by Barabási and coworkers and many others on scaling, robustness, power-laws and growth processes enjoyed immediate impact. The mathematical beauty and elegance of networks and graph theory finding renewed interest and delivering often spectacular results from statistical mechanics, gene networks and ecosystems, through to social sciences and politics.

Amaral *et al.* (2000) studied a variety of diverse real-world networks and presented evidence of the occurrence of three classes of small-world networks: (i) scale-free, (ii) broad-scale and (iii) single-scale networks. The latter two have constraints limiting the addition of new nodes and the nature of these constraints influences the emergence of the different classes.

Power-laws have been studied for a long time and although these recent network discoveries are perhaps not as astonishing as sometimes claimed, the research of Barabási and others has been highly influential and dramatically changed the way in which we now view and try to understand complex systems. The topological features of scale-free networks and the remarkable consequences thereof have had a significant impact in Biology, especially on the understanding of protein–protein interactions, regulatory networks and metabolic pathways, and the interplay between parts that has led to the development of Systems Biology.

In this manuscript, we discuss another scale-invariant property that relates directly to the numerical values that one associates with metabolic and regulatory networks. We show that the first-digit distribution of the kinetic constants follows Benford's law and furthermore that a power-law underlies the data.

In 1881, the astronomer Simon Newcomb noticed that the first pages of a book of logarithms showed higher usage than the later pages. Perhaps, his surprising conclusion was that in a table of physical constants, numbers are more likely to begin with a smaller rather than a larger digit. Newcomb proposed a logarithmic law to describe the frequency of the digits. This work was published under the title 'Note on the Frequency of Use of the Different Digits in Natural Numbers', Newcomb (1881). In 1938, this phenomenon was rediscovered by physicist Frank Benford (Benford, 1938), who based his observation similar to Newcomb on usage of logarithm books and came up with the same logarithmic equation for the leading digit distribution,

$$P(d) = \log_{10}(1 + 1/d). \quad (1)$$

$P(d)$ denotes the probability of observing d as the leading digit of a number. The logarithmic base reflects the base of the number system one is working with. This first-digit phenomenon and the resulting counter-intuitive fact that in real life data there is a 30% probability that the first digit is a one which is now known as Benford's law. Benford himself put his observation to the test and gathered many diverse datasets with many thousands of samples to validate his argument. Benford's law can be generalized to a sequence of digits and can be shown to follow the same equation, thus introducing another unexpected effect: the digits in sequences become dependent on the previous digits. For instance, the probability of the first two digits being 1 and 5 is $p(15) = \log_{10}(1 + 1/15) = 0.028$ and the probability for 5 and 5 is $p(55) = 0.008$. So the probability of the second digit being a 5 conditional on the first digit being a 1 is $p(5|1) = p(15)/p(1) = 0.093$, whereas the conditional probability of a 5 given a 2 as the first digit is 0.099. The probability of the second digit being a 1 given the first digit is a 1 is 0.126 and the probability of the second digit being a 1 given the first digit is a 5 is 0.106.

It has been shown that data are scale-invariant if and only if they follow Benford's law (Hill, 1996). Although the rigorous

*To whom correspondence should be addressed.

proof of Benford's law is more involved and was shown for the first time in 1996 (Hill, 1996) over a century after its first observation and 50 years after its rediscovery, it can be understood with intuitive arguments. When estimating a parameter for which we have no prior knowledge, we would expect the distribution of digits to remain the same regardless of the choice of units. The insensitivity to units is, by definition, a scale-invariant property which in turn implies the distribution of digits will vary proportionately with magnitude, i.e. if we have arrived at a number whose first digit is 1, we must increase this number by 100% to change the first digit to a two. The change from 2 to 3, requires only a 50% increase, 3 to 4 a 30% increase, and so on. This introduces a logarithmic scale to the distribution of significant digits.

Scale-free networks follow a power-law in terms of their topology (connectivity) and scale-invariant data follow a power-law with regard to the probability of digits or sequences of digits. In this article, we demonstrate that biological pathway data follow a power-law and are therefore scale-invariant.

2 RESULTS AND DISCUSSION

To evaluate the scale-invariance of enzyme kinetic data, we downloaded all deposited models from the BioModels Database (Le Novére *et al.*, 2006), June version 2007, and extracted all kinetic parameters. Although a number of collections exist for kinetic data, the BioModels database has the advantage of being a high-quality resource of manually-curated and validated models for biological pathways. This version of the BioModels database contains 113 validated pathway entries with a total of 7684 reactions.

We analysed the distribution of first digits of the rate and equilibrium constants, the result of which is plotted in Figure 1. As may be observed, the curve follows Benford's law closely. Whenever Benford's law holds, scaling the distribution will not change the first-digit distribution and the data are said to be scale-invariant.

