Eur. Phys. J. B **50**, 259–264 (2006)
DOI: 10.1140/epjb/e2006-00121-2

THE EUROPEAN PHYSICAL JOURNAL B

Non-extensive trends in the size distribution of coding and non-coding DNA sequences in the human genome

Th. Oikonomou[1,2,a] and A. Provata[1,b]

[1] Institute of Physical Chemistry, National Center for Scientific Research "Demokritos", 15310 Athens, Greece
[2] School of Medicine, Department of Biological Chemistry, University of Athens, Goudi, 11527 Athens, Greece

Received 25 October 2005 / Received in final form 1st December 2005
Published online 12 April 2006 –

Abstract. We study the primary DNA structure of four of the most completely sequenced human chromosomes (including chromosome 19 which is the most dense in coding), using non-extensive statistics. We show that the exponents governing the spatial decay of the coding size distributions vary between $5.2 \leq r \leq 5.7$ for the short scales and $1.45 \leq q \leq 1.50$ for the large scales. On the contrary, the exponents governing the spatial decay of the non-coding size distributions in these four chromosomes, take the values $2.4 \leq r \leq 3.2$ for the short scales and $1.50 \leq q \leq 1.72$ for the large scales. These results, in particular the values of the tail exponent q, indicate the existence of correlations in the coding and non-coding size distributions with tendency for higher correlations in the non-coding DNA.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Da Systems obeying scaling laws – 87.14.Gg DNA, RNA

1 Introduction

During recent years numerous studies on the statistics of genomic sequences have demonstrated various degrees of complexity in the primary structure of DNA. In particular, Peng et al. in 1992 demonstrated the existence of long range correlations using the "DNA walk" model [1]. Similar conclusions were reached by Li et al. [2] and Voss [3] using the $1/f$ spectrum and later by studies on the size distribution of Purine (Adenine, Guanine) and Pyrimidine (Thymine, Cytocine) clusters in coding and non-coding regions of different organisms [4,5]. Other studies manifested long range correlations and power laws in the primary structure of DNA using a variety of statistical methods ranging from wavelets to linguistic approaches [6].

In recent studies, one of the present authors (AP) and coworkers have shown that the long range distributions of Pyrine and Pyrimidine clusters in the non-coding regions of higher eucaryotes are related to similar long range distributions present at a higher level of genomic organisation: the level of coding and non-coding alternating regions [7].

Non-extensive statistical mechanics is particularly fitted to describe complex structures which present long range correlations, power laws and fractality [8]. In particular, non-extensive statistics have been used to describe successfully complex spatiotemporal structures in diverse fields such as high energy physics, turbulence, biological systems, anomalous diffusion, classical and quantum chaos, interacting particle systems and reactive dynamics [9].

Classical statistical mechanics uses the Boltzmann Gibbs (BG) Entropy, S_{BG}, defined as:

$$S_{BG} = -\sum_{i=1}^{W} p_i \ln p_i \tag{1}$$

to describe the properties of systems at equilibrium. In equation (1), p_i denotes the probability of the ith microscopic state and the average runs over the total number of states W. This BG entropic form can not successfully describe systems in which self-organisation, long range features and scaling are observed. As a generalisation of equation (1), Tsallis and coworkers [10] have introduced the non-extensive entropy, defined as:

$$S_q = \frac{1 - \sum_{i=1}^{W} p_i^q}{q-1}, \text{ for } q \neq 1 \tag{2}$$

where q is the non-extensivity exponent. Note that for $q = 1$ the classical BG statistics (Eq. (1)) is recovered and thus departure of the exponent q from the value 1 signals departure from BG statistics.

[a] e-mail: thoikonomou@chem.demokritos.gr
[b] e-mail: aprovata@limnos.chem.demokritos.gr

In relation to non-extensive statistics, long range behaviour may be obtained by a non-linear equation expressed as [11]:

$$\frac{d\xi}{ds} = -\kappa_q \xi^q, \quad \text{for} \quad (\kappa_q \geq 0, \quad q \neq 1)\,. \tag{3}$$

In particular, for $q > 1$ long range behaviour is manifested, while for $q = 1$ the well known exponential law is obtained. The solution of equation (3) is:

$$\xi(s) = [1 - (1-q)\kappa_q(s-1)]^{1/(1-q)}\,, \text{ for } (\kappa_q \geq 0,\ q > 1) \tag{4}$$
$$= \exp(-\kappa_1(s-1)), \quad \text{for} \quad (\kappa_1 \geq 0,\ q = 1)$$

with initial condition $\xi(1) = 1$. Thus for $q > 1$ a long range law (power law) is obtained, while for $q = 1$ a short range (exponential) law emerges.

