New Journal of Physics
The open–access journal for physics

The tails of rank-size distributions due to multiplicative processes: from power laws to stretched exponentials and beta-like functions

G G Naumis and G Cocho
Instituto de Física, Universidad Nacional Autónoma de México (UNAM), Apartado Postal 20-364, 01000, México, Distrito Federal, Mexico
E-mail: naumis@fisica.unam.mx

New Journal of Physics **9** (2007) 286
Received 17 May 2007
Published 28 August 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/8/286

Abstract. Although power laws have been used to fit rank distributions in many different contexts, they usually fail at the tail. Here we show that many different data in rank laws, like in granular materials, codons, author impact in scientific journals, etc are very well fitted by a β-like function ($\{a, b\}$ distribution). Since this distribution is indeed ubiquitous, it is reasonable to associate it with some kind of general mechanism. In particular, we have found that the macrostates of the product of discrete probability distributions imply stretched exponential-like frequency-rank functions, which qualitatively and quantitatively can be fitted with the $\{a,b\}$ distribution in the limit of many random variables. We show this by transforming the problem into an algebraic one: finding the rank of successive products of a given set of numbers.

Power-laws in rank distribution frequencies seem to be ubiquitous in physics, biology, geography, economics, linguistics, etc [1, 2]. For example, the frequency of words in different languages obey the Zipf power law [1]. In physics, we can cite the rank distribution of stick-slip events in sheared granular media [3], earthquakes (known in the field as the Gutenberg–Richter law [3]), radionuclides half-life time and nuclides mass number [4]. Other complex systems like networks [5], biological clocks [6] and metabolic networks [7] share as well the same phenomenology. Zipf discovered his rank law by analyzing manually the frequencies of words in the novel 'Ulysses' by James Joyce. It contains a vocabulary of 29 899 different word types. However, when larger corpora are used a deviation from a power law is observed for larger

PII: S1367-2630(07)50604-X

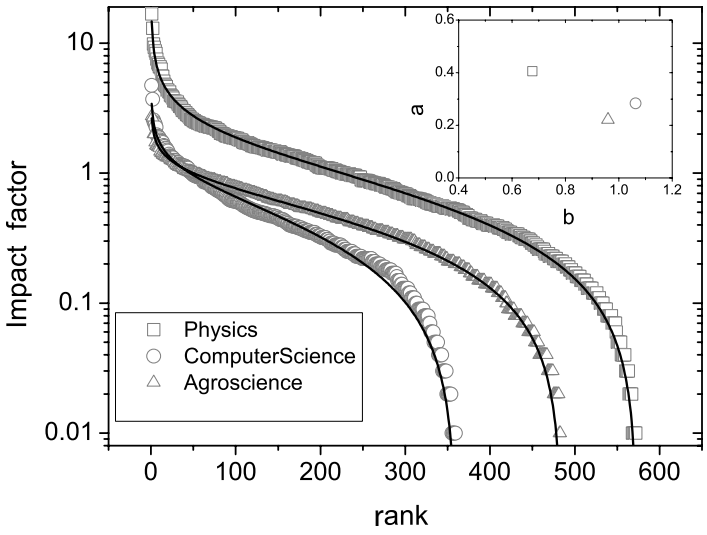

Figure 1. Impact factor as a function of the rank for physics, computer science and agroscience. Fits using the β-like function are shown as solid lines. Inset: values of a and b. The size of the symbols is proportional to the error.

ranks [8]. Such deviation is also found in many physical systems [9] and is known as the tail of the distribution [10]. As a matter of fact, when the highest rank of the data is finite, the power law is cut and finite size effects should be present. Usually, for each case a different ad hoc fitting function is proposed [3]. Another path is to construct a rank-size distribution from the cumulative distribution [10], by which method others have fitted the probability distributions with stretched exponential [9] and log-normal distributions [11]. However, for low ranks deviations are also observed, and unfortunately all of the previous expressions do not fit the data at *both ending tails*, at which different kinds of processes are set in once a crossover region is reached. Thus, multiscaling physical modeling seems to be a key issue as in turbulence, where Kolmogorov's power law is observed only in the inertial regimen. In one tail (small length scales) energy dissipation plays the main role, while energy injection dominates at big scales. One can conjecture that similar ideas are behind many other complex physical systems, since we report that many rank laws are extremely well parametrized, with a two exponent β function-like formula with parameters $\{a, b\}$,

$$f(r) = K\frac{(R-r+1)^b}{r^a}, \tag{1}$$

where a and b are fitted from the data, r is the rank and R is the maximal r. If $f(r)$ is normalized to 1, then $K \equiv 1/\sum_{r=1}^{R}(R-r+1)^b/r^a$. For $R \gg 1$, K can be transformed into an integral that yields $K \approx \Gamma(b-a+2)/\Gamma(1-a)\Gamma(1+b)$. We will show that $f(r)$ is related to a kind of central limit theorem. In fact, Moyano *et al* [12] have commented that the rather ubiquitous presence of the Tsallis q-distributions is maybe due to a q-generalized central limit theorem for a class of non independent, correlated, product of probability distributions [13].

