ELSEVIER

15 July 2002

Physics Letters A 299 (2002) 565–570

PHYSICS LETTERS A

www.elsevier.com/locate/pla

Lévy statistics in coding and non-coding nucleotide sequences

Nicola Scafetta [a,b,*], Vito Latora [c], Paolo Grigolini [b,d,e]

[a] *Pratt School EE Department, Duke University, P.O. Box 90291, Durham, NC 27708, USA*
[b] *Center for Nonlinear Science, University of North Texas, P.O. Box 311427, Denton, TX 76203-1427, USA*
[c] *Dipartimento di Fisica e Astronomia, Universitá di Catania, and INFN, Corso Italia 57, 95129 Catania, Italy*
[d] *Dipartimento di Fisica dell'Università di Pisa and INFM, Piazza Torricelli 2, 56127 Pisa, Italy*
[e] *Istituto di Biofisica CNR, Area della Ricerca di Pisa, Via Alfieri 1, San Cataldo, 56010 Ghezzano-Pisa, Italy*

Received 16 April 2002; received in revised form 16 April 2002; accepted 23 May 2002

Communicated by A.R. Bishop

Abstract

The diffusion entropy analysis measures the scaling of the probability density function (pdf) of the diffusion process generated by time series imagined as a physical source of fluctuations. The pdf scaling exponent, δ, departs in the non-Gaussian case from the scaling exponent H_V evaluated by variance based methods. When $\delta = 1/(3-2H_V)$ Lévy statistics characterizes the time series. With the help of artificial sequences that are proved to be statistically equivalent to the real DNA sequences we find that long-range correlations generating Lévy statistics are present in both coding and non-coding DNA sequences. © 2002 Elsevier Science B.V. All rights reserved.

PACS: 05.40.+j

Keywords: Time series analysis; Lévy statistics; DNA

The recent progress in experimental techniques of molecular genetics has made available a wealth of genome data (see, for example, the NCBI's Gen-Bank data base of Ref. [1]), and raised the interest for the statistical analysis of DNA sequences [2–5]. These pioneer papers mainly focused on the controversial issue of whether long-range correlations are a property shared by both coding and non-coding sequences or are only present in non-coding sequences. The results of more recent papers [6,7] yield the convincing conclusion that the former condition applies. However, some statistical aspects of the DNA sequences are still obscure, and it is not yet known to what extent the dynamic approach to DNA sequences proposed by the authors of Ref. [8] is a reliable picture for both coding and non-coding sequences. The later work of Refs. [9] and [10] established a close connection between long-range correlations and the emergence of non-Gaussian statistics, confirmed by Mohanti and Narayana Rao [6]. According to the dynamic approach of Refs. [8,11] this non-Gaussian statistics should be Lévy, but this property has not yet been assessed with compelling evidence. The reason for this failure is that the scaling detection has been based upon the evaluation of the variance. In this Letter we aim at filling this gap and we show that the diffusion entropy analysis

[*] Corresponding author.
E-mail address: ns2002@duke.edu (N. Scafetta).

PII: S0375-9601(02)00730-2

(DEA) [12–14] realizes the goal of evaluating the genuine scaling value of the probability distribution. Finally, we prove that the joint use of the DEA and of the detrended fluctuation analysis (DFA), a widely used variance based method, applied to DNA sequences by the authors of Ref. [15], allows us to:

(1) establish the presence of long-range correlations in coding as well as in non-coding sequence;
(2) assess the Lévy nature of the resulting non-Gaussian statistics.

More specifically, we analyze the two DNA sequences studied in Ref. [15]. These two sequences are the human T-cell receptor alpha/delta locus, Gen-Bank name HUMTCRADCV, a non-coding chromosomal fragment of $M = 97630$ bases (composed of less than 10% of coding regions), and the Escherichia Coli K12, Gen-Bank name ECO110K, a genomic fragment with $M = 111401$ bases consisting of mostly coding regions (it contains more that 80% of coding regions). We build up a random walk trajectory in the x-space with the following prescription [5]. The site position t is interpreted as "time". The walker $x(t) = x(t-1) + \xi(t)$ takes a step up ($\xi(t) = +1$) or down ($\xi(t) = -1$) for each pyrimidine and purine, respectively, at time t. Thus a DNA sequence becomes equivalent to a single trajectory from which we have to derive many distinct trajectories as we shall show below. Fig. 1(a) and (b) show the two DNA walks.