For phenomena which exhibit crossover between two different regimes in the short and long length scales, a further phenomenological generalisation of equation (3) maybe introduced by addition of terms carrying different powers [11]. The simplest one carries only one additional term and is:

$$\frac{d\xi}{ds} = -\kappa_q \xi^q - (\lambda_r - \kappa_q)\xi^r, \quad \text{for} \quad (q \leq r). \tag{5}$$

Note that equation (3) is recovered for $\kappa_q = \lambda_r$ $(\forall r)$. The solution of equation (5) can not be written in a simple form but it may be shown that it consists of two distinct power law regions, one governed by the exponent q and one by the exponent r [11].

In Figure 1 we present the size distribution of coding and non-coding DNA sequences in chromosome 16. To avoid local fluctuations running averages are considered over 15 Base Pairs (bps). For clarity only the first 1000 points are shown. The maximum size of coding regions is of the order of 7000−8000 bps (reaches $\sim 20\,000$ bps for chromosome 19) while the maximum sizes of the non-coding regions reach $\sim 10^8$ bps. The coding size distributions are rich in small segments of the order of 100−110 bps and then fall fast, while the non-coding ones have a similar maximum in the small scales and fall relatively slower. For comparison we also present the size distribution of chromosome 17, in Figure 2, in double logarithmic scale where the entire s-range is shown.

Comparing Figures 1 and 2 we note that the size distribution of non-coding DNA, has a complex form but we may clearly distinguish two regions: one region at the short length scales which is bell-shaped and which mostly describes the introns (non-coding regions within genes) and one region at the larger scales which contains a long tail and which describes mostly the non-coding intergenic regions. It is thus natural, at the phenomenological level, to use equation (5) for the description of the complex shape of the size distribution of non-coding DNA hoping to capture these two trends, the introns and the intergenic regions.

In the current study we use non-extensive statistics to study globally the size distributions of coding and non-coding sequences in the human genome which is now near completion. We have selected to study four of the most complete human chromosomes including chromosome 19 which contains the highest percentage of coding. In the next section we concentrate on the primary structure of the human genome and we give details on the particular data we use. In Sections 3 and 4 we present the analysis of the size distribution of coding and non-coding DNA, respectively. We conclude by summarising our main results and discussing some open problems.

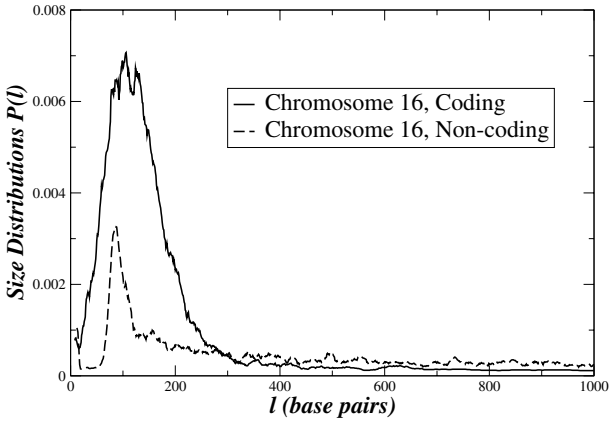

Fig. 1. Running average over 15 points of the size distribution of coding and non-coding DNA in chromosome 16. Only sizes of up to 1000 bps are shown.

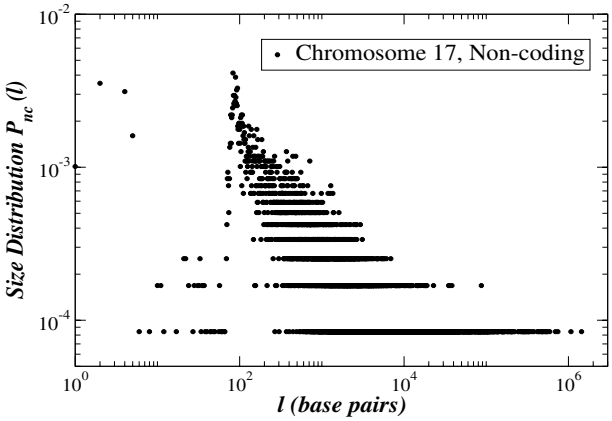

Fig. 2. The size distribution of non-coding DNA in chromosome 17 in a double logarithmic scale (all data).

