Linguistic Features of Noncoding DNA Sequences

R. N. Mantegna,[1] S. V. Buldyrev,[1] A. L. Goldberger,[2] S. Havlin,[1] C.-K. Peng,[2] M. Simons,[2] and H. E. Stanley[1]

[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*

[2]*Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215*

(Received 26 April 1994)

We extend the Zipf approach to analyzing linguistic texts to the statistical study of DNA base pair sequences and find that the noncoding regions are more similar to natural languages than the coding regions. We also adapt the Shannon approach to quantifying the "redundancy" of a linguistic text in terms of a measurable entropy function, and demonstrate that noncoding regions in eukaryotes display a *smaller entropy* and *larger redundancy* than coding regions, supporting the possibility that noncoding regions of DNA may carry biological information.

PACS numbers: 87.10.+e, 05.40.+j, 06.50.–x, 72.70.+m

Recently there has occurred an explosion of activity on the interface between statistical mechanics and biological physics [1–4]. One attractive problem concerns applications of statistical mechanics to the sequence of base pairs forming DNA [2]. An intriguing puzzle is related to the fact that in higher organisms, only a small fraction of the DNA sequence is used for coding proteins; the possible function—if any—of the noncoding regions remains unclear [5]. In this Letter, we find that noncoding sequences have certain statistical features in common with natural languages.

A remarkable feature of languages is Zipf's law [6]. In the Zipf analysis, one calculates the histogram that gives the total number of occurrences of each word in a text. If all the words in the text are arranged in rank order, from most frequent to least frequent, then such a histogram is found to be linear on double logarithmic paper, with a slope $-\zeta$, with $\zeta \approx 1$ for all languages studied. Attempts to understand the origin of the Zipf law are connected to the hierarchical structure of language [7].

A second common feature of languages is redundancy: Letters or even entire words can be omitted or changed without the text becoming nondecipherable; e.g., a text with typing errors does not become unintelligible. The notion of redundancy was quantified in the classic work of Shannon [8], who introduced the concept of the entropy from which the redundancy can be computed.

An open question is the biological role of the apparently "silent" noncoding regions of DNA sequences. For this reason, it is natural to enquire whether there are linguistic features of these noncoding sequences. Previously, linguistic analysis has been applied primarily to the coding regions [9,10].

In order to adapt the Zipf and Shannon analyses to DNA, the concept of word must first be defined. In the case of coding regions, the words are the 64 3-tuples ("triplets") which code for the amino acids, AAA, AAT, ..., GGG. However, for noncoding regions, the words are not known. Therefore we begin by considering the word length n as a free parameter, and perform analyses not only for $n = 3$ but also for all values of n in the range 3–8. The different n-tuples are obtained for the DNA sequence by shifting progressively by 1 base a window of length n; hence, for a DNA sequence containing L base pairs, we obtain $L - n + 1$ different words. For the DNA alphabet of 4 characters (A,C,G,T), the number of possible n-tuples is 4^n.

To avoid any bias in DNA sequence selection, we performed Zipf analysis for *each* of the 37 sequences of eukaryotes, eukaryotic viruses, prokaryotes, and bacteriophages comprised of more than 50 000 base pairs (bp) apiece, from GenBank Release No. 81.0 (15 February 1994). In addition, we analyzed bacteriophages λ and T7 (which are just under 50 000 bp), as well as the recently published *C. elegans* sequence comprising 2.2×10^6 bp [11].

Frequency tables of n-tuples have been used to build reference tables to construct successful coding sequence finder algorithms [12]. Here, we study the functional form of the "word" frequency versus the rank in analogy to the Zipf analysis of natural languages. To implement the Zipf analysis, we first rank order the total number of occurrences of each n-tuple, and then we plot their relative occurrence against rank. We analyzed each of the sequences separately, as well as grouping them by category (Table I). Figure 1 shows the Zipf plot for the 14 different sequences of mammalian origin from GenBank No. 81.0. These sequences, comprising 1.1×10^6 bp, are mainly noncoding (4.7% coding regions). The Zipf plot is linear over 2 decades, with slope $\zeta = 0.283 \pm 0.002$, where the error bars represent one standard deviation. If we weight the 14 separate sequences equally by averaging together the ζ values for each sequence, we find 0.32 ± 0.04. We next compare the Zipf analysis for a single representative sequence that is primarily *noncoding* [Fig. 2(a), top curve] and one that is primarily *coding* [Fig. 2(a), bottom curve]. The Zipf exponent of the noncoding sequence ($\zeta = 0.362 \pm 0.003$) is 76% larger than the Zipf exponent for the coding region ($\zeta = 0.206 \pm 0.002$).

