

# Markov Processes: Linguistics, Zipf's Law and Long-Range Correlations

I. Kanter and D. A. Kessler

*Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

(October 20, 1994)

It is shown that a 2-parameter random Markov process constructed with  $N$  states and biased random transitions give rise to a stationary distribution where the probabilities of occurrence of the states,  $P(k)$   $k = 1, \dots, N$ , exhibit the following three universal behaviors which characterize biological sequences and texts in natural languages; (a) the rank-ordered frequencies of occurrence of words are given by Zipf's law,  $P(k) \propto 1/k^\rho$ , where  $\rho(k)$  is slowly increasing for small  $k$ , (b) the frequencies of occurrence of letters is given by  $P(k) = A - D \ln(k)$  and (c) long-range correlations are observed over long but finite intervals, as a result of the quasi-ergodicity of the Markov process.

PACS numbers: 87.10.+e, 05.40.+j, 02.50.Ga

Recently an explosion of activity has centered on the statistical nature of biological sequences and texts in various languages. [1] The activities concentrate on the observation that all examined sequences of biological systems, such as DNA or amino-acids chains, [2] or texts in both natural and artificial languages, [3] deviate remarkably from **random** sequences. It is of course not surprising at all that sequences such as texts in English, for instance, are **locally** constrained due to some grammatical rules. However, the fact that local grammatical rules affect the **global** statistical nature of sequences, on large scales, is intriguing for a physicist. These observations are not only interesting in their own right, they have important practical implication. For example, in higher organisms, only a small fraction of the DNA sequence is used for coding proteins, the possible function or the importance of the information content, if any, of the non-coding regimes is unclear. Also, the significance of long-range correlations (LRC) in a chain of letters or words, i.e. a story, has been discussed extensively. [1] In this paper, we will present a simple 2-parameter model which serves to explain and elucidate these remarkable statistical properties.

Observations on the statistical nature of these sequences has led to the claim that they obey (at least) the following three universal laws:

**A. Zipf's law for Words:** A long sequence is defined over a vocabulary, a set of semantic units; words in a natural language and the 64 3-tuples ('triplets') which code the amino acids. The frequencies of occurrence of each semantic unit is calculated and the units are ordered in decreasing rank order,  $P(1) \geq P(2) \geq \dots \geq P(N)$  where  $N$  is the size of the vocabulary. It is found in

many languages that the following approximate scaling behavior exists

$$P(k) = \frac{A}{k^\rho} \quad (1)$$

where  $A$  is a constant and  $\rho$  is estimated to be  $\sim 1.0$  for natural languages [4] and much smaller ( $\sim 0.3$ ) for DNA sequences. [5]

While the general wisdom is that the data fits well to Zipf's law, a more careful analysis reveals significant systematic deviations. The logarithmic derivative,  $\rho \equiv -d \ln(P)/d \ln(k)$ , is in fact non-constant over even relatively small intervals of  $k$ , i.e. intervals extending over no more than one decade. The attempt to fit the data to a generalized Zipf's law, [6]  $P(k) = A/(D+k)^\rho$  is a sign of this deviation from the simple Zipf's law. For illustration, the probabilities of occurrence of words in the Bible are plotted in Fig. 1, where the local logarithmic slope,  $\rho(k)$ , is shown in the insert. It is clear that  $\rho$  is an increasing function of  $k$ , and there is no macroscopic regime or even an entire decade where  $\rho$  is a constant. This situation is in fact the typical one and has been verified for many other natural and linguistic sequences. [7] As a natural consequence of this variation of  $\rho$ , any attempt to fit the data with a single  $\rho$  will be very sensitive to the details of the fitting. It should be also pointed out that the behavior of  $\rho$  is inconsistent with the sharp crossover characteristic of the generalized Zipf's law above.

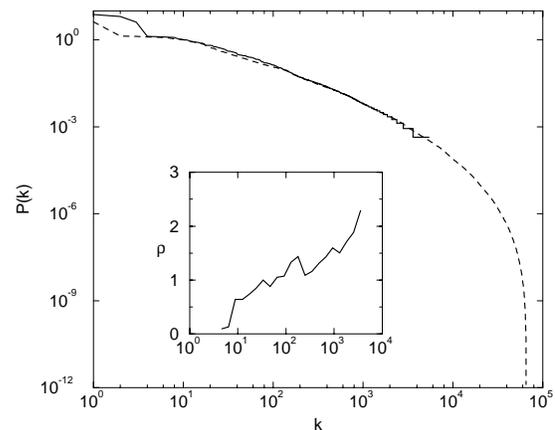


FIG. 1. Rank-ordered frequencies of words,  $P(k)$ , for the Bible (solid line) and for the Markov Process defined in the text with  $L = 16$ ,  $x = B = .92$  (dashed line). Both sets of  $P$ 's are normalized to  $P(10) = 1$ . Inset: Logarithmic derivative  $\rho(l)$  for the Bible data.

One may attempt to attribute these deviations from Zipf's law to, for example, the statistical fluctuations in the data and/or the finite length of the strings studied. However, while Zipf's law is attractive in its simplicity we have no compelling reasons to expect it to hold, much less a theoretical framework in which to analyze deviations. The purpose of this paper is to present such a framework. Our fundamental starting point will be to suggest a simple natural process for the generation of strings. This natural mechanism as we shall see gives rise to a distribution which fits the data better than Zipf's law, and furthermore provides a general framework for explaining other universal features of strings described below.

**B. A law for letters:** A similar ordering of the frequencies of occurrence of letters,  $P(k)$ , over the alphabet is found, for many languages, to follow the universal behavior

$$P(k) = A - D \ln(k) \quad (2)$$

where  $A$  and  $D$  are constants. This behavior holds for biological sequences and was recently found to be correct for over 100 natural languages where the size of the alphabet ranged between 14 – 60. [8]

**C. Long-Range Correlation (LRC):** For each long sequence mapped into a binary string the mean square fluctuation,  $F^2(m)$ , is calculated in the following way

$$F^2(l) = \ll D(l_0, l)^2 \gg - \ll D(l_0, l) \gg^2 \quad (3)$$

where  $D(l_0, l)$  is the sum of the binary string in the window  $[l_0, l_0 + l]$ ,  $\ll \dots \gg$  stands for the average over non-overlapping windows of size  $l$ . The mapping to binary strings is done in the following ways: (a) The four alphabet letters ( $A, C, G$ , and  $T$ ) of DNA, or the 26 letters in English are divided equally at random into two groups which are mapped to '0' and '1'; or (b) the binary representation (ASCII) of the letters is used. A universal behavior was again observed for all examined biological sequences and natural texts,

$$F(l) \sim l^\alpha \quad \alpha > 1/2 \quad . \quad (4)$$

for some intermediate regime  $0 < l \ll l_{max}$ , where  $l_{max}$  is the size of the binary sequence. Note that for random sequence  $\alpha = 0.5$ , where for the strings examined the typical  $\alpha$  is in the range 0.6-0.9.

In the following we will show that it is possible that these three universal laws are fundamentally related, and are a result of the statistical nature of random discrete-time Markov processes (MPs). Conceptually, a MP is the simplest natural algorithm for the stochastic production of strings. Consider the production of a string composed of tokens chosen from among  $N$  possibilities. If the probability distribution for choosing the next token is only a function of the current last one in the string, then the construction can be considered as a MP. We denote the

transition probabilities between tokens, or states,  $i$  and  $j$  as  $W(i, j)$  where, of course,  $\sum_j W(i, j) = 1$ . The probabilities of occurrence of each token in the (infinite) string is given by  $P(1), P(2), \dots, P(N)$ , the stationary ensemble of the MP.

The following discussion will concentrate, for simplicity, on a subset of possible MPs:

(a) Number of states: For simplicity,  $N$  is fixed to be equal to  $2^L$  with  $L$  integer, and each state is identified by an  $L$ -bit binary number between 0 and  $N - 1$ .

(b)  $W$  is a sparse matrix. The number of non-vanishing elements in each row,  $C$ , is finite and does not scale with  $N$ . The particular case which will be exhaustively examined in the following is  $C = 2$ . If the  $P$ 's are considered as dynamical variables, this type of system is known as a highly diluted asymmetric network [9], however with the following two notable differences: (1) the sum of the  $P$ 's is restricted to be unity and (2) the activation function for the update of the  $P$ 's is linear in the MP case.

(c) Meaningful connectivity: The cornerstone of the discussed MPs is that the transition matrix,  $W$ , differs from a random graph. In a language, for instance, the semantic structure guarantees that words are not haphazardly followed by a random selection of other words. Rather, "meaningfulness" strongly constrains the choice of successive words. We attempt to model this fact in the following simple manner: The two states,  $m_0$ ,  $m_1$ , connected to the state  $m$  are given by

$$m_0 = 2m \bmod N \quad ; \quad m_1 = 2m \bmod N + 1 \quad (5)$$

where  $m = 0, 1, \dots, 2^L - 1$ . In words, the  $L - 1$  rightmost bits of state  $m$  are shifted one bit to the left, and the rightmost bit is set equal to either 0 or 1. Thus each successive word is closely related to the one before. Note that under this construction both the outward and inward connectivity of each state are equal to two.

(d) Strength of transition probabilities: In order to reduce the number of free parameters, the two weights, transition probabilities, going out from each state can take only the values  $x$  and  $1 - x$ .

(e) Bias: For each state there are two options for the transition probabilities. We pick  $W(m, m_0) = 1 - x$  and  $W(m, m_1) = x$  with probability  $B$  and vice-versa with probability  $1 - B$ . The bias,  $B$ , is chosen for simplicity to be a constant independent of the state, like a constant external magnetic field. Note that the ensemble of possible MPs is due to the freedom given by  $B$ .

On the **average** the bias  $B$  and the strength  $x$  play the same role, since the average drift towards 1 from each state is  $x_{eff} = xB + (1 - x)(1 - B)$ , which is symmetric in these two parameters. However, the effect of **fluctuations** in  $x$  and  $B$  play a different role, and indeed all three universality behaviors mentioned above measure fluctuations since they are sensitive to the whole distribution of  $\{P(k)\}$ .

The interesting regime in the unit square of  $(B, x)$  is  $(0.5-1, 0.5-1)$ , since the process obeys a global inversion symmetry  $x \rightarrow 1-x$  and  $B \rightarrow 1-B$ , and furthermore changing only  $x \rightarrow 1-x$  or  $B \rightarrow 1-B$  is like changing the role between 0 and 1.

For the case  $x = 0.5$  one can easily show that for any  $B$  the stationary solution is  $P(k) = 1/N$ , independent of  $k$ . In general the analytical solution is difficult since we are interested in the nature of the  $N$  entries of the **eigenvector** of  $\lambda = 1$  which are ordered in a decreasing order, but a priori this order is not known and its combinatorial complexity is  $N!$ . Fortunately, the interesting limit of  $B = 1$  can also be solved analytically. In this limit, from the detailed balance equations and from the overall normalization constraint one can easily deduce that all states which contain  $j$  1's have the same probability

$$q^j / (1+q)^L \quad (6)$$

where  $q = (1-x)/x$ . Hence the distribution  $\{P(k)\}$  is constructed from  $L$  plateaus, where the ratio between the values of two subsequent plateaus is fixed to be  $q$ . The local logarithmic slope,  $\rho(k)$  between plateaus  $j-1$  and  $j$ , where the rank order  $k = \sum_{i=1}^j \binom{L}{i}$ , is given by

$$\rho(k) = -\ln(q) / [\ln(\sum_{i=1}^j \binom{L}{i}) - \ln(\sum_{i=1}^{j-1} \binom{L}{i})] \quad (7)$$

In the limit where  $L \gg j$  and  $j \gg 1$  one can verify that

$$\rho(k) \sim -\ln(q) / \ln(L/j) \quad (8)$$

From these equations the following conclusions can be derived. (a)  $|\rho|$  is an increasing function of  $k$ . (b)  $\rho$  is approximately constant for  $\ln j < \ln L$ . (c) For a fixed  $k$ ,  $\rho$  is a function of  $x$  only through  $\ln(q(x))$ . (d) For  $j > L/2$ , the size of the plateaus decreases with  $j$ , and one can verify explicitly from Eq. (7) that the tail of the distribution decays exponentially. (e) For large  $L$  and small  $k$ ,  $\rho(k)$  goes to zero as  $1/\ln(L)$ .

An interesting question now is whether these features of the distribution of  $\{P(k)\}$  also characterize the case where the bias  $B < 1$ . Since the average drift towards 1 is given by  $x_{eff}(B, x) = xB + (1-x)(1-B)$ , it is plausible that this parameter plays an important role. Since  $x$  enters only through  $q$ ,  $\rho$  should depend only on  $q_{eff} = (1-x_{eff})/x_{eff}$ . Indeed, it was found in our simulations that the local slope  $\rho(k; x)$  for all  $x$  are related to a universal curve,  $\rho^*(k)$

$$\rho^*(k) = \rho(k, x) / \ln(q_{eff}(x)) \quad (9)$$

Typical results for  $B = 0.8$  and various  $x$  are plotted in Fig. 2, where the data is ensemble-averaged (the variation between realizations is at any rate small for reasonably large  $L$ ). This data explains the origin of the Zipf

hypothesis, namely the value of  $\rho$  appears fairly constant for small  $k$ ; in fact the relative changes in  $\rho$  are quite large. Note that as  $x - 1/2 = \epsilon \rightarrow 0$ , the rescaling of the local slope vanishes as  $\epsilon$  and a flat distribution is obtained. From Eqs. (8,9) it is clear that the local slope is an increasing function of  $x$  and  $B$ , a result which is expected qualitatively since the tree of coming inputs into each state serves as an **amplifier** whose strength increases with these two parameters. [10] We stress that for the whole distribution the variation of the local slope is visible and for illustration an approximate fit for the data from the Bible is presented in Fig. 1 and for a DNA sequence in the inset of Fig. 3. Note that the tail of the DNA data is also fit well by our model. There is no tail in the data from the Bible, however, presumably due to the finite length of the text relative to the vocabulary. Note that the fitted parameters are not uniquely determined by the  $P(k)$  distribution; essentially only  $x_{eff}$  is fixed.

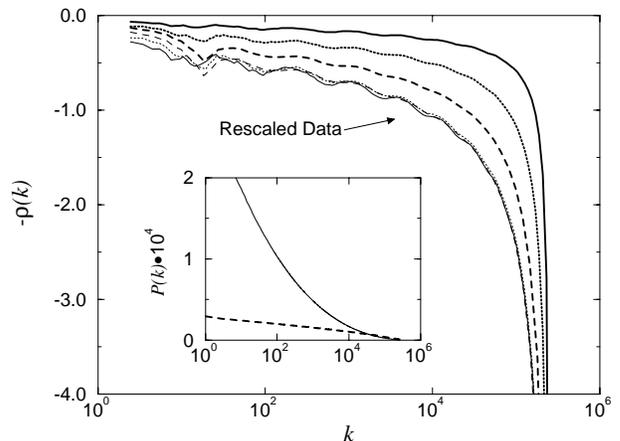


FIG. 2.  $\rho(k)$  for  $L = 18$  and fixed  $B = 0.85$  for  $x = 0.6$  (solid line),  $x = 0.7$  (dotted line), and  $x = 0.8$  (dashed line) and after rescaling by  $\ln(q_{eff}(x, B))$ . Inset:  $P(k)$  for  $L = 18$  and  $x = 0.8$  for  $B = 0.7$  (solid line) and  $B = 0.5$  (dashed line).

As  $B$  decreases toward 0.5, the effect of finite  $N$  becomes stronger so that increasingly large values of  $N$  must be used to achieve the data collapse as in Fig. 2. In the limiting case of  $B = 0.5$ , there is no overall preference for 0 or 1, and therefore  $P(k)$  for each state  $k$  becomes more similar and we expect a distribution similar to  $N$  random numbers whose sum is constrained to unity. [11] There are of course correlations due to the different topology of the tree of inputs to each state, in addition to the fluctuations arising from the random arrangement of the amplification factors  $x$  and  $1-x$  on each tree. [10] However, due to the overall lack of amplification in the trees, the fluctuation effects dominate, giving rise to the above estimate. Indeed in our simulations with  $6 \leq L \leq 20$  we found that (see inset of Fig. 2)

$$P(k) = A - D \ln(k) \quad (10)$$

for essentially the whole range of  $k$ . The dependence of the coefficients  $A$  and  $D$  on  $N$  can be shown from normalization to be  $A(N) \propto \ln N/N$  and  $D(N) \propto 1/N$ . Furthermore,  $A$  and  $D$  are found to increase with the strength  $x$ . [7]

The scaling behavior, Eq. (10), for the probabilities of states in unbiased MPs is the same as the probabilities of occurrence of letters in the alphabet of natural languages or biological sequences. A sequence of letters is affected by short-range correlations due to some phonetic preferences, and which are represented in the MP by local preferences for each state. However, the 'meaning' of a language is embedded only in words, and hence letters are expected to be a result of a process without any 'meaning' or 'tendency', such as an unbiased MP.

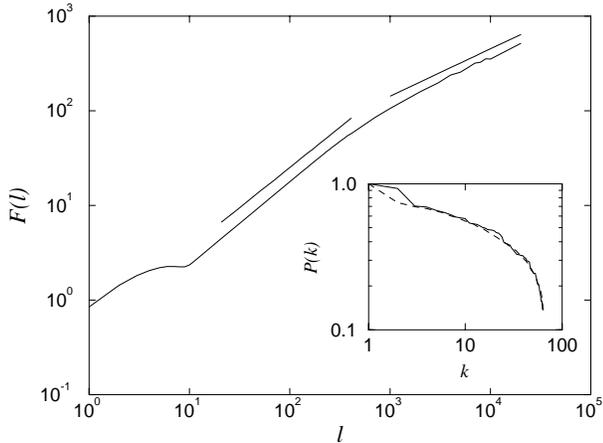


FIG. 3.  $F(l)$  as a function of  $l$  for  $L = 10$ ,  $B = 0.6$ , and  $x = 0.995$ , measured on a sequence of size  $10^8$ . Power laws with exponents 0.85 and 0.5 are drawn for comparison. Inset:  $P(k)$  for the Yeast III chromosome (solid line) and for a MP with  $L = 6$ ,  $x = B = 0.68$ .

Note that the same scaling behavior, Eq. (10), was proven to describe also a set of  $N$  ordered random numbers under the **global constraint** that their sum is unity. [11] However, we believe that the procedure represented by a random MP is a more general and natural explanation for the distribution of letters, since (a) short-range correlations due to phonetic preferences which exist between letters are naturally taken into account only by a MP, (b)  $B = 0.5$  is a limiting case of the biased MP which characterizes the distribution of "words," (c) the MP contains a free parameter  $x$  which can affect the slope  $D$ .

Using the MP to construct random strings, LRC were found for windows up to size  $\sim 10^3$  with  $0.5 < \alpha \leq 1$ . For larger windows, however,  $\alpha \rightarrow 1/2$ , see Fig. 3. These LRC on intermediate scales are a result of the quasi-ergodicity of the MP. For illustration, in the limit  $x \rightarrow 1$  it is clear that the graph of transition probabilities is constructed from loops with weights equal to  $x$ , where dis-

tinct loops are weakly coupled by transitions of strength  $1 - x \rightarrow 0$ . For small windows, the chain follows only one loop of the graph, which in general is biased towards '0' or '1', following the bias  $B$ . For such windows, it is expected that  $\alpha \approx 0.5$ . For intermediate window sizes, large fluctuations are possible. These fluctuations, arising from visiting an atypical loop many times, will dominate the process, giving rise to an  $\alpha > 0.5$ , similar to a Levy flight for random walks. [12] On the largest scales, the finite ergodicity time limits the maximal size of the fluctuations, which eventually become smeared out, and  $\alpha \rightarrow 1/2$ . This picture was indeed verified for some DNA sequences. [7] Furthermore, this picture predicts that the LRC will weaken when  $x$  decreases and as  $B$  approaches unity, an effect which was indeed observed. Increasing  $N$  also weakens the apparent LRC as the probability for an atypical loop decreases. However, a quantitative estimation of the crossover scale beyond which LRC cease is still an open question. Although the vocabulary of a natural language ( $\sim 10^5$ ) is larger than that of DNA ( $\sim 10^2$ ), both  $\alpha$  and the crossover scale are larger for natural text. [7] This is a result of the larger  $x$  for natural languages ( $\sim .9$ ) compared to that for DNA sequences ( $\sim .65$ ).

We note that the entropy of the "languages" described by our MP can be easily calculated and shown to be an extensive quantity. [7] Also, the generic features of our model persist even for the case of a fully connected transition matrix  $W$ , when only a finite number of transitions from a given state are of order unity. [7] It is tempting to speculate that the phenomenological parameters  $x$  and  $B$  can shed light on the actual processes responsible for the generation of the strings encountered in nature, or serve as an aid in classifying different processes. Among the interesting questions are the relationship between different languages, and its possible connection to the historical links between them.

- 
- [1] See, e.g. W. Ebeling and T. Pöshel, *Europhys. Lett.* **26**, 241 (1994) and references therein.
  - [2] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simon and H. E. Stanley, *Nature* **356**, 168 (1992).
  - [3] A. Schenkel, J. Zhang and Y. C. Zhang, *Fractals* **1**, 47 (1993).
  - [4] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, (Addison-Wesley Press 1949).
  - [5] R. N. Mantegna, et. al., *Phys. Rev. Lett.* in press.
  - [6] B. Mandelbrot, *Fractals*, (Freeman, San Francisco, 1977).
  - [7] I. Kanter and D. A. Kessler, unpublished.
  - [8] S. Shtrikman, private communication.
  - [9] B. Derrida, E. Gardner, and A. Zippelius, *Europhys. Lett.* **4**, 167 (1987).

- [10] A detailed explanation of this argument one can find, for instance, in I. Kanter, Phys. Rev. Lett. **60**, 1891 (1988).
- [11] M. Yu. Borodovsky and S. M. Gusein-Zade, J. Biomolecular Structure, **6**, 1001 (1989).
- [12] M. F. Schlesinger, Ann. Rev. Phys. Chem. **39**, 269 (1988).