2 The human genome data

Although officially the human genome project is announced to be near completion, in the international EMBL and GenBank genomic data bases the sequence data deposited varies from 98.91% for chromosome 17 to 43.1% for chromosome Y. The unknown base pairs are usually denoted by the letter N = (unknown base pair) and they are either isolated or appear in clusters. The meaning of N is not unique. It might denote a base pair which resists to sequencing methods completely or partially. Resisting

partially means that partial information on the base is known, for example being a Purine or a Pyrimidine. Another case is that the various laboratories which verify the sequencing may not agree on this base pair.

In the current project we analyse the complete primary structures of human chromosomes 6, 16, 17 for which the N percentage is the smallest and also chromosome 19, which contains the highest percentage of coding DNA, 3.8%. The sequenced percentage presented in the data bases and the coding percentage of these are shown in Table 1. After downloading the chromosomes we isolate the coding and non-coding segments and calculate the respective size distributions for each one of them. A representative plot is shown in Figure 2. Due to the heavy fluctuations in the data we prefer to work with the cumulative distributions $\tilde{P}(s)$ defined as:

$$\tilde{P}(s) = \int_s^\infty P(l)dl \tag{6}$$

where $P(l)$ is the usual distribution of coding or non-coding regions of size l. In general, due to summation the cumulative distributions have better statistical properties than the usual distribution functions while they keep the main data trends. Notice that, if the distribution $P(l)$ has the exponential (short range) form its cumulative $\tilde{P}(s)$ will also have the exponential form. If the distribution function $P(l)$ has a power law form with exponent $-1-\mu$ as in equation (7), then the cumulative distribution will have a power law form with exponent $-\mu$ (see equation (7)),

$$P(l) \sim l^{-1-\mu} \implies \tilde{P}(s) = \int_s^\infty l^{-1-\mu} dl = s^{-\mu}, \;\; 0 \le \mu \le 2. \tag{7}$$

Cumulative diagrams of the four coding and non-coding cumulative size distributions are shown in Figures 3 and 4, respectively. The non-extensive analysis of these distributions follows in the next two sections.

3 Sizes of coding DNA sequences

As we have already seen in Figure 1 the coding size distributions have a bell-shape and their tails in the large scales fall relatively fast. To give a quantitative account for the decay of the distribution tails we plot the cumulative size distributions in Figure 3 (solid lines).

To describe the shape of the four curves we use the phenomenological non-extensive description of equation (5) and the corresponding curves are also shown in the same figures (dashed lines). The fit is performed by numerically solving equation (5) since its exact solution is only known for very specific values of $q = 0, 1$ as shown in reference [11]. We scan the parameter space $(q, r, \kappa_q, \lambda_r)$, with step sizes (0.005, 0.1, 0.00001, 0.00001) respectively and we determine the parameter values which best fit the DNA data (both in the short and long size scales). The theoretical lines approximate well the data. The exponents q which describe the tails of the distributions vary between $1.45 < q < 1.50$ for the four chromosomes and their specific values are given in Table 1. The non-extensive exponent q corresponds to power law tails of the form equation (7) with exponent μ given by

$$\mu = -1/(1-q). \tag{8}$$

Thus the tails of the coding size distributions present short range correlations, since $\mu \ge 2$. The exponent r which expresses the small scale characteristics, takes values between $5.2 < r < 5.6$ for these chromosomes. Similar results have also been observed for the other human chromosomes. The similarity of the two exponents in the four chromosomes indicate that the same (or similar) dynamical, evolutionary processes have created the coding parts of all chromosomes during evolution. This dynamics must be of conservative type in short time scales, since coding DNA changes very slowly (behaves as an almost-closed system) and this is consistent with short range correlations [7].

Table 1. Non-extensive exponents and parameters describing the coding size distributions.

Chromosome	Sequenced %	Coding %	q	r	κ_q	λ_r	$\mu = 1/(q-1)$
6	97.86	1.03657	1.50	5.2	0.018	0.00012	2.00
16	88.81	1.67416	1.45	5.7	0.017	0.00009	2.22
17	98.91	2.48184	1.45	5.4	0.018	0.00010	2.22
19	87.43	3.39768	1.50	5.6	0.018	0.00012	2.20

4 Sizes of non-coding DNA sequences

The cumulative size distributions of the non-coding DNA in the four chromosomes are shown in Figure 4 (solid lines). We observe that the four distributions have as common characteristic a long tail which can be expressed in the form of a pure power law [7]. In the smaller scales the behaviour is characterised by a different exponent which is very similar for the four distributions.

