Correlations in binary sequences and a generalized Zipf analysis

Andras Czirók,[1,2] Rosario N. Mantegna,[1,3] Shlomo Havlin,[1,4] and H. Eugene Stanley[1]
[1] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2] *Department of Atomic Physics, Eötvös University, Puskin u. 5-7, Budapest, 1088 Hungary*
[3] *Dipartimento di Energetica ed Applicazioni di Fisica, Università, di Palermo, Palermo, I-90128, Italy*
[4] *Department of Physics, Bar-Ilan University, Ramat-Gan, Israel*

(Received 12 January 1995)

We investigate correlated binary sequences using an n-tuple Zipf analysis, where we define "words" as strings of length n, and calculate the normalized frequency of occurrence $\omega(R)$ of "words" as a function of the word rank R. We analyze sequences with short-range Markovian correlations, as well as those with long-range correlations generated by three different methods: inverse Fourier transformation, Lévy walks, and the expansion-modification system. We study the relation between the exponent α characterizing long-range correlations and the exponent ζ characterizing power-law behavior in the Zipf plot. We also introduce a function $P(\omega)$, the frequency density, which is related to the inverse Zipf function $R(\omega)$, and find a simple relationship between ζ and ψ, where $\omega(R) \sim R^{-\zeta}$ and $P(\omega) \sim \omega^{-\psi}$. Further, for Markovian sequences, we derive an approximate form for $P(\omega)$. Finally, we study the effect of a coarse-graining "renormalization" on sequences with Markovian and with long-range correlations.

PACS number(s): 05.40.+j

I. INTRODUCTION

Two topics of current research in statistical mechanics are long-range correlations and Zipf analysis. Stochastic processes with long-range power law correlations have been observed in many systems. Examples are critical phenomena [1], Lévy walks [2], DNA sequences [3–5], heartbeat intervals [6], natural languages [7], and fractional stochastic processes [8].

Complex systems have been studied using the Zipf analysis, originally introduced in the context of natural languages [9]. In conventional Zipf analysis, one calculates the normalized occurrence ω of each word in a given text, and assigns a rank R to each word, with $R = 1$ being the most frequent, $R = 2$ the second most frequent, and so on. A Zipf plot is a log-log plot of the function $\omega(R)$, and for natural languages approximates a straight line of slope roughly -1. Zipf analysis has also been extended to other systems [10], such as the distribution of city sizes [11], DNA base pair sequences [12], and the size distribution of industrial firms [13,14].

In this paper we investigate the relation between correlations in binary sequences ("texts") and a modification of Zipf analysis termed n-tuple Zipf analysis. Unlike conventional Zipf analysis, the words are n-tuples, i.e., strings of length n. We study model sequences with both Markovian (short-range correlated) and long-range-correlated sequences generated by three methods of current interest [15–17].

In Sec. II we discuss Zipf analysis and introduce the quite different n-tuple Zipf analysis. In Sec. III we introduce a function, the frequency density, which is related to the inverse Zipf function. Section IV is devoted to the study of Markovian sequences. Section V applies the n-tuple Zipf analysis to long-range correlated sequences. In Sec. VI we discuss the effects of a coarse-graining "renormalization," while in Sec. VII we study the relation between ζ and the long-range correlation exponent α.

II. N-TUPLE ZIPF ANALYSIS

For texts of natural languages, a basic unit is defined: the word. A word is a string of characters between two separators (usually "space" and/or punctuation marks). For all the nontechnical natural languages studied, the histogram of word occurrence ω vs rank R decreases approximately as a power law:

$$\omega(R) \sim R^{-\zeta}, \tag{1}$$

with an exponent $\zeta \approx 1$ [9]. Since there exist "texts" (long strings of characters carrying information) that are not comprised of *natural* words, it is of interest to modify the Zipf analysis, by defining a "word" to be an n-digit-long string of the text. In this case the set of possible words is finite, e.g., for binary sequences, there exist $N = 2^n$ different n-tuples.

