J. theor. Biol. (1997) **184,** 25–29

Is DNA a Language?

ANASTASIOS A. TSONIS†, JAMES B. ELSNER‡, AND PANAGIOTIS A. TSONIS

† *Department of Geosciences, University of Wisconsin-Milwaukee, Milwaukee, WI* 53201-413,
‡ *Department of Meteorology, Florida State University, Tallahassee, FL* 32306-3034,
and the Department of Biology, The University of Dayton, Dayton, OH 45469, *U.S.A.*

(*Received on* 10 *April* 1996, *Accepted in revised form on* 12 *August* 1996)

DNA sequences usually involve local construction rules that affect different scales. As such their "dictionary" may not follow Zipf's law (a power law) which is followed in every natural language. Indeed, analysis of many DNA sequences suggests that no linguistics connections to DNA exist and that even though it has structure DNA is not a language. Computer simulations and a biological approach to this problem further support these results.

1. Introduction

The evolution of the genetic information and the generation of genes is one of the most challenging problems facing evolutionary and molecular biologists. The principles by which nature produced the genetic information and subsequent generation of DNA sequences are still not well understood. DNA sequences are strings of the bases (nucleotides) A, T, C, G and are characterized as coding (intron-less) or non-coding (intron-containing) sequences. Since the early 1970s, scientists have attempted to discover some kind of order or hidden structure in DNA sequences to discriminate between coding and non-coding regions, to find translation initiation sites, to explore and understand function in genes, etc. With the advent of DNA sequencing techniques in the late 1970s, researchers had the opportunity to probe DNA for such "order". As a result several important issues have been raised including the periodicity of three (Tsonis *et al.*, 1991; Shepherd, 1982), the suggestion of spectra appropriate to $1/f^a$ noises (Voss, 1992) and of long-range correlations (Peng *et al.*,1992, Nee, 1992; Prabhu & Claverie, 1992; Tsonis *et al.*, 1993; Tsonis & Elsner, 1995; Karlin & Brendel, 1993), the existence of local construction rules (Tsonis *et al.*, 1996) and recently the connection of non-coding sequences to linguistics (Mantegna *et al.*, 1994).

A remarkable feature of languages is Zipf's law (Zipf, 1949). This law dictates that the frequency, f, of each word in a text and its rank (the most frequent word having rank 1, the second most frequent word having rank 2 and so on), r, are related according to the power law $f, \propto r^{-k}$ with $k \approx 1$ for all languages [see also a recent review by Casti (1995)]. A similar analysis motivated by this law was applied lately (Mantegna *et al.*, 1994) to DNA sequences and it was concluded that non-coding sequences exhibit linguistic-type features as they obey similar power law (but with much smaller k). A potential problem, however, of this analysis is that the power law was assumed to exist a priori and no comparison to other possibilities was offered.

The power law, $y = Cx^a$, can be viewed as the solution of the differential equation $dy/y = a dx/x$. Such an equation specifies that if x changes from x_1 to x_2 ($x_2 > x_1$), then y is magnified by a factor $(x_2/x_1)^a$. Such properties are appropriate to fractal sets (Mandelbrot, 1983). DNA sequences