0031-9007/94/73(23)/3169(4)$06.00

TABLE I. Linguistic analysis of the 37 sequences in GenBank Release No. 81.0 of 15 February 1994 that exhibit more than 50 000 base pairs (bp) apiece: eukaryotes, eukaryotic viruses, prokaryotes, and bacteriophages; also shown are *C. elegans* and bacteriophages λ and T7. These 40 sequences are partitioned into groups: Group IA contains 14 sequences from mammals (GenBank accession codes are HSG6PHDH, HSMHCAPG, HUMGHCSA, HUMHBB, HUMHDABCD, HUMHPRTB, HUMMMDBC, HUMNEUROF, HUMRETBLAS, HUMTCRADCV, HUMVITDBP, MMBGCXD, MUSTCRA, and RATCRYG). Group IB contains the 74 subsequences comprising part of the recently reported [11] 2.2×10^6 bp sequence of chromosome III of *C. elegans,* and 3 sequences of invertebrate (CEC07A9, CELTWIMUSC, and DROABDB). The sequence DROABDB has been analyzed separately as information about putative coding regions is still not available. Group IC consists of the 315 338 bp yeast chromosome III sequence (SCCHRIII). Group II contains 11 sequences of eukaryotic viruses (ASFV55KB, EBV, HE1CG, HEHCMVCG, HEVZVXX, HS1ULR, HSECOMGEN, HSGEND, IH1CG, VACCG, and VVCGAA). Group III contains 7 sequences of bacteria (BSGENR, ECO110K, ECOHU47, ECOUW82, ECOUW85U, ECOUW87, and ECOUW89), while group IV contains 3 sequences of phage (LAMCG, MLCGA, and PODOT7). The groups are arranged in increasing order of percent of *coding* regions. The exponent ζ and the r^2 coefficient are obtained by least squares fit to the data over the range 1–1000. The quantity $R(4)$ refers to the percent redundancy for a typical value of n, $n = 4$.

	Length	% coding	ζ	r^2	$R(4)$
		I. Eukaryotes			
A. Mammals (14 sequences)					
All	1078 100	5	0.283 ± 0.002	0.98	2.7
Coding	50 687	100	0.208 ± 0.004	0.98	2.1
Noncoding	1027 413	0	0.289 ± 0.002	0.98	2.8
B. Invertebrates					
1. *C. elegans*					
Complete sequence	2176 983	29	0.465 ± 0.002	0.99	4.4
Coding	633 029	100	0.244 ± 0.004	0.93	2.8
Noncoding	1543 954	0	0.537 ± 0.003	0.98	5.6
2. Other invertebrates (3 sequences)					
All	120 966	32	0.403 ± 0.004	0.99	3.8
Coding	38 361	100	0.21 ± 0.01	0.91	2.9
Noncoding	82 605	0	0.477 ± 0.006	0.98	5.1
C. Yeast chromosome III					
Complete sequence	315 338	67	0.289 ± 0.003	0.98	2.4
Coding	211 091	100	0.225 ± 0.005	0.95	2.0
Noncoding	104 247	0	0.391 ± 0.005	0.98	4.0
		II. Eukaryotic viruses (11 sequences)			
All	1616 928	84	0.263 ± 0.002	0.98	0.8
Coding	1361 411	100	0.194 ± 0.001	0.98	0.5
Noncoding	255 517	0	0.362 ± 0.003	0.99	1.5
		III. Prokaryotes (7 sequences)			
All	784 344	83	0.203 ± 0.002	0.97	1.3
		IV. Bacteriophages (3 sequences)			
All	140 735	87	0.158 ± 0.003	0.97	0.8

Of further interest are those sequences possessing *both* coding and noncoding regions of sufficient length to permit accurate statistical analysis; in these we are able to analyze the coding and noncoding regions separately, and we find that the Zipf exponent for the noncoding region is about 50% larger than that for the coding region. Such analysis for the 2.2×10^6 bp sequence of *C. elegans* chromosome III is shown in Fig. 2(b) and for the complete 315 000 bp sequence of yeast chromosome III in Fig. 2(c).