To carry out this n-tuple Zipf analysis, we move a window of length n along the sequence, one character at a step [18], and record the occurrence of each n-tuple. We calculate the normalized frequency of occurrence $\omega(R)$, where

$$\sum_{R=1}^{N} \omega(R) = 1. \tag{2}$$

In the case of a long unbiased sequence, where each symbol is an independent random variable, all possible n-tuples approach the same frequency $1/N$, so the Zipf plot

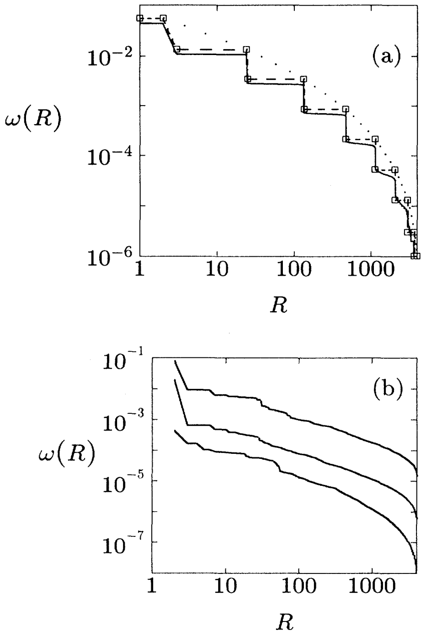

FIG. 1. (a) Zipf plot (frequency vs rank) of a Markovian sequence with transition probability $\varrho(1,1) = 0.80$. Solid line: values obtained from numerical calculations, dashed line with squares: calculated using Eq. (6). (The steps are due to the fact that ω is determined by the number of consecutive digit pairs with both digits different, hence many words have the same frequency of occurrence.) Note that $\omega(R)$ is normalized by $\sum_R \omega(R) = 1$. (b) Zipf plots of long-range correlation sequences generated by three different methods having the same exponent $\alpha = 0.80 \pm 0.02$. From the top to the bottom we have IFT, Lévy, and EMS sequences; for purposes of clarity, the three curves are offset by a constant amount. Similar power-law behavior is observed in the Zipf plot in the interval $(10, 300)$ for all the curves.

becomes horizontal and $\zeta = 0$.

Markovian and long-range correlated (LRC) sequences have an interesting profile. In Fig. 1 we show Zipf plots for Markovian and long-range correlation sequences, with $n = 12$ (so the total number of the words is $N = 2^{12} = 4096$) [19]. For the Markovian case, a steplike plot is found, whereas for the long-range correlation case, the Zipf plot displays approximate power-law behavior in the interval $10 \lesssim R \lesssim 300$.

III. FREQUENCY DENSITY AND THE INVERSE ZIPF FUNCTION

A second useful quantity can be calculated from the analysis outlined above: the frequency density function $\mathcal{P}(\omega)$, where $\mathcal{P}(\omega) d\log\omega$ is the probability of finding an n-tuple with logarithmic frequency between $\log\omega$ and $\log\omega + d\log\omega$. The frequency density $\mathcal{P}(\omega)$ and the inverse Zipf function $R(\omega)$ are related by

$$R(\omega) = N \int_{\omega}^{0} \mathcal{P}(\omega') d\log\omega'. \tag{3}$$

In general, $\mathcal{P}(\omega)$ is not monotonic. However the tail of $\mathcal{P}(\omega)$ is related to the function $\omega(R)$, since from (3), we note that if $\omega(R) \sim R^{-\zeta}$, then $\mathcal{P}(\omega) \sim \omega^{-\psi}$, when the exponents ζ and ψ are related via

$$\psi = 1/\zeta. \tag{4}$$

Our numerical simulations show that the tails of this distribution are well approximated by power law for long-range correlated sequences. Although $\omega(R)$ and $\mathcal{P}(\omega)$ are mathematically related, $\mathcal{P}(\omega)$ is of theoretical interest because it does not require the concept of rank. Moreover, $\mathcal{P}(\omega)$ can be approximated analytically for the case of Markovian texts, as we shall see in the following section.

IV. UNBIASED MARKOVIAN SEQUENCES

First we investigate the Zipf function $\omega(R)$ for an unbiased binary Markovian sequence. Denote the digit in position i of the sequence by d_i, where d_i can have the values 0 or 1. For the simplest Markovian sequence, the probability distribution of digit d_i is determined only by digit d_{i-1}. We denote by $\varrho(u,v)$ the conditional probability that a certain digit v follows another digit u. For unbiased binary sequences $\varrho(0,0) = \varrho(1,1) = p$ and $\varrho(1,0) = \varrho(0,1) = q$, with $p + q = 1$. If each digit is an independent random variable, then $p = 1/2$, while if $p > 1/2 > q$ short-range Markovian correlations are present. The probability $p_i(v)$ that the digit d_i is v is given by

$$p_i(v) = \varrho(u,v) p_{i-1}(u) + \varrho(v,v) p_{i-1}(v). \tag{5}$$

To calculate the Zipf function $\omega(R)$, we first calculate the frequency ω of a given n-tuple, which depends only on the number k of consecutive digit pairs in that word with both digits different:

$$\omega_k = \tfrac{1}{2} q^k p^{n-1-k} \qquad [k = 0, 1, \ldots, n-1]. \tag{6}$$