An insufficient number of large networks have been deposited in the BioModels database to evaluate this behaviour on individual pathways in a statistically sound manner. Nevertheless, this is a probablistic law and as such this behaviour should be approximately valid for individual biological networks. Indeed, an analysis over the larger networks from the BioModels database, Figure 1B, shows that this is the case. The degree with which these individual pathways follow Benford's law varies hugely as one might expect for such small samples, they are, however, clearly non-uniform and display a preference for smaller digits.

For distributions that have sufficient spread in logarithmic space, i.e. vary over several orders of magnitude, we would expect Benford's law to hold and the data to be approximately scale-invariant. A number of first digit distributions derived from other biological databases are given in Table 1. The lowers digits are often more frequent, however, there are also large deviations from Benford's law.

If we had to hazard a guess at estimating a kinetic rate constant, which value or range would we choose and how confident would we be in that estimate? With no further knowledge all we know is that these parameters are positive and

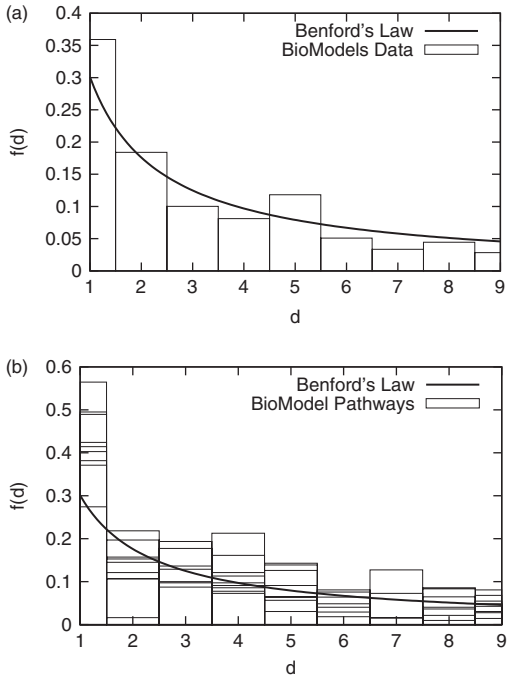

Fig. 1. The frequency, $f(d)$, of the first significant digit, d, of kinetic parameters plotted against the logarithmic law of Equation (1). (**A**) shows all the kinetic data from the BioModels database, whereas (**B**) shows data from large individual pathways (between 62 and 264 rate constants) extracted from the BioModels database.

Table 1. First digit distributions

Digit	Benford	k	δG	E_B	ADP	MW
1	0.301	0.359	0.310	0.129	0.393	0.240
2	0.176	0.184	0.150	0.193	0.247	0.194
3	0.125	0.100	0.103	0.254	0.123	0.164
4	0.097	0.081	0.090	0.208	0.067	0.135
5	0.079	0.118	0.088	0.158	0.041	0.095
6	0.067	0.051	0.072	0.023	0.028	0.061
7	0.058	0.034	0.081	0.020	0.031	0.043
8	0.051	0.044	0.060	0.006	0.034	0.035
9	0.046	0.028	0.046	0.009	0.036	0.032

This table lists the first digit distributions for kinetic constants k (Le Novére *et al.*, 2006), free energies of unfolding δG (Kumar *et al.*, 2006), experimental binding energies E_B (Puvanendrampillai and Mitchell, 2003), and the molecular weights MW of proteins (Wu *et al.*, 2006).

cover many orders of magnitude. In setting up prior distributions, we can include this ignorance about the scale of the parameters. If the scale of a number, x, is not known then it makes sense to require that the prior distribution, $p(x)$, that we assign should not vary upon scaling,

$$p(x)\mathrm{d}x = p(ax)a\mathrm{d}x, \tag{2}$$

in which a is a positive multiplicative scale factor. This implies that $p(x) \propto 1/x$, which results in a improper distribution known as the Jeffreys prior (Jaynes, 2003). Another way of expressing Jeffreys' $1/x$ prior is that the probability over the logarithm of

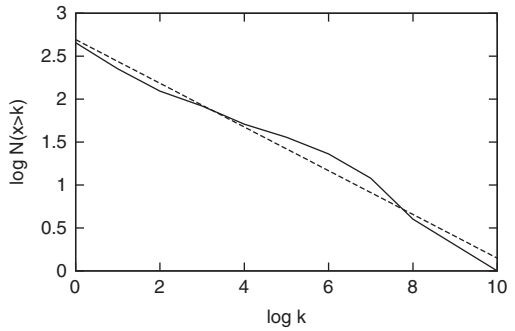

Fig. 2. A cumulative histogram. This logarithmic plot shows the number of data entries, $N(x>k)$, above a given value, k, for the full range of kinetic values. For power-laws this plot should produce a straight line. A least squares fit (shown in dashed) determined the slope to -0.254 ± 0.012.

x is constant, $p(\log x) = \text{const}$. The probability of observing d as the leading digit is therefore proportional to $\log(d+1) - \log(d)$ with a logarithmic base of 10 if we use the decimal number system, $\log_{10}(1 + 1/d)$. Benford's law is thus a consequence of scale-invariance and the process underlying the extraction of the first digit. It can further be shown that scale-invariance implies base-invariance and that these characteristics lead to Benford's law (Hill, 1995). We can conclude that a lack of prior knowledge does not translate into a uniform probability distribution. This result has implications for parameter estimation and sampling strategies.