The results of the Zipf analysis for all 40 DNA sequences analyzed are summarized in Table I. The averages for each category support the observation that ζ is consistently larger for the noncoding sequences, suggesting that the noncoding sequences bear more resemblance to a natural language than the coding sequences. To confirm that our methods indeed work on "natural language," we use the same algorithm developed for DNA sequences to analyze actual texts. First, we analyzed a collection of articles taken from an encyclopedia comprising 500 000 letters. We found a power-law behavior using n-tuples instead of real words, but we find $\zeta \approx 0.57$, smaller than the value $\zeta \approx 0.85$ found for Zipf analysis of the *same* text using the actual words. Moreover, we find roughly

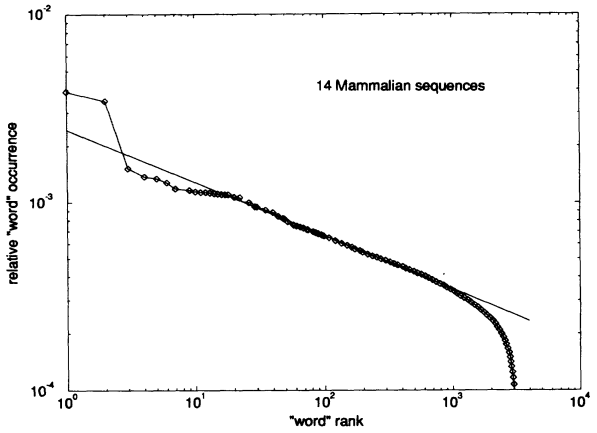

FIG. 1. Double logarithmic plots of the relative number of "word" occurrences of 6-tuples as a function of the rank of the word for the 14 different sequences of mammalian origin from GenBank No. 81.0. Similar results are found for n-tuples with n ranging from 4 to 8. The straight line is the best fit of the data lying in the range 1 to 1000. The exponent $\zeta = 0.283$, significantly larger than the value $\zeta = 0$ expected for a control sequence of random numbers.

the same value of ζ for a range of values of n from 3 to 5.

We also used the same DNA algorithm on a binary executable file of the Unix Operating System containing 9 000 000 bits, which is analyzed in terms of 12-tuples (0 and 1). The Zipf plot is linear over roughly three decades from rank 1 to rank 1000, and we find $\zeta = 0.77 \pm 0.05$. Finally, we analyzed a "control," consisting of a sequence of bits chosen with equal probability to be 0 or 1; the relative occurrence is the same for all words, and the Zipf plot has the expected slope of zero ($\zeta = 0$).

The redundancy is a manifestation of the *flexibility* of the underlying code. To quantitatively characterize the redundancy implicit in the DNA sequence [13], we utilize the approach of Shannon, who provided a mathematically precise definition of redundancy. Shannon's redundancy is defined in terms of the entropy of a text—or, more precisely, the "n entropy" $H(n) = -\sum_{i=1}^{4^n} p_i \log_2 p_i$, which is the entropy when the text is viewed as a collection of n-tuple words [8]. The redundancy is defined through as $R \equiv \lim_{n\to\infty} R(n)$, where $R(n) \equiv 1 - H(n)/kn$; here $k = \log_2 4 = 2$ [14]. We calculated the Shannon n entropy $H(n)$ for $n = 1, 2, \ldots, 6$. The maximum value of n for which it is possible to determine $H(n)$ is $n = 6$—even