The number of such words is

$$\mathcal{N}_k = 2\binom{n-1}{k}. \tag{7}$$

For correlated sequences, it follows from (6) that $\omega_k > \omega_{k+1}$. The ranks of the $\mathcal{N}_k$ words occurring with a frequency ω_k are in the interval

$$\left[\sum_{j=0}^{k-1} \mathcal{N}_j,\ \sum_{j=0}^{k} \mathcal{N}_j\right]. \tag{8}$$

Figure 1(a) compares the results of a single simulation of a sequence comprising $L = 10^6$ digits for words of length $n = 12$ with the exact results of Eqs. (6)–(8). The probability $\mathcal{P}_k \equiv \mathcal{N}_k/N$ of finding a word with a frequency ω_k is

$$\mathcal{P}_k = \binom{n-1}{k} 2^{-(n-1)}. \tag{9}$$

We next introduce a new parameter $-1 \leq \kappa \leq 1$ defined by

$$\kappa \equiv \frac{2k}{n-1} - 1, \tag{10}$$

and express k and $n-1-k$ in terms of κ:

$$k = \frac{n-1}{2}(1+\kappa) \quad \text{and} \quad n-1-k = \frac{n-1}{2}(1-\kappa). \tag{11}$$

Substituting (11) into (9), and applying Stirling's approximation for $n \gg 1$ and $\kappa \ll 1$, we find

$$\log \mathcal{P}(\kappa) = -\frac{n-1}{2}\kappa^2 + O(\kappa^4). \tag{12}$$

To relate $P(\kappa)$ to $\mathcal{P}(\omega)$, we substitute (11) into Eq. (6) and find that $\log \omega$ is linear in κ,

$$\log \omega_\kappa = \frac{n-1}{2}\left[\log q + \log p + \kappa(\log q - \log p)\right] - \log 2, \tag{13}$$

Thus the approximate probability density $\mathcal{P}(\omega)$ of Markovian sequences on a double-logarithmic plot is a parabola [20]. In Fig. 2(a) we show Eq. (12) (solid line) together with the values obtained from a simulated sequence of length 10^6 digits. In our simulations, we varied p in the interval (0.5,0.9). The agreement between the approximation and numerical simulations is quite good. From numerical simulations we note that Markovian sequences did not show power-law behavior in the wings of the $\mathcal{P}(\omega)$ function. Since $\mathcal{P}(\omega)$ is not a power law, Eq. (3) implies that $R(\omega)$ [and hence $\omega(R)$] is also not a power law [21].

V. LONG-RANGE CORRELATED SEQUENCES

We generate long-range correlation sequences by using three different algorithms, each of which is described in Appendix A:

a. Inverse Fourier transformation (IFT) [3,15,22–24],

b. Lévy walks [16], and

c. The expansion-modification system (EMS) [17].

Long-range correlated sequences are characterized by the correlation exponent α [6]. We measure α by calculating the average width $w(\ell)$ of a digitized walk in a window of length ℓ, and using the scaling relation $w(\ell) \sim \ell^\alpha$ [3]. In Appendix A, we also show that the IFT and Lévy methods yield sequences of real numbers $x(t)$ sampled with a fixed interval Δt, from which we obtain the cor-

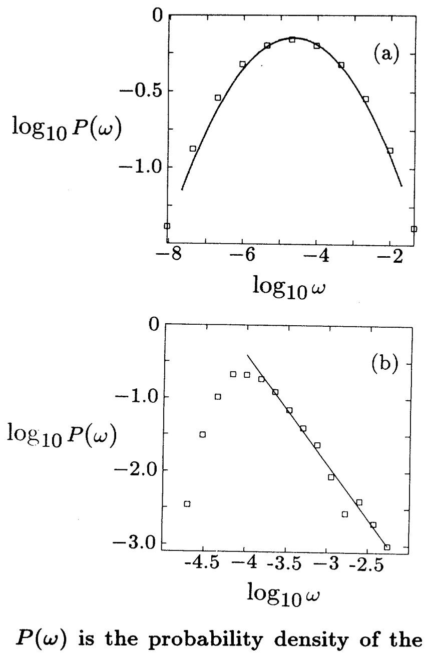

FIG. 2. $P(\omega)$ is the probability density of the number of words with a frequency in the interval $\log \omega$ and $\log \omega + d \log \omega$. (a) For a Markovian sequence: the solid line is the parabola given by Eqs. (12) and (13) while the squares represent the measured values. (b) For a long-range correlation sequence generated by IFT with $\alpha = 0.74$ power-law decay is observed in the tail. The straight line is the best fit in the interval $(10^{-4}, 10^{-2})$, its inverse slope is $1/\chi = 0.66$, close in the measured value $\zeta = 0.64$, in agreement with Eq. (2) (see Fig. 3).