As an example of the phenomenology that we have found in rank laws, here we present three representative results. Figure 1 shows a semilog plot of the impact factor against the rank of scientific journals, taken from a recent study [14], compared with the fits given by equation (1). The fits are excellent, all with correlation coefficients above 0.98. Notice that we use a semilog

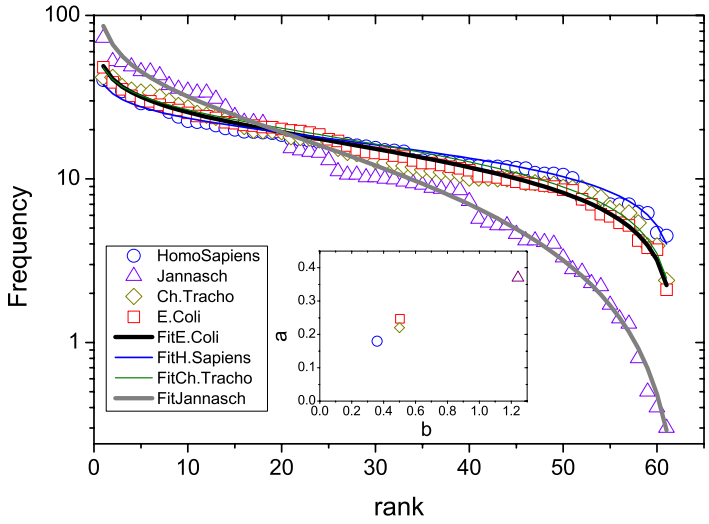

Figure 2. Frequency of codons (normalized to 1000) as a function of the rank for the genome of four different species, with their correponding fits shown as solid lines. Inset: values of a a and b used for the fits in the beta-like distribution.

plot instead of the usual log–log plot representation due to the fact that in such a way, the tails are easily observed since the ranks are treated in a linear fashion. A usual plot in the log–log representation will also reveal a significant deviation from the power law at the tails, as can be observed in several works [1]–[3].

Similar excellent fits are also obtained for codon usage in genomes, as shown in figure 2, where we plot the logarithm of the frequency of codons (normalized to 1000) as a function of the rank for different representative organisms, taken from a well known genome database [15]. Using equation (1), we have made similar plots for at least 10 organisms with a correlation parameter bigger than 0.99.

In figure 3, we show the rank-ordered distribution of stick-slip events in a slowly sheared granular media taken from [3], fitted using equation (1). Although a modified power law was proposed in [3] to explain the results, the present fit gives a better correlation coefficient. We have verified that equation (1) can be used with excellent results in order to correct the Gutenberg–Richter law, Bénard convection cells and in many different fields, like architecture, population, music or roads [16].

As the $\{a, b\}$ distribution is indeed ubiquitous, it is reasonable to try to associate it with the product of correlated probability distributions [12]. We have not found such a class; however, here we will show that the product of discrete probability distributions imply stretched exponential-like frequency-rank functions, that qualitatively and quantitatively can be fitted very well with the $\{a, b\}$ distribution.

In the dynamics of scientific journal impact factor there are many important issues: the ability to select a good problem for investigation, the gift for writing clear papers, etc. Similar comments would be valid for the dynamics of granular media. Perhaps, the presence of all these factors implies products of probabilities which obey the conditions of the hypothetical central limit theorem for the product of correlated probability distributions [12, 17]. To be concrete, let us proceed in the same spirit of the central limit theorem, in which a given observable is just the result of many different random processes. Each realization of an observable, is determined by