The basic tenet of many techniques, currently used to analyze time series, is the detection of scaling [16, 17]. Scaling is the property of diffusion processes relating the space variable x to the time variable t via the key relation $x \propto t^H$. The symbol H stands for Hurst, as a recognition by Mandelbrot of the earlier work of Hurst [17], and is interpreted as a scaling parameter. It has to be pointed out that Mandelbrot's arguments are based on the so-called fractional Brownian motion (FBM), an extension of ordinary Brownian motion to anomalous diffusion. According to the authors of Ref. [18], it is convenient to adopt two distinct symbols, H_H and H_V, to denote the values afforded by the Hurst method and variance, respectively. If the FBM condition applies, it is shown [18] that $H_V = H_H = H$ and the scaling detected by the variance, H_V, is equal to the scaling of the distribution. In this case, the departure from ordinary diffusion is given by $H \neq 1/2$, with no departure, though, from Gaussian statistics. When the FBM condition does not apply, H_V measures *only* the scaling of the variance, when it exists, and it may depart from H_H [18] and from the scaling of the distribution as well. In this case Mandelbrot's argument does not apply and only the DEA can establish the scaling of the distribution [12–14]. It is worth stressing at this stage that *our definition of scaling is given by the asymptotic time evolution of the probability distribution of x, obeying the property*

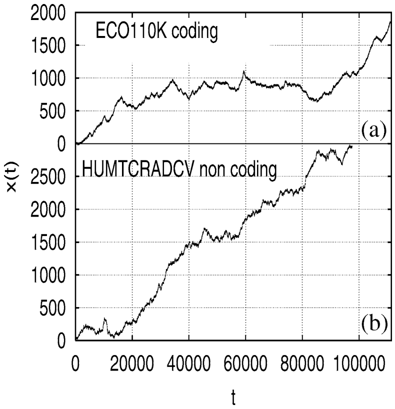

Fig. 1. In (a) we report the DNA walk relative to ECO110K, a coding genomic fragment. In (b) we report the DNA walk relative to HUMTCRADVC, a non-coding chromosomal fragment.

$$p(x,t) = \frac{1}{t^\delta} F\left(\frac{x}{t^\delta}\right), \tag{1}$$

where the symbol δ denotes the scaling exponent of the distribution, which exists also when the second moment of $F(y)$ is divergent.

The diffusing DNA walk trajectories are generated in the following way. For each time t we can construct $M - t + 1$ trajectories of length t,

$$x_j(t) = \sum_{i=j}^{j+t-1} \xi_i, \quad j = 1, 2, \ldots, M - t + 1, \tag{2}$$

where $x_j(t)$ represents the position of the trajectory j at time t. In the variance methods, scaling is studied by direct evaluation of the time behavior of the variance of the diffusion process. If the variance scales, we have

$$\sigma_x^2(t) \propto t^{2H_V}, \tag{3}$$

where H_V is the scaling exponent of the variance. When the asymptotic limit of diffusion process becomes Lévy, the function $p(x,t)$ at large but finite times becomes as close as possible to an ideal distribution with an infinite second moment. In this case, the method of analysis based upon the evaluation of the variance, rather than providing the genuine scaling, affords information on the truncation of the tails generated by the observation at finite times [11].

We note that this choice of trajectories is based on a window of size t, the left side of which moves from the position $j = 1$ to the position $M - t + 1$. The DFA rests on a much smaller number of non-overlapping windows, whose left side is located at the positions $1, t + 1, 2t + 1, \ldots,$ and so on. For any of these non-overlapping windows the DFA considers only the difference between the actual sequence value and a local trend [15]. The DEA [12–14] uses, on the contrary, the overlapping windows of Eq. (2). The choice of the overlapping windows of Eq. (2), in addition to increasing the statistical accuracy of the analysis, fits the rules for the calculation of the Kolmogorov–Sinai (KS) entropy [19,20]. As a matter of fact, the DEA shares with the KS, not only the use of the Shannon entropy indicator, but also the same prescription to convert one single trajectory into a large set of distinct trajectories. The DEA monitors the spreading of the trajectories of Eq. (2), interpreted as the source of entropy increase, to detect scaling, whereas the KS focuses on the entropy increase associated to the random dynamics responsible for the fluctuations of the variables ξ [21]. If the spreading of x-trajectories is independent of biases, if any exist, the DEA determines the scaling associated to this spreading without requiring de-trending, since the scaling is determined by the entropy increase and this is virtually independent of biases.