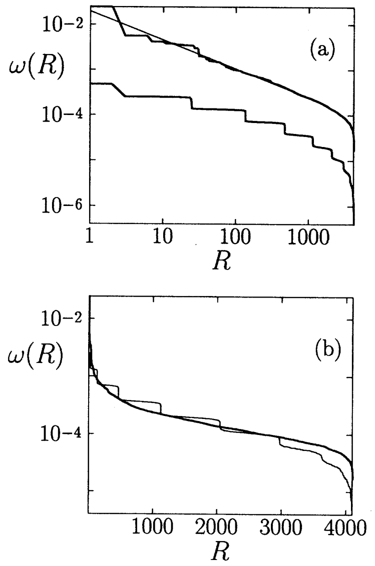

FIG. 3. Comparison of the Zipf plot of a long-range correlation (IFT) sequence with $\alpha = 0.74$ with its Markovian approximation where $\varrho(1,1) = 0.65$ (shifted to allow comparison). (a) Double logarithmic scale: A power-law decay over 3 decades with $\zeta = 0.64$ is observed in the case of long-range correlation sequences. The straight line is the best fit in the interval (1,1000). For better visualization the Markovian plot is shifted down by one decade. (b) Log-linear scale: Differences between the long-range correlated (smooth line) and the Markovian (jagged line) sequence are more relevant for the most and least frequent words only.

responding binary sequences d_i using

$$d_i = \begin{cases} 1 & \text{if } x((i+1)\Delta t) \geq x(i\Delta t), \\ 0 & \text{if } x((i+1)\Delta t) < x(i\Delta t). \end{cases} \tag{14}$$

In Fig. 3, the Zipf plot of a long-range correlation sequence generated by IFT is compared with its Markovian approximation, i.e., a Markovian sequence having the same conditional probabilities $\varrho(i,j)$ as the long-range correlation sequence possesses. Numerical simulations show that the long-range correlation sequence is better fit by a power law than the Markovian approximation.

We also found a power-law behavior in the wings of the frequency density $\mathcal{P}(\omega)$ [Fig. 2(b)], and the measured values of the fitted slopes ψ satisfy Eq. (4). However, if we examine the two functions on a log-linear scale [Fig. 3(b)], we notice that most of the Zipf plot of the long-range correlation sequence is determined by the Markovian probabilities alone, and the long-range correlated nature influences mainly the frequency of the most-frequent and least-frequent words.

VI. COARSE GRAINING

In order to better distinguish between long-range correlated and Markovian sequences, we suggest the following renormalization procedure. We perform a series of coarse-grainings on the original d_i sequence. In each course graining step we replace the triplets $\{111, 110, 101, 011\}$ with "1" and the triplets $\{000, 001, 010, 100\}$ with "0" ("majority rule").

We expect that for scale-invariant structures (such as power law long-range correlated sequences), the statistical properties—including the Zipf plot—should not change after the application of coarse graining. In contrast, for Markovian sequences this renormalization procedure eliminates the short-range correlations. In fact, we show in Appendix B that the $\varrho(u,v)$ transition probabilities of the Markov matrix of the renormalized sequences converge to $1/2$ after a few coarse-graining steps.