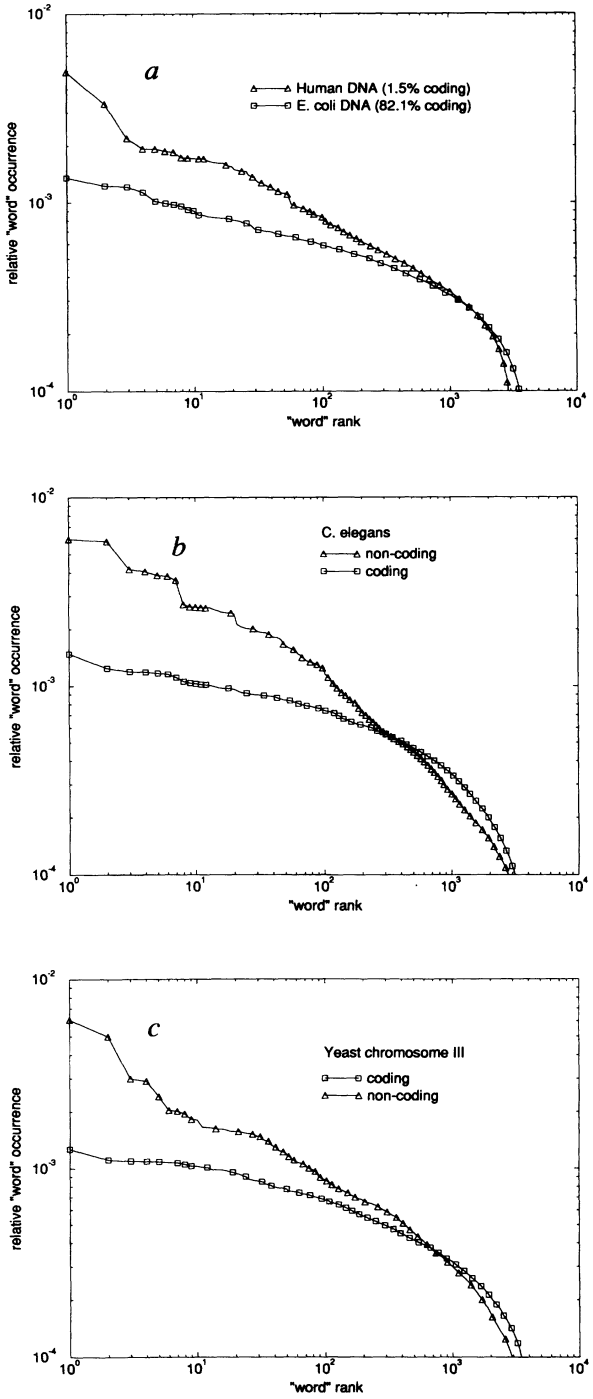

FIG. 2. Zipf analysis of 6-tuples of DNA sequences. (a) The first (bottom curve) is the longest sequence of bacteria DNA available (ECOVW89, 176 195 bp, 82.1% of which are coding regions), while the second is the longest sequence of mammalian DNA available (HUMRETBLAS, 180 000 bp, 1.5% of which are from coding regions). The primarily rich coding sequence has a lower value of ζ ($\zeta = 0.206$) and shows an approximately linear behavior only over roughly 1 decade, while the primarily noncoding sequence has $\zeta = 0.362$, 76% larger, and, moreover, displays linearity over roughly 2 decades. (b) Separate analyses of the noncoding (1 543 954 bp) and putative coding (633 029 bp) regions of the recently published 2.2×10^6 bp *C. elegans*; the Zipf exponent for the noncoding regions ($\zeta = 0.537 \pm 0.003$) is 2.2 times larger than the Zipf exponent for the coding regions ($\zeta = 0.244 \pm 0.004$). (c) Separate analyses of noncoding and putative coding regions of the 315 338 bp sequence of yeast chromosome III; the Zipf exponent for noncoding regions ($\zeta = 0.391 \pm 0.005$) is 1.7 times larger than the Zipf exponent for the coding region ($\zeta = 0.225 \pm 0.005$). We note that for sequences composed of primarily coding regions, the data are well fitted by a logarithmic function [10]. We also find that simple explanations, such as a two-step Markov process, cannot fully account for the observed Zipf-like behavior.