To evaluate the Shannon entropy of the diffusion process at time t, we partition the x-axis into cells of size $\epsilon = 1$, and we define $S(t)$ as

$$S(t) = -\sum_i p_i(t) \ln\bigl[p_i(t)\bigr], \tag{4}$$

where $p_i(t)$ is the probability that x can be found in the ith cell at time t,

$$p_i(t) \equiv \frac{N_i(t)}{(N - t + 1)}, \tag{5}$$

and $N_i(t)$ is the number of trajectories found in the cell i at a given time t. The connection between $S(t)$ and scaling becomes evident in the continuous approximation, where the trajectories of the DNA walk of Eq. (2) are described by the continuous equation of motion

$$\frac{dx}{dt} = \xi(t). \tag{6}$$

Here $\xi(t)$ is the dichotomous variable assuming the values $+1$ and -1, and t is thought of as a continuous time. In this case, the Shannon entropy reads

$$S(t) = -\int_{-\infty}^{\infty} dx\, p(x,t) \ln\bigl[p(x,t)\bigr] \tag{7}$$

and, after a simple algebra, the earlier illustrated scaling property, Eq. (1), yields

$$S(t) = A + \delta \ln(t), \tag{8}$$

where

$$A \equiv -\int_{-\infty}^{\infty} dy\, F(y) \ln\bigl[F(y)\bigr]. \tag{9}$$

It becomes thus evident why the DE would detect the true scaling even if the process under study were a perfect realization of statistics with divergent second moment. Eq. (8) states that the scaling exponent δ is the slope of the entropy against the logarithmic time.

Let us now consider the two following possibilities:

(1) If $\xi(t)$ is an uncorrelated dichotomous variable, $F(y)$ has a Gaussian form:

$$F_{\text{Gauss}}(y) = \frac{\exp(-y^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}}. \tag{10}$$

The variance scaling exponent is $H_V = 0.5$ and the diffusion entropy of Eq. (8) reads

$$S(t) = \frac{1}{2}\bigl[1 + \ln\bigl(2\pi\sigma^2\bigr)\bigr] + \frac{1}{2}\ln(t). \tag{11}$$

(2) If, instead, $\xi(t)$ has the power-law correlation function $\Phi_\xi(t) \sim 1/t^\beta$, with $0 < \beta < 1$, the distribution density of sojourn times in one of the two states $+1$ or -1, $\Psi_\xi(t)$, is known [11,13] to get the form $\Psi_\xi(t) \sim 1/t^\mu$, with $\mu = \beta + 2$, and the $F(y)$ gets the form of a stable Lévy distribution [22]. The scaling exponent of the probability density function is $\delta = 1/(\mu - 1)$ [11,13] whereas the scaling exponent of the variance is $H_V = (4 - \mu)/2$ [11,13]. The entropy of the diffusion process is

$$S(t) = A_{\text{Levy}} + \frac{1}{\mu - 1}\ln(t). \tag{12}$$

For both cases we expect $S(t)$ to be a linear function of $\ln(t)$, with slope $\delta = 0.5$ that coincides with H_V, and $\delta = 1/(\mu - 1)$ that does not coincide with H_V.

We are now ready to consider the applications to the two DNA sequences. In Fig. 2(a) we show that the DEA of the non-coding sequence HUMTCRADCV results in what apparently seems to be a time-dependent scaling, in conflict with the result provided by the DFA analysis [15], which yields $H_V = 0.615$ throughout the whole time regime. The apparent time dependence of δ is pointed out by means of two straight lines of different slopes: the slope in the short-time regime $\delta = 0.615$, coinciding with the DFA value, while the real asymptotic scaling is $\delta = 0.565$ corresponding to $\mu = 2.77$ (see Eq. (12)). As made evident in the remainder of this Letter, the apparent time dependence of δ is actually the manifestation of an extended regime of transition from "dynamics to thermodynamics", different from that of the coding sequence, but yielding in the asymptotic limit, in both cases, Lévy statistics.