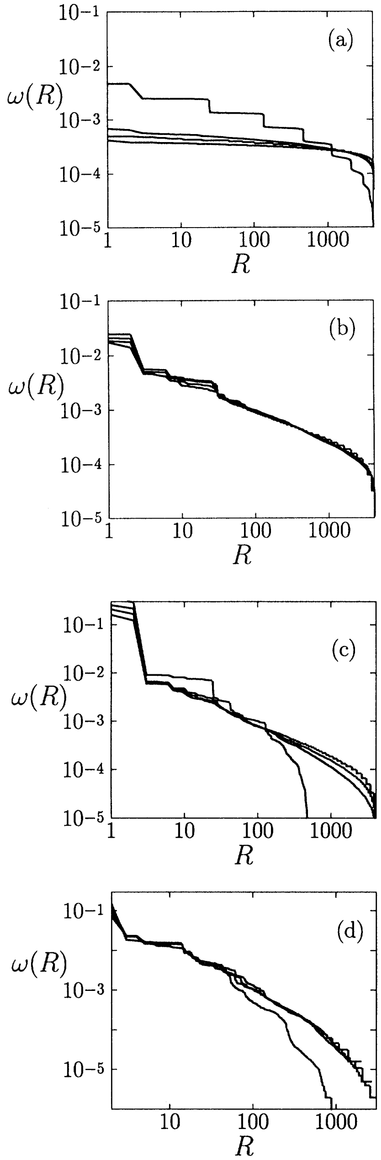

FIG. 4. Zipf plots obtained by consecutive renormalizations. (a) Markovian sequence with $\varrho(1,1) = 0.65$, the coarse graining destroys the correlations and the curve converges to a horizontal line, corresponding to the Zipf plot of a random sequence that is uncorrelated and unbiased. If the sequence is long-range correlated [IFT, $\alpha = 0.74$ (b); Lévy, $\alpha \approx 0.9$ (c); and EMS, $\alpha = 0.87$ (d)], the renormalization procedure leads to a power-law Zipf plot. In the cases of (c) and (d) strong short-range correlations are present in the original sequences, so that the power-law behavior of the Zipf plots can be observed only after a few coarse-graining steps.

In Fig. 4 we show typical Zipf plots after the numerical renormalization procedure. The Zipf plots of the coarse-grained Markovian sequence approach a horizontal line [Fig. 4(a)], while the long-range correlation sequences show stable slopes after 2–3 renormalization steps [Figs. 4(b)–4(d)]. Note, that in Figs. 4(c) and 4(d) a power-law behavior is observed only after the first coarse-graining step. This observation reflects a peculiarity of the Lévy and EMS sequences, for which a strong short-range correlation is also present, which is already destroyed almost completely after the first coarse-graining step.

To characterize the deviation from the pure power-law behavior, we study (Fig. 5) a "local exponent" $\alpha_L(\ell)$, which is the logarithmic derivative of the width $W(x)$:

$$\alpha_L(\ell) \equiv \left(\frac{d \log W(x)}{d \log x} \right)_{x=\ell} . \tag{15}$$

For an ideal long-range correlation sequence $\alpha_L(\ell) = \alpha = \text{const}$. For Lévy sequences, however, $\alpha_L(\ell)$ for small ℓ ($\ell \approx 10$) is larger than the asymptotic value [see Fig. 5(b)]. Due to this strongly correlated behavior at short range, some words do not occur in the finite sequence studied, thereby strongly affecting the Zipf plot. After a few coarse grainings the asymptotic behavior becomes dominant, and the Zipf plots [Fig. 4(c)] show a power-law behavior characterized by a constant ζ.

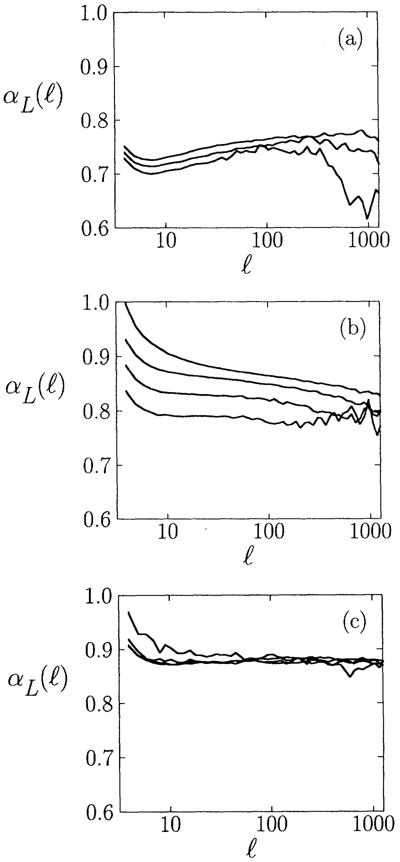

FIG. 5. Local α as a function of the window length ℓ for different long-range correlated sequences: (a) IFT, (b) Lévy, and (c) EMS. Different lines refer to consecutive renormalization steps. The initial value of α_L (low values of ℓ) is affected by the strong short-range correlations in (b) and (c). The asymptotic value of α_L is affected by the finite size of the sequence in (b).