To describe the shape of the four curves we use the phenomenological non-extensive description of equation (5) and the corresponding curves are shown in the same figures (dashed lines). The fit was performed numerically as in the previous section. The theoretical lines are very faithful approximations to the data. The exponents q which describe the long tails of the distributions are very close for the four chromosomes and their corresponding values are given in Table 2. Their values vary between $1.50 < q < 1.72$. The non-extensive exponent q corresponds to a power law of the form equation (7) with exponent μ being within the bounds $0 \le \mu \le 2$, which indicates long range correlations. These values of μ can be verified by directly measuring the tail slopes in Figure 4. In the case of chromosome 19, which (up to now) contains the highest coding percentage amongst all human chromosomes, the value of μ calculated through equation (8) is equal to 2, which is border line case between short and

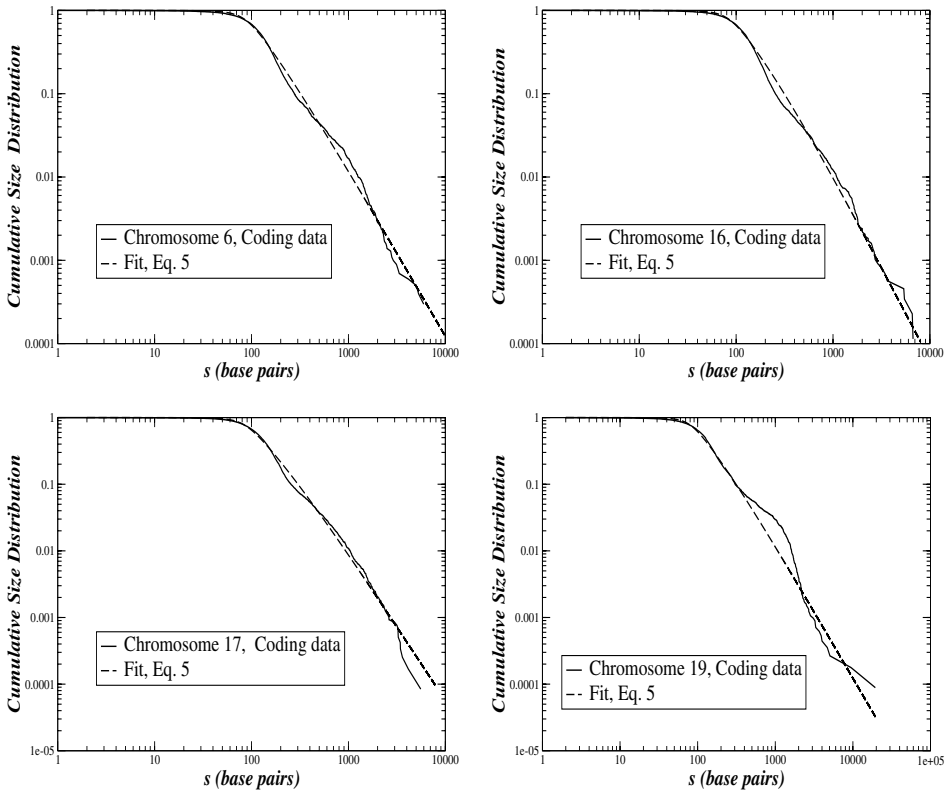

Fig. 3. The cumulative size distributions of coding DNA in chromosomes 6, 16, 17 and 19 (solid lines) and the non-linear fits using equation (5) (dashed lines).