In Fig. 2(b) we consider the more delicate problem of a coding sequence. Here the misleading interpretation in terms of a time-dependent scaling would lead us to conclude that at short times $\delta = 0.52$ and at long times $\delta = 0.67$, thereby implying a kind of transition from normal to anomalous diffusion corresponding to $\mu = 2.5$. We note that our short-time result agrees with the DFA value [15], namely $H_D = 0.51$, but conflicts with the corresponding long-time DFA value, $H_D = 0.75$. Actually, we prove that this conflict is a consequence of the fact that the scaling of the probability density function, Eq. (1), may not be detected by the variance methods like the DFA, whereas the DEA makes it emerge as the asymptotic limit of a long lasting transition from dynamics to thermodynamics, characterized by Lévy statistics, in both coding and non-coding sequences.

To prove this important fact, we model the DNA sequences by adopting the copying mistaken map (CMM) of Ref. [8]. As pointed out more recently [10], this model is equivalent to the generalized Lévy walk (GLW) [5]. The GLW, in turn, fits very well the observation made by the authors of Ref. [15], that the transition to super-diffusion in the long-time region is a manifestation of random walk patches with bias. The CMM corresponds to a picture where nature builds up the real DNA sequence, either coding or non-coding, by using two different sequences. The former is a random sequence (RS) equivalent to assigning to any site the value +1 or −1, with equal probability. The latter sequence, on the contrary, is highly correlated and is obtained as follows. First of all, a sequence of integer numbers $l > 0$ is drawn, with the inverse power law distribution:

$$p(l) = \frac{C}{(T + l)^{\mu}}, \quad 2 < \mu < 3. \tag{13}$$

Any drawing corresponds to fixing the length of a sequence of patches. To any patch is then assigned a sign, either +1 or −1, by tossing a coin. This prescription is virtually the same as that adopted to build up the symbolic sequence of Ref. [23], and corresponds to the intermittent condition of the Manneville map [24,25]. We call this correlated sequence intermittent randomness sequence (IRS). As shown in Refs. [11, 22], the diffusion process generated by the IRS is a

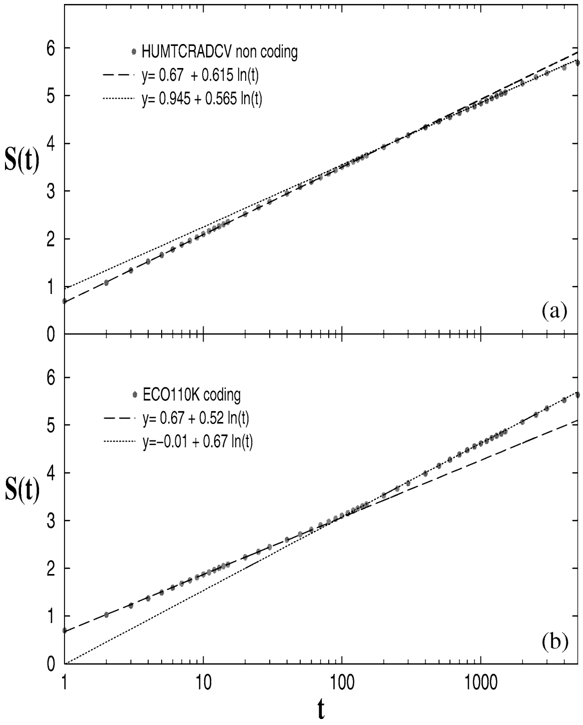

Fig. 2. The diffusion entropy analysis for the two DNA sequences results in a scaling changing with time. For the HUMTCRADCV, the non-coding chromosomal fragment, the slope of the straight line is $\delta = 0.615$ at short-time regime and $\delta = 0.565$ at long-time regime. For ECO110K, the coding genomic fragment, slopes are $\delta = 0.52$ at short-time regime and $\delta = 0.67$ at long-time regime.