VII. CONNECTION BETWEEN ζ AND α

Next we investigate the relationship between the Zipf exponent ζ and the long-range correlation exponent α. In the Zipf analysis, we used a word length of 12 bits, which provided large enough scaling regime in the Zipf plot, and did not require extremely long sequences to minimize finite-size effects. To measure ζ, we fitted with power law the frequency vs the rank on different intervals, typically between 3 and 300. The difference between the values measured on the same sequence but on different fitting intervals gives an estimate for the error of the exponent, which is about 10%. In all cases, we confirmed that that the lengths of our sequences are long enough to exclude finite-size effects. The typical length of the sequences used in the measurement was 3×10^6–10^7 digits and we could observe relevant finite-size effects only below the length scale of 10^4 digits [19]. The results are shown in Fig. 6, where we plot the values of the exponent ζ vs α.

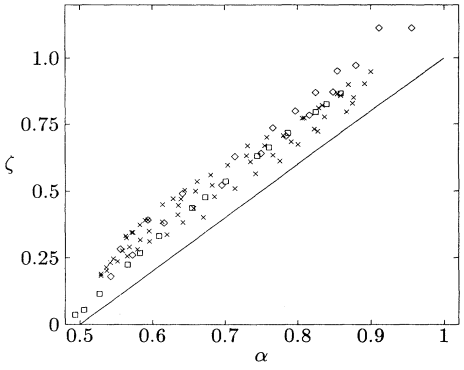

FIG. 6. Numerical investigation of the connection between the exponents α and ζ for long-range correlation sequences (IFT □, Lévy ×, and EMS ◇), applying one or more renormalization steps. The estimated error of the exponents ζ is about 10%. The solid line represents the conjectured "bound" $\zeta = 2\alpha - 1$.

VIII. CONCLUSIONS

The numerical study of the relation between α and ζ in sufficient long-range correlated binary sequences seems to suggest a simple relation between these two parameters. However, the correlation exponent α and the Zipf exponent ζ in long-range correlated binary sequences provide information on quite different scales: α is obtained by investigating the scaling properties over the entire length of the sequence, whereas ζ is obtained by investigating a "short-range" property, the frequency of n-tuples. Our study shows that the long-range correlation exponent α may be related to the exponent ζ measured by the n-tuple Zipf analysis, provided corrections-to-scaling terms are very small so that $\alpha_L \approx \alpha$ over a wide range of scales.

The literature has addressed the problem that the power law observed by performing Zipf analysis can be shallow [10,25]. In fact, a trivial power law is observed in random text if one selects a character as "space" and performs the Zipf analysis [25]. However, this conclusion does not apply to our n-tuple word definition. In our analysis, we do not have a "space" and the n-tuple Zipf analysis of an unbiased random sequence gives $\zeta = 0$.

On the other hand, an n-tuple Zipf analysis of a Markovian sequence can give results roughly similar to the one observed for long-range correlated sequences. This is due to the fact that short-range correlated sequences can mimic the local behavior of the long-range correlated sequences. However by introducing a coarse-graining procedure of the binary texts we are able to distinguish between Markovian and long-range correlation texts. In fact only for long-range correlation sequences is the power-law behavior observed after several coarse-grainings.

ACKNOWLEDGMENTS

We thank A.-L. Barabasi, S. V. Buldyrev, A. L. Goldberger, C.-K. Peng, and especially T. Vicsek for sharing their ideas in many stages of this work. This work was supported by Grants from Foundation Pro Cultura, NSF and NIH.

APPENDIX A: METHODS OF GENERATION LONG-RANGE CORRELATION SEQUENCES

1. Inverse Fourier transform (IFT)

A sequence of real numbers $x(t)$ is generated by inverse Fourier transforming a sequence of complex numbers

$$u(f) = |f|^{-\beta/2}\eta(f), \tag{A1}$$

where $\eta(f)$ is a Gaussian stochastic noise of amplitude A obeying the conditions

$$\langle \eta(f) \rangle = 0 \tag{A2}$$

and

$$\langle \eta(f)\eta^*(f') \rangle = A^2\delta(f - f'), \tag{A3}$$

where the symbols $\langle\ \rangle$ indicate averages over different realizations of the noise. The correlation exponent is related to the parameter β through the relation

$$\alpha = (1 + \beta)/2. \tag{A4}$$

2. Lévy walk

Lévy walks show also long-range correlated behavior on a sufficiently large time (length) scale. The general form of the probability density of a jump of length r in the time interval $(t, t + dt)$ in a Lévy walk is

$$\phi(r, t) = Cr^{-\mu}\delta(r - t^{\nu}), \tag{A5}$$

where t is the time needed to perform a jump to the distance of r. We have to convert the $x(t)$ coordinates of the Lévy walk into binary sequences, so we set $\nu = 1$ to have a constant velocity, and discretized both time and the possible length of the jumps. The resulting $x'(t)$ walk can be identified with the walk of the binary sequence d_i. It is possible to tune α in the range of (0.55,0.9) varying μ in the interval $(2, 3)$.