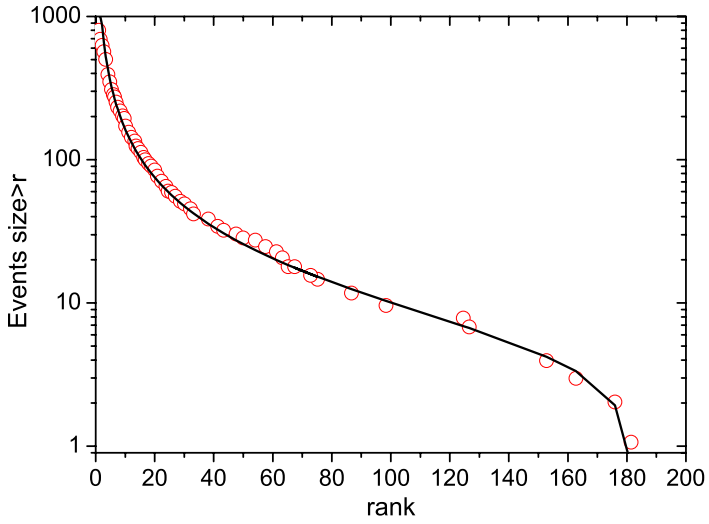

Figure 3. Rank-ordered distribution of stick-slip events in a slowly sheared granular media. Circles are data taken from [3], and the solid line is a fit using equation (1), with $a = 1.08$ and $b = 0.40$.

the actual values taken by the random variables in the involved random processes. For example, the distribution of heights in a population is determined by genetics, aliments, health care, etc, but the height of a particular person is just a realization with certain values of the random variables in each related processes. A similar thing happens in multiplicative processes, an observable is built from realizations made in each process. As in the case of the central limit theorem, each process has a different probability distribution. However, when many random variables are considered, only the first moments of these distributions turn out to be important, and as a consequence, a Gaussian appears with or without different probability distribution functions. Thus, in order to simplify the problem, here we will only consider the case of N processes that are identical, where each of them can have s different states with probability p_j with $j = 1, \ldots, s$. When N such processes are composed, the full state space may be considered to consist of all s^N possible strings of length N, and there are s^Npossible states of the whole system. One can reduce the probability of each of these states to just $(N+s-1)!/(s-1)!(N)!$ different values that we call the *reduced probabilities* $x_N(n_1, n_2, \ldots, n_s)$. The multiplicity of the states is given by a multinomial coefficient $N!/(n_1!n_2!n_3!\ldots n_s!)$, where n_j is the number of subsystems in the state j. The probability of an observable for the whole system is,

$$P_N(n_1, n_2, \ldots, n_s) = \frac{N!}{n_1!n_2!n_3!\ldots n_s!} x_N(n_1, n_2, \ldots, n_s), \tag{2}$$

with $n_1 + n_2 + n_3 + \cdots + n_s = N$. However, we are interested in the *rank of the different values of the resulting observable*, not in the probability distribution. To tackle this problem, we notice that each different value of $x_N(n_1, n_2, \ldots, n_s)$ corresponds to a *different observable*, since if one assumes that a certain observable (X) is a one to one function of $n_1, n_2, \ldots, n_s$, then each value of $X(n_1, n_2, \ldots, n_s)$ can be mapped to $x_N(n_1, n_2, \ldots, n_s)$ and $X(n_1, n_2, \ldots, n_s) = X(x_N(n_1, n_2, \ldots, n_s))$. From the previous considerations, it is clear that any rank hierarchy of $x_N(n_1, n_2, \ldots, n_s)$ will be inherited to $X(n_1, n_2, \ldots, n_s)$ in most physical cases, where one can

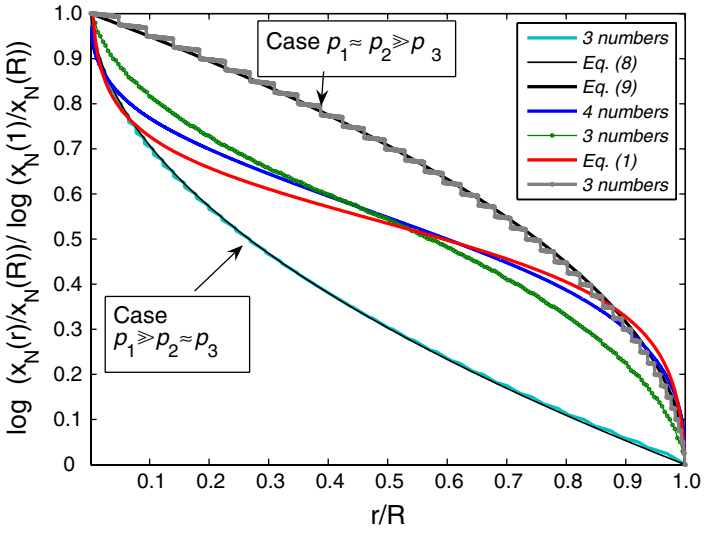

Figure 4. Normalized logarithm of numbers obtained after 40 iterations of succesive multiplication, as a function of the rank (also normalized) for 3 numbers with $p_1 \sim p_2 \gg p_3$, $p_1 \gg p_2 \sim p_3$ and $p_1 \sim p_2 \sim p_3$ (with label '3 numbers'). We also plot the case of an initial set of 4 numbers. Equations (1), (9) and (10) are indicated as solid lines. Notice how equations (9) and (10) dominate at the tails.