Lévy diffusion. According to the CMM, nature builds up the real DNA sequence by adopting for any site of the real sequence the nucleotide occupying the same site in the RS, with probability p_R, or the corresponding one of the IRS, with probability $p_L = 1 - p_R$. The same prescription is used for modeling both the coding and non-coding DNA sequences, the only difference being the different value of p_R/p_L, i.e., the ratio of the uncorrelated to the correlated weight. We note also that the coding DNA sequence is characterized by $p_R \gg p_L$. The Lévy diffusion is faster than ordinary diffusion, and therefore is expected to become predominant, and so ostensible at long times, even when $p_R \gg p_L$. Of course, upon increase of p_R, Lévy statistics become ostensible at longer and longer times. As shown in Figs. 3(a) and (b), the DEA of HUMTCRADCV and ECO110K is perfectly reproduced by a CMM with $\mu = 2.77$ and $\mu = 2.5$, respectively. For the coding sequence $p_R = 0.943$, i.e., the random component is predominant, while for the non-coding sequence $p_R = 0.560$. It is worth to notice that with such values of p_R the CMM also accounts for the correct slope of $S(t)$ vs. $\ln(t)$ in the short-time regime.

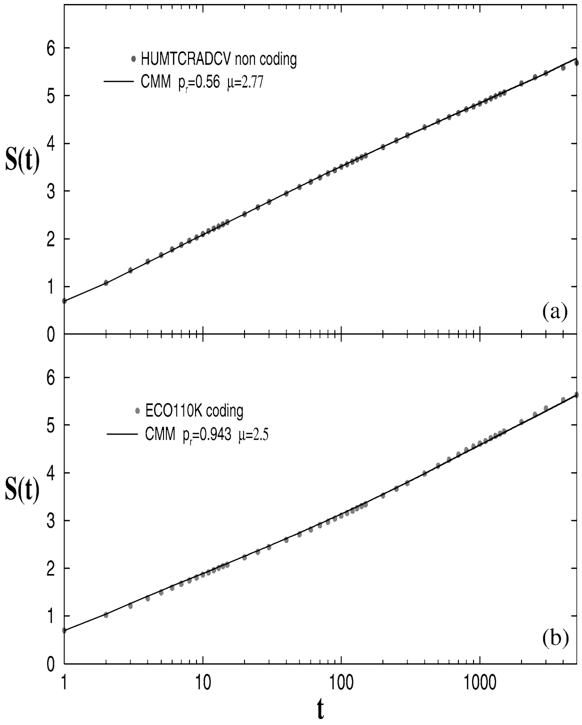

Fig. 3. CMM simulation of the two DNA sequences. (a) shows the comparison between the DE analysis of HUMTCRADCV and an artificial sequence corresponding to the CMM model with $p_R = 0.56$, $T = 0.43$, $\mu = 2.77$. (b) shows the comparison between the DE analysis of ECO110K and an artificial sequence corresponding to the CMM model with $p_R = 0.943$, $T = 45$, $\mu = 2.5$.

Finally, we want to prove the crucial DEA property: The DEA detects the scaling exponent δ of the probability density function, rather than the second moment scaling H_V. In the Lévy walk case the variance may be evaluated [11,13], and the scaling exponent δ and variance exponent H_V are related the one to the other by

$$\delta = \frac{1}{3 - 2H_V}. \tag{14}$$

We see that in the case of the non-coding sequence the DEA yields an asymptotic scaling which is slightly smaller than the short-time scaling. This corresponds to the transition from the short-time Gaussian condition to the long-time Lévy condition, namely to the transition from $\delta = H_V = 0.61$, at short times, to the value $\delta = 1/(\mu - 1) = 0.565$ of the Lévy regime, at long times, with delta related now to $H_V = 0.61$ by Eq. (14). In the coding case we see that the scaling detected by the DEA method is $\delta = 0.67$ that again is related to $H_V = 0.75$ through Eq. (14). Finally, in Table 1 we show the values of the scaling exponents H_V and δ for a set of different coding and non-coding sequences, and the values δ_H evaluated by using the Lévy relation (14). The measured scaling exponent δ coincides with the evaluated exponent δ_H in all cases.