Rate constants have a finite range, so true scale-invariance in the form of a log-uniform distribution will only ever be approximately realized. A more sensible prior distribution to assume is one which is uniform in logarithmic space with a cutoff value around the diffusion limit or to introduce an exponentially decaying distribution that defines the boundaries. The distribution we observe for the rate constants from the BioModels database resembles a Normal distribution in log space, with a mean at -0.37 and a standard deviation of 2.06. There are other definitions of scale-invariance, the most popular being the requirement that a function of x, $f(x)$, scales with λ^a upon multiplication of x by λ. A power-law is a polynomial function which is scale-invariant, the log-uniform distribution being a special case with a power-law exponent of -1. A common technique to detect power-law behaviour is to study the rank/frequency plot (cumulative histogram). In Figure 2, we show the kinetic data from the BioModels database in this double logarithmic plot and fit a straight line via least squares. Such data are said to follow Zipf's law or the Pareto distribution which provides additional evidence for the scale-invariant character of the data.

Based on the assumption that people unaware of Benford's law would tend to make up numbers randomly in a uniform fashion, Hal Varian suggested in 1972 that the first-digit phenomenon could be used to detect possible fraud in economic data. Benford's law is now used to analyse accounts, insurance, and economic data to detect potential anomalies, (Varian, 1972). Similarly, on close inspection of individual biological pathway data, strong deviations from Benford's law may be observed in some cases, such as the BioModels entry BIOMD0000000014. This model has a first digit distribution consisting only of 1s (48%) and 5s (52%) and can be detected immediately in a Benford plot. Many of these parameters were estimated from an evolutionary optimization approach ignorant of Benford's law rather than experimentally determined. The analysis of further deviations, reveals that digit distributions from parameter estimation programs tend to over-populate numbers with a leading digit of 1 and 5, compared to Benford's law. These methods may produce useful approximations but their precise values are not realistic. We suggest that this is a consequence of the sampling strategy, i.e. the distribution and the coarseness of discretization. As previously shown, if the scale of the data is not known, then the log-uniform distribution is a reasonable prior probability assignment and it is from this distribution that should be sampled.

We have thus shown that kinetic data from biological pathways follow Benford's law closely and are therefore approximately scale-invariant within a finite region. This finding places the numerical data on a similar power-law foundation that is thought to exist for the topology of networks.

ACKNOWLEDGEMENTS

We thank Dr Martin Howard for critical reading of the manuscript, Robin Mason, Dr James Brown and Dr Irilenia Nobeli for helpful discussions and advice, and especially one anonymous referee for insightful criticism and constructive suggestions. SG and RJM are grateful for support from the John Innes Centre and the UK Biotechnology and Biological Sciences Research Council.

Conflict of Interest: none declared.

REFERENCES

Amaral,L.A.N. *et al.* (2000) *Proc. Natl Acad. Sci. USA*, **97**, 11149–11152.
Barabáasi,A.-L. and Albert,R. (1999) *Science*, **286**, 509–512.
Benford,F. (1938) *Proc. Amer. Phil. Soc.*, **78**, 551–572.
Hill,T.P. (1995) *Proc. Amer. Math. Soc.*, **49**, 1609–1625.
Hill,T.P. (1996) *Stat. Sci.*, **10**, 354–363.
Jaynes,E.T. (2003) *Probability Theory: The Logic of Science*. G Larry Bretthorst, Cambridge.
Kumar,M.D. *et al.* (2006) *Nucl. Acids Res.*, **34**, D204–D206.
Newcomb,S. (1881) *Amer. J. Math.*, **4**, 39–40.
Le Novère,N. *et al.* (2006) *Nucl. Acids Res.*, **34**, D689–D691.
Nye,J. and Moul,C. (2007) *BE J. Macroeconomics*, **7**, 1 (Article 17).
Puvanendrampillai,D. and Mitchell,J.B.O. (2003) *Bioinformatics*, **19**, 1856–1857.
Thiruv,B. *et al.* (2005) *BMC Struct. Biol.*, **5**, 12.
Varian,H. (1972) *Am. Stat.*, **26**, 65.
Wu,C.H. *et al.* (2006) *Nucl. Acids Res.*, **34**, D187–D191.