3. Expansion-modification system (EMS)

Li's expansion-modification system is especially suitable for our purposes as it provides binary sequences and we do not have to face the digitization problem. We applied the following recursive rule to build up binary strings (discussed by Li in detail): In one step we substitute each digit "1" by the string "00" or "11" with probabilities p and $1 - p$; while each digit "0" by strings "10" or "00" with probabilities q and $1 - q$, respectively. As we consider unbiased sequences only, this yields to the constraint $2q = p$. Using this method we were able to generate sequences with a typical length of 8 million digits, tuning α with q in the range from $\alpha \approx 1$ (where $q \approx 1$) to $\alpha \approx 0.5$ (where $q \approx 0.25$).

APPENDIX B: RENORMALIZATION PROCEDURE ON MARKOVIAN SEQUENCES

We show that the conditional probabilities $\varrho(u, v)$ of the renormalized sequences converges to 1/2 during the consecutive coarse-graining steps using a somewhat simpler rule than the majority rule used in the text, viz., replacing the doublets $\{11, 10\}$ with "1" and the doublets $\{00, 01\}$ with "0." We denote by prime the appropriate variables for the renormalized sequence, and by $p_{k,k+1}(u, v)$ the probability having a digit u and v at position k and $k + 1$, respectively. By definition the Markovian probability of the coarse-grained sequence is

$$\varrho'(1, 1) = \frac{\langle p'_{k,k+1}(1, 1)\rangle_k}{\langle p'_k(1)\rangle_k}. \tag{B1}$$

As we consider unbiased sequences (this symmetry is kept by the renormalization rules), $\langle p'_k(1)\rangle_k = 1/2$. Using the coarse-graining rules we can express $\langle p'_{k,k+1}(1, 1)\rangle_k$ by the appropriate word frequencies of the original $d(i)$ sequence:

$$\langle p'_{k,k+1}(1, 1)\rangle_k = \omega_{1111} + \omega_{1110} + \omega_{1011} + \omega_{1010}, \tag{B2}$$

where ω_{ijkl} represents the frequency of the word "$ijkl$." If $p = \varrho(1, 1)$ is given, then we can easily calculate these probabilities:

$$\langle p'_{k,k+1}(1, 1)\rangle_k = \tfrac{1}{2}\left(p^3 + p^2q + pq^2 + q^3\right), \tag{B3}$$

where $q = 1 - p$. Writing both p and p' in the form of $1/2 + \epsilon$ and $1/2 + \epsilon'$, respectively, we can derive the recursion rule for ϵ:

$$\begin{aligned}\tfrac{1}{2} + \epsilon' = {}& \left(\tfrac{1}{2} + \epsilon\right)^3 + \left(\tfrac{1}{2} + \epsilon\right)^2\left(\tfrac{1}{2} - \epsilon\right) \\ & + \left(\tfrac{1}{2} + \epsilon\right)\left(\tfrac{1}{2} - \epsilon\right)^2 + \left(\tfrac{1}{2} - \epsilon\right)^3\end{aligned} \tag{B4}$$

which simplifies to

$$\epsilon' = 2\epsilon^2. \tag{B5}$$

This recursion leads to $\lim_{n\to\infty}\epsilon^{(n)} = 0$, and the convergence is faster than exponential.

[1] M. Cassandro and G. Jona-Lasinio, Adv. Phys. **27**, 913 (1978).

[2] M. F. Shlesinger, G. M. Zaslavsky, and J. Klafter, Nature (London) **363**, 31 (1993).

[3] C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992); C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger,

Phys. Rev. E **49**, 1685 (1994). A variant of the "DNA walk" we also used was introduced in a different context by M. Ya. Azbel, Phys. Rev. Lett. **31**, 589 (1973).

[4] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[5] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992); **71**, 1777 (1993); S. V. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, F. Sciortino, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **71**, 1776 (1993); see also the recent GenBank analysis in S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).