In conclusion, this Letter affords two important results. The first is the discovery of a way to detect the scaling of the probability density function that may be different from the variance scaling detected by the DFA. As a second result, we prove, with the help of artificial sequences generated by the CMM and with the joint use of the DEA, that measures δ, and the DFA that measures H_V, that in the long-time limit of both coding and non-coding sequences, the emerging anomalous diffusion belongs to the Lévy basin of attraction. In other words, in accordance with the expectation of [10], both non-coding and coding DNA sequences, though after a different transient process, reach the same stable “thermodynamic” regime, characterized by Lévy statistics.

Table 1
Values of the scaling exponents H and δ for a set of different coding and non-coding sequences

	N	H_V	δ_H	δ
Non-coding				
HUMTCRADCV	97630	0.61	0.56	0.56
CELMYUNC	9000	0.71	0.63	0.635
CHKMYHE	31109	0.78	0.69	0.70
DROMHC	22663	0.72	0.64	0.65
HUMBMYHZ	28437	0.58	0.54	0.54
Coding				
ECO110K	111401	0.74	0.66	0.66
ECOTSF	91430	0.74	0.66	0.66
LAMCG	48502	0.85	0.77	0.76
CHKMYHN	7003	0.74	0.66	0.66
DDIMYHC	6680	0.68	0.61	0.61
DROMYONMA	6338	0.69	0.62	0.64
HUMBMYH7CD	6008	0.63	0.57	0.58
HUMDYS	13957	0.69	0.62	0.62

In the first column we report the Gen-Bank name of the sequence [1], and in the second column the length N of the sequence. For all measures the error is ± 0.01. δ_H in the fourth column is the theoretical value for δ if the Lévy condition applies, Eq. (14). If the length of the genome is larger than 20,000, the fitted region is $100 < l < 2000$. If the length of the genome is shorter than 20,000, the statistics are not very good for large l. In this case, the fitted region is $20 < l < 200$.

References

[1] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/.

[2] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168.

[3] W. Li, Int. J. Bifurc. Chaos Appl. Sci. Eng. 2 (1992) 137;
W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655;
W. Li, T. Marr, K. Kaneko, Physica D 75 (1994) 392.

[4] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.

[5] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 47 (1993) 4514.

[6] A.K. Mohanti, A.V.S.S. Narayana Rao, Phys. Rev. Lett. 84 (2000) 1832.

[7] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J.F. Muzy, A. Arneodo, Phys. Rev. Lett. 86 (2001) 2471.

[8] P. Allegrini, M. Barbi, P. Grigolini, B.J. West, Phys. Rev. E 52 (1995) 5281.

[9] P. Allegrini, M. Buiatti, P. Grigolini, B.J. West, Phys. Rev. E 57 (1998) 4588.

[10] P. Allegrini, M. Buiatti, P. Grigolini, B.J. West, Phys. Rev. E 58 (1998) 3640.

[11] P. Allegrini, P. Grigolini, B.J. West, Phys. Rev. E 54 (1996) 4760.

[12] P. Grigolini, D. Leddon, N. Scafetta, Phys. Rev. E 65 (2002) 046203.

[13] N. Scafetta, P. Grigolini, cond-mat/0202008.

[14] N. Scafetta, P. Hamilton, P. Grigolini, Fractals 9 (2001) 193.

[15] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E 49 (1994) 1685.

[16] B.B. Mandelbrot, Fractal Geometry of Nature, W.H. Freeman, New York, 1988.

[17] J. Feders, Fractals, Plenum, New York, 1988.

[18] A. Montagnini, P. Allegrini, S. Chillemi, A. Di Garbo, P. Grigolini, Phys. Lett. A 244 (1998) 237.

[19] C. Beck, F. Schlögl, Thermodynamics of Chaotic Systems, Cambridge University Press, Cambridge, 1993.

[20] J.R. Dorfman, An Introduction to Chaos in Nonequilibrium Statistical Mechanics, Cambridge University Press, Cambridge, 1999.

[21] V. Latora, M. Baranger, Phys. Rev. Lett. 82 (1999) 520.

[22] M. Annunziato, P. Grigolini, Phys. Lett. A 269 (2000) 31.

[23] M. Buiatti, P. Grigolini, L. Palatella, Physica A 268 (1999) 214.

[24] P. Gaspard, X.J. Wang, Proc. Natl. Acad. Sci. USA 85 (1988) 4591.

[25] G. Aquino, P. Grigolini, N. Scafetta, Chaos Solitons Fractals 12 (2001) 2023.