suppose that $X(x_N(n_1, n_2, \ldots, n_s))$ can be expressed as a power series in $x_N(n_1, n_2, \ldots, n_s)$,

$$X(x_N(n_1, n_2, \ldots, n_s)) = X_0 + X_1 x_N(n_1, n_2, \ldots, n_s) + \cdots, \tag{3}$$

where X_0 and X_1 are constants. Up to first order, this assumption means that X is proportional to $x_N(n_1, n_2, \ldots, n_s)$. The rank features are thus reduced to study the hierarchy present in $x_N(n_1, n_2, \ldots, n_s)$. In the general case of interacting processes, the addition of a new one leads to a relationship of the type $x_{N+1}(n_1, n_2, \ldots, n_s) = f(x_N(n_1, n_2, \ldots, n_s))$, while for independent processes,

$$x_N(n_1, n_2, \ldots, n_s) = p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots p_s^{n_s}. \tag{4}$$

For the last case, the rank structure can be reduced to the following algebraic problem: take s numbers $p_1, p_2, \ldots, p_s$ at random (normalization can be imposed at the end of the process), labeled in such a way that $p_1 > p_2 > \cdots > p_s$, and multiply each number by all the other ones. With these resulting numbers, repeat the process N times, to obtain a set of numbers that have the form $p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots p_s^{n_s}$, with the restriction $n_1 + n_2 + \cdots + n_s = N$. If the resulting set is arranged in decreasing order, we can assign a rank (r) to each one according to its order in the hierarchy. The rank $r = 1$ is assigned to the number p_1^N, while the lowest $r = R$ corresponds to p_s^N. For example, choose three random numbers p_1, p_2 and p_3 and then form all of the possible products: p_1^2, $p_1 p_2$, $p_1 p_3$, p_2^2, $p_2 p_3$, p_3^2, then repeat the procedure. In figure 4 we present a plot of $\log x_N(n_1, n_2, n_3)$ as a function of r for $N = 40$. Surprisingly, as shown in the figure, the resulting ranks are well fitted by the same two parameter β-like function. The message from this numerical experiment is simple: if this product is seen as a multiplicative process where each number is the probability of making a certain choice in during the process, the result has a well determined hierarchy.

The problem that remains is how to calculate $x_N(n_1, n_2, \ldots, n_s)$ in terms of the rank. One can think of each set of values $(n_1, n_2, \ldots, n_s)$ as coordinates in an s-dimensional lattice, that live on a subspace of dimension [18] $s-1$, and the rank is a parametrization of a path between the lattice points in such a way that $\ln x_N(n_1, n_2, \ldots, n_s)$ decreases in each step,

$$\ln x_N(r) = n_1(r) \ln p_1 + n_2(r) \ln p_2 + \ldots + n_s(r) \ln p_s, \tag{5}$$

where the starting point is always $(N, 0, \ldots, 0)$ and the end is at $(0, 0, \ldots, N)$. For $s = 2$, the solution is easy to find. Using that $n_1 + n_2 = N$, it follows that,

$$x_N(r) = p_1^N \left(\frac{p_2}{p_1}\right)^{r-1} = p_1^N \mathrm{e}^{-C(r-1)}, \tag{6}$$

with $C = |\ln(p_2/p_1)|$. Equation (6) shows that the numbers decay in a pure exponential way.

The case $s = 3$ can be easily visualized as a trajectory in a triangle. The solution for any set p_1, p_2, p_3 is complicated, since the paths are usually complex. However, one can work out the cases $p_1 \sim p_2 \gg p_3$ and $p_1 \gg p_2 \sim p_3$; these cases provide the clue to solve others.