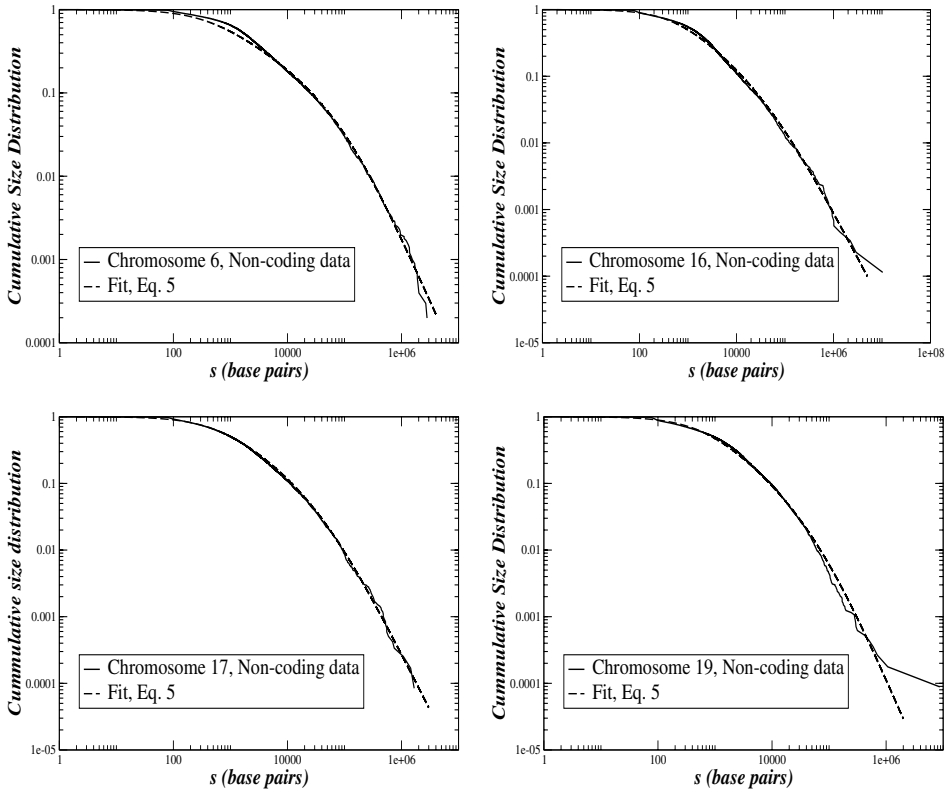

Fig. 4. The cumulative size distributions of non-coding DNA in chromosomes 6, 16, 17 and 19 (solid lines) and the non-linear fits using equation (5) (dashed lines).

Table 2. Non-extensive exponents and parameters describing the non-coding size distributions.

Chromosome	q	r	κ_q	λ_r	$\mu = 1/(q-1)$
6	1.65	3.2	0.00009	0.00120	1.54
16	1.72	2.7	0.00021	0.00124	1.39
17	1.59	2.7	0.00021	0.00118	1.69
19	1.50	2.4	0.00018	0.00124	2.00

long range correlations. On the other hand, the exponent r which expresses the small scale characteristics, takes values between $2.4 < r < 3.2$ for these chromosomes. Similar results have also been observed for all other human chromosomes.

The different small and large scale behaviour observed in the size distribution of the non-coding indicates that different evolutionary mechanisms are involved in the formation of small non-coding segments (which are usually found as introns in the genes) and in the large non-coding areas, or intergenic regions which are found between genes and between families of genes. The intergenic regions are extended non-coding regions which can support extensive (massive) influx and outflux of genomic material. Thus the ensemble of intergenic regions acts as an open system which supports exchange with the environment. In open systems, out of equilibrium, power laws and long range spatial correlations emerge naturally via non-extensive or edge of chaos evolutionary dynamics. Open aggregating systems, with influx mechanisms similar to the ones involved in genomic evolution and which lead to long range spatial correlations are presented in reference [7]. On the other hand, the non-coding segments found within genes, introns, are less supportive to external influences because often they include functional strings. Thus they behave more like closed systems and their evolutionary paths and stationary state statistics are expected to be similar to coding DNA. Further studies are needed to clarify to what extent the short and long scale behaviour found in the non-coding DNA are related to the intron and intergenic region statistics.

5 Conclusions

We have studied the size distribution of all known coding and non-coding sequences in human chromosomes 6, 16, 17 and 19. The first three were selected as representatives of the most completely sequenced chromosomes while chromosome 19 has the highest, up to date, coding percentage. We have found that the spatial decay of the non-coding size distributions is consistent with non-extensive statistics as expressed by the non-linear equation (5). We have observed two distinct regions in the non-coding: one large scale region, related to the intergenic non-coding DNA which presents a power law exponent $1.5 \leq q \leq 1.72$, and a second short scale region related to the introns (non-coding DNA within genes) which presents a power law exponent $r > 2.4$. The correlation exponents observed in the coding size distributions are between $1.45 \leq q \leq 1.50$. This is consistent with earlier observations of correlations in the size distributions of higher eucaryotes [6,7]. All other human chromosomes demonstrate similar characteristics.