[6] C. K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993).

[7] A. Schenkel, J. Zhang, and Y.-C. Zhang, Fractals **1**, 47 (1993); M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, and N. Shnerb, Fractals **2**, 7 (1994); W. C. Ebeling and A. Neiman, Physica A **215**, 233 (1995).

[8] B. B. Mandelbrot and J. W. Van Ness, SIAM Rev. **10**, 422 (1968).

[9] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, New York, 1949).

[10] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983).

[11] M. Gell-Mann, *The Quark and the Jaguar* (Freeman, New York, 1994).

[12] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994). Recent work [Mantegna *et al.* (unpublished)] suggests that the conclusions of this article may hold for the three eukaryotic chromosomes recently submitted to the GenBank (invertebrates), but the evidence is somewhat less conclusive for higher forms of life.

[13] J.-P. Bouchaud, in *Proceedings of 1994 International Conference on Lévy Flights*, edited by M. F. Shlesinger, G. Zaslavsky, and U. Frisch (Springer, Berlin, 1995).

[14] M. H. R. Stanley, S. V. Buldyrev, S. Havlin, R. Mantegna, M. A. Salinger, and H. E. Stanley, Econ. Lett. (in press); M. H. R. Stanley, 1994 Westinghouse report (unpublished); H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and M. H. R. Stanley, in *Proceedings of 1994 International Conference on Lévy Flights* [13].

[15] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993).

[16] G. Zumofen, A. Blumen, J. Klafter, and M. F. Shlesinger, J. Stat. Phys. **54**, 1519 (1989); S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[17] W. Li, Phys. Rev. A **43**, 5240 (1991).

[18] In this way, for a sequence of length L, we obtain a total of $\omega = L - n + 1$ words. If, however, we move the box n characters at a time, then we obtain n different "reading frames," each of which contains $\omega = L/n$ words. In coding DNA, e.g., $n = 3$ and there are three distinct reading frames.

[19] Note that for an n-tuple Zipf analysis with $n = 12$ a simulation of a symbolic text of 10^7 bits is in fact a very large simulation. The reason is that such a system size allows one to anticipate (in a first approximation) an average number of more than 2500 occurrences for each possible n-tuple (since the quotient of 10^7 and 2^{12} is roughly 2500). Also, we examined sequences with a wide range of lengths L, ranging from 10^3 to 10^6. It turned out that if $L > 10^5$, then the uncertainty of ζ is below 3%. This means that the error due to the finite size of the samples is much less than the error caused by the nonperfect power-law scaling. We also tested the effect of varying word length n from 10 to 13 bits. We found that the change in the exponent is smaller than 4% for this range of word length.

[20] We checked the *quality* of our approximations by performing numerical simulations of Markovian processes. One example of these investigations is shown in Fig. 2(a).

[21] Nevertheless, one can quantify the "steepness" of the Zipf plot of Markovian sequences by calculating the slope of a line connecting $\omega(1)$ and $\omega(N/2)$

$$\zeta_M \equiv \frac{\log \omega(1) - \log \omega(N/2)}{\log N/2}.$$

To calculate $\omega(1)$ we note that the two most frequent words consist of the same digits, so $\omega(1) \approx p^n$. Due to the symmetry of the frequency density function of Eq. (12) the words giving the maximum of the probability density function have a rank of $N/2$. Since the maximum is located at $\kappa = 0$, the frequency corresponding to these words, from (13), is $\omega(N/2) \approx p^{n/2} q^{n/2}$. Since $N = 2^n$, we find

$$\zeta_M = \frac{\log p - \log q}{2 \log 2}.$$

For an uncorrelated sequence, $p = q = 1/2$ and $\zeta_M = 0$. The Zipf plot becomes steeper with increasing p. In the $p \to 1$ limit, $\zeta_M \to \infty$, in agreement with our expectations; only one of two words (111...11 or 000...00) are possible, yielding a single peak of height N for the Zipf plot.

[22] H. Makse, S. Havlin, H. E. Stanley, and M. Schwartz, Chaos Solitons Fractals **6**, 295 (1995).

[23] S. Prakash, S. Havlin, M. Schwartz, and H. E. Stanley, Phys. Rev. A **46**, R1724 (1992).

[24] S. Havlin, R. Blumberg-Selinger, M. Schwartz, H. E. Stanley, and A. Bunde, Phys. Rev. Lett. **61**, 1438 (1988).

[25] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992).