Consider the limit $p_1 \sim p_2 \gg p_3$, and $\delta_{21}^2 \gg \delta_{31}$, where we define $\delta ij \equiv p_i/p_j$. This rank sequence is similar to an odometer with an increased range after each turn due to the hierarchy $1 > \delta_{21} > \delta_{21}^2 > \delta_{31} > \delta_{21}\delta_{31} > \delta_{31}^2 > \cdots > \delta_{31}^N$. For example, when $N = 2$ this leads to the following table that contains the number $x_N(r)$ as a function of the rank and the corresponding path,

$x_N(r)$	n_2	n_3	r	$n_{2M}(r)$
p_1^2	0	0	1	–
$p_1^2\delta_{21}$	1	0	2	–
$p_1^2\delta_{21}^2$	2	0	3	2
$p_1^2\delta_{31}$	0	1	4	–
$p_1^2\delta_{21}\delta_{31}$	1	1	5	1
$p_1^2\delta_{31}^2$	0	2	6	0

The sequence of the path goes as follows, first $n_2(r)$ is increased one by one as $n_3(r)$ remains constant, until it reaches a maximal value called $n_{2M}(r)$ which in fact determines the basic shape of the curve $x_N(r)$, since δ_{31} only produces small jumps (see figure 4). Once $n_2(r)$ increases from zero to $n_{2M}(r)$, a new cycle begins with $n_2(r) = 0$ and $n_3(r+1) = n_3(r) + 1$. As a result, after a large number of steps,

$$R - r = \sum_{j=1}^{n_{2M}(r)} j = \frac{n_{2M}(r)(n_{2M}(r)+1)}{2}, \tag{7}$$

where R is the maximal rank. Then, by solving the resulting quadratic equation and since $1 \ll N \ll R$, we get that $n_{2M}(r) \approx N(R-r+1)^{1/2}$. The corresponding value of $n_3(r)$ can be obtained from the condition $n_2(r) + n_3(r) \leqslant N$. Finally, the number as a function of the rank is given by,

$$x_N(r) \approx \left[p_1 \left(\frac{p_2}{p_1}\right)^{(1-r-1/R)^{1/2}} \left(\frac{p_3}{p_1}\right)^{1-(1-r-1/R)^{1/2}} \right]^N. \tag{8}$$

Furthermore, equation (8) can be written as an stretched exponential as follows,

$$x_N(r) \approx p_3^N \exp\left[D\left(1-\frac{r-1}{R}\right)^{1/2}\right], \tag{9}$$

with $D = N|\ln(p_2/p_3)|$. This curve shows an excellent agreement with the numerical results (see figure 4). Notice in figure 4 how this formula works better as the rank approaches R. The case $p_1 \gg p_2 \sim p_3$ can be tackled in a similar way. The result is,

$$x_N(r) \approx p_1^N \exp\left[-E\left(\frac{r}{R}\right)^{1/2}\right], \tag{10}$$

with $E = N|\ln(p_1/p_3) - \ln(p_2/p_3)|$. The comparison with the numerical results is also excellent (see figure 4).

In the general case, when p_1, p_2 and p_3 have the same order of magnitude, as for example in figure 4, there are two tails for $r \to 1$ and $r \to R$. The tail at low r is basically produced by the hierarchy in the biggest probabilities, i.e. by numbers where $n_1 \sim N$ in which equation (10) gives the upward curvature. In a similar way, the tail for r near R is produced by the lowest probability hierarchy, $n_3 \sim N$, controlled basically by equation (9). The main effect upon these tails when $p_1 \sim p_2 \sim p_3$ is that the sequence of ordering is not uniform [18], and there is a change in the exponent $1/2$ that appears in the stretched exponential. Equation (10) is thus transformed into,

$$x_N(r) \approx p_1^N \exp\left[-E_s\left(\frac{r}{R}\right)^{\alpha}\right], \tag{11}$$

where $\alpha < 1/2$ and E_s is a constant that depends on s. In a similar way, the exponent $1/2$ in equation (9) is replaced by an exponent $\beta < 1/2$. Furthermore, these generic exponents for the tails also appear for $s > 3$ since from the polynomial equivalent to equation (7) one gets $\beta \approx 1/(s-1)$. These changes are the result of the increasing number of cycles in the 'odometer' that we have discussed, as is also clear from the change in the exponents that transform from 1 to $1/2$ as s goes from $s = 2$ to $s = 3$. A simple procedure to combine each of the tails represented by equations (9) and (10) is obtained by making the observation that only one of the exponentials dominates at a given tail, while the other tends toward a constant, i.e. the logarithm derivative of equation (10),

$$\left(\frac{\mathrm{d}\ln x_N(r)}{\mathrm{d}r}\right) = -\frac{\alpha E_s}{R}\left(\frac{r}{R}\right)^{\alpha-1}, \tag{12}$$