A more detailed analysis could involve the use of more terms with different exponents in equation (5), in order to capture more details such as the exponent which govern non-coding distances between families of homologous genes (they may be governed by one of the current exponents, q or r, or by a third one). Also, the study of the statistics of genes and intergenic regions separately may indicate different characteristic exponents, which emerge as a result of different evolutionary paths.

It is true that today the human chromosomes may be close to full sequencing but their complete annotation will take much longer. This means that there are still coding sequences which are not discovered within the genome. Thus we expect that with the advancement of DNA annotation, which is the next major step in genomics after sequencing, we will be able to give more precise, final values to the exponents q and r for the human genome. Also the study in parallel of the genomes of other organisms, as they become sequenced and annotated, will allow for a comparative analysis of genomic data between different classes of organisms.

The authors would like to thank Prof. C. Tsallis for suggesting this approach and Prof. K. Trougkos for helpful discussions.

References

1. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature **356**, 168 (1992)
2. W. Li, K. Kaneko, Europhys. Lett. **17**, 655 (1992)
3. R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992)
4. R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E **52**, 2939 (1995)
5. A. Provata, Y. Almirantis, Physica A **247**, 482 (1997)
6. A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo, Y. d' Aubenton-Carafa, E. Bacry, P.V. Graves, J.F. Muzy, C. Thermes, Physica D **96**, 291 (1996); R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994); A. Czirók, R.N. Mantegna, S. Havlin, H.E. Stanley, Physical Review E **52**, 446 (1995); A.A. Tsonis, J.B. Elsner, P.A. Tsonis, J. Theor. Biol. **151**, 323 (1991); S. Karlin, V. Brendel, Science **259**, 677 (1993); H. Herzel, E.N. Trifonov, O. Weiss, I. Grosse, Physica A **249**, 449 (1998); S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev E **47**, 4514 (1993)
7. Y. Almirantis, A. Provata, J. Stat. Phys. **97** 233 (1999); A. Provata, Y. Almirantis, J. Stat. Phys. **106**, 23 (2002); P. Katsaloulis, T. Theoharis, A. Provata, Physica A **316**, 380 (2002)

8. *Nonextensive Statistical Mechanics and its Applications*, edited by S. Abe, Y. Okamoto, Series Lecture Notes in Physics (Springer-Verlag, Berlin, 2001); *Non Extensive Statistical Mechanics and Physical Applications*, edited by G. Kaniadakis, M. Lissia, A. Rapisarda, Physica A **305**, No 1/2 (Elsevier, Amsterdam, 2002)
9. I. Bediaga, E.M.F. Curado, J. Miranda, Physica A **286**, 156 (2000); C. Beck, Phys. Rev. Lett. **87**, 180601 (2001); N. Arimitsu, T. Arimitsu, Europhys. Lett. **60**, 60 (2002); M. Peyrard, I. Daumont, Europhys. Lett. **59**, 834 (2002); A. Upadhyaya, J.-P. Rieu, J.A. Glazier, Y. Sawada, Physica A **293**, 549 (2001); P.A. Alemany, D.H. Zanette, Phys. Rev. E **49**, R956 (1994); C. Tsallis, S.V.F. Levy, A.M.C. Souza, R. Maynard, Phys. Rev. Lett. **75**, 3589 (1995); A.R. Plastino, A. Plastino, Physica A **222**, 347 (1995); M.L. Lyra, C. Tsallis, Phys. Rev. Lett. **80**, 53 (1998); E.P. Borges, C. Tsallis, G.F.J. Ananos, P.M.C. Oliveira, Phys. Rev. Lett. **89**, 254103 (2002); Y.S. Weinstein, S. Lloyd, C. Tsallis, Phys. Rev. Lett. **89**, 214101 (2002); V. Latora, A. Rapisarda, C. Tsallis, Phys. Rev. E **64**, 056134 (2001); G.A. Tsekouras, A. Provata, C. Tsallis, Phys. Rev. E **69**, 016120 (2004)
10. C. Tsallis, J. Stat. Phys. **52**, 479, (1988); E.M.F. Curado, C. Tsallis, J. Phys. A **24** L69 (1991); C. Tsallis, R.S. Mendes, A.R. Plastino, Physica A **261**, 534 (1998)
11. C. Tsallis, G. Bemski, R.S. Mendes Phys. Lett. A **257**, 93 (1999)