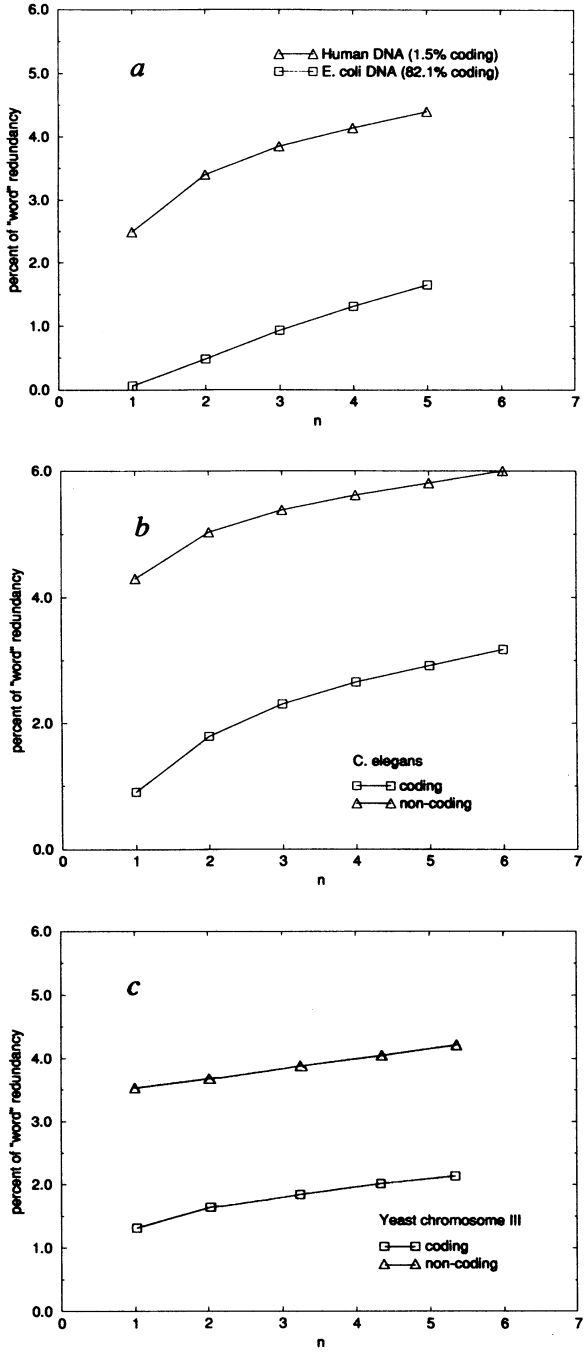

FIG. 3. The Shannon entropy analysis consists of calculating the probability p_i of a given n-tuple and forming the functions $H(n)$ and $R(n)$. Shown in (a) is the percent redundancy $[100R(n)]$ for the same DNA sequences shown in Fig. 2(a). The redundancy is larger for the primarily noncoding sequence. In (b) we show the percent redundancy for the coding (squares) and noncoding (triangles) portions of the *C. elegans* 2.2×10^6 bp sequence. An approximate relation between the Zipf behavior and redundancy can be obtained if a truncated power law is assumed for the Zipf distribution. If the amount of used n-tuples is increasing with n exponentially ($r_0 \propto e^{an}$ where r_0 is the cutoff observed in the Zipf plot), then the redundancy will be asymptotically given by $R(n) \simeq 1 - [\log_2 r_0 + F(\zeta)]/2n$, where $F(\zeta)$ can be calculated analytically.

for very long sequences (e.g., *C. elegans*)—due to the extremely slow convergence to the final value. For shorter sequences, reliable values of $H(n)$ are obtainable only up to a value of n less than 6.

Figure 3 shows $R(n)$ for selected sequences. For sufficiently high values of n (for example, $n = 4$), we see that the redundancy is consistently larger for the primarily noncoding sequences. In fact, for most of the sequences consisting primarily of coding regions, we find that $R(n)$ is quite close to the value $R(n) = 0$ which we find for a control sequence of random numbers.

Thus we find that *noncoding* sequences show two similar statistical properties to those of both natural and artificial languages: (a) Zipf-like scaling behavior, and (b) a nonzero value of Shannon's redundancy function $R(n)$. These results are consistent with the *possible* existence of one (or more than one) structured biological language(s) present in noncoding DNA sequences.

We wish to thank J. J. Schwartz and F. Sciortino for discussions and the NSF, the NIH, the Mathers Charitable Foundation, and the Israel-U.S. Binational Science Foundation for support.

[1] B. J. West and W. Deering, Phys. Rep. (to be published); A. Yu. Grosberg and A. R. Khokhlov, *Statistical Physis of Macromolecules* (AIP, New York, 1994); B. J. West, *Fractal Physiology and Chaos in Medicine* (World Scientific, Singapore, 1990); M. F. Shlesinger, Annu. Rev. Phys. Chem. **39**, 269 (1988).

[2] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, and H. E. Stanley, in *Fractals in Science,* edited by A. Bunde and S. Havlin (Springer, Berlin, 1994).

[3] M. F. Shlesinger and B. J. West, Phys. Rev. Lett. **67**, 2106 (1991); A. Schenkel, J. Zhang, and Y-C. Zhang, Fractals **1**, 47 (1993).

[4] E. Ben-Jacob, O. Shochet, A. Tenenbaum, I. Cohen, A. Czirtók, and T. Vicsek, Nature (London) **368**, 46 (1994).

[5] A. M. Lambowitz and M. Belfort, Annu. Rev. Biochem. **62**, 587 (1993).

[6] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Cambridge, MA, 1949).

[7] B. Mandelbrot, Word **10**, 1 (1954).

[8] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **30**, 50 (1951).

[9] E. N. Trifonov, Bull. Math. Bio. **51**, 417 (1989).

[10] M. Yu. Borodovsky and S. M. Gusein-Zade, J. Biomolecular Structure & Dynamics **6**, 1001 (1989).

[11] R. Wilson *et al.,* Nature (London) **368**, 32 (1994).

[12] J.-M. Claverie and L. Bougueleret, Nucleic Acids Res. **14**, 127 (1985); J. Fickett and C.-S. Tung, *ibid.* **20**, 6441 (1992).

[13] L. L. Gatlin, J. Theoret. Biol. **10**, 281 (1966).

[14] A similar definition was given in H. Almagor, J. Theor. Biol. **117**, 127 (1985).