is nearly zero if $r \to R$ for $R \gg 1$. Analyzing the limit $r \to R$ gives a similar result for equation (9). From these considerations, a simple way to produce a function with the required tails at both ends when $r \to R$ and $r \to 1$ is the following,

$$x_N(r) \approx C_1 \exp\left[D_s\left(1-\frac{r-1}{R}\right)^{\beta}\right] \exp\left[-E_s\left(\frac{r}{R}\right)^{\alpha}\right], \tag{13}$$

where C_1 is a constant and D_s is another constant that depends on s. Finally, equation (13) can be simplified when many processes are present, since $s \gg 1$ and as a consequence $\alpha \to 0$ and $\beta \to 0$. For this limit, in the logarithm derivative equation (12) one can neglect α with respect to 1 and $\ln x_N(r) \approx -\alpha E_s \ln(r/R)$. A similar procedure can be done in the tail $r \to 1$, in which β is neglected with respect to 1. Combining both tails in a sole expression, we get the fitting

function used in this work: $x_N(r) = K(R - r + 1)^b / r^a$, where

$$b = \beta D_s \quad \text{and} \quad a = \alpha E_s, \tag{14}$$

and K can be obtained by self-consistency. The previous law is the limiting form of two stretched exponentials when the number of states of processes involved is big, as can be confirmed by comparing the cases with 3 and 4 numbers in figure 4. In conclusion, we have shown a simple formula that allows to fit many different rank phenomena which arises as a limiting case for products of random variables. A task that remains is how to get the coefficients a and b from physical principles, using for example master equations and the concept of multiscaling modeling. A key observation for such a study is that for expansion-modification algorithms in DNA models, $a > b$ if the expansion probability of the genetic code is bigger than the mutation rate [19]. Thus, a and b represent the relative influence of two general mechanisms, where each of them dominate at a given tail. According to some preliminary results, a seems to be related to a certain funnel type of energy landscape, as in protein folding, which leads to a certain deterministic sequence, while b is associated with a many valley landscape, as seen in spin glasses. This last opposite effect provides much more variability in the sequence of results. Such correlation is consistent with associating b to the stochastic component of the dynamics and a with the most deterministic features [19]. In a forthcoming paper, we will analyze specific trends in a and b for different classes or systems.

Acknowledgment

This work was supported by DGAPA UNAM project IN108502, and CONACyT 48783-F and 50368.

References

[1] Li W 1991 *Phys. Rev.* E **43** 5240
See also: Li W 2003 http://www.nslij-genetics.org/wli/zipf/
[2] Benguigui L and Blumenfeld-Lieberthal E 2006 *Int. J. Mod. Phys.* C **17** 1429
[3] Bretz M, Zaretzki R, Field S B, Mitarai N and Nori F 2006 *Europhys. Lett.* **74** 1116
[4] Audi G, Bersillon O, Blachot J and Wapstra A H 1997 *Nucl. Phys.* A **624** 124
[5] Fortunato S, Flammini A and Menczer F 2006 *Phys. Rev. Lett.* **96** 218701
[6] Yang A C C, Hseu S S, Yien H W, Goldberger A L and Peng C K 2006 *Phys. Rev. Lett.* **90** 108103
[7] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabasi A L 2000 *Nature* **407** 651
[8] Le Quan H, Sicilia-García E I, Minj J and Smith F J 2002 *Proc. 17th Int. Conf. on Computer Lingustics (Montreal)*
[9] Laherrere J and Sornette D 1998 *Eur. Phys. J.* B **2** 525
[10] Sornette D 2004 *Critical Phenomena in Natural Sciences* 2nd edn (Berlin: Springer Verlag)
[11] Montroll E W and Shlesinger M F 1983 *J. Stat. Phys.* **209** 32
[12] Moyano L G, Tsallis C and Gell-Mann M 2006 *Europhys. Lett.* **72** 355
[13] Marsh J A, Fuentes M A, Moyano L G and Tsallis C 2006 *Physica* A **372** 183
[14] Popescu I 2003 *Glottometrics* **6** 83
[15] Codon Usage Database, NCBI-GenBank. Online at http://www.kazuza.or.jp/codon
[16] Cocho G and Martínez-Mekler G to be published
[17] Manrubia S C and Zanette D H 1999 *Phys. Rev.* E **59** 4945
[18] Naumis G and Cocho G to be published
[19] Mansilla R and Cocho G 2000 *Complex Syst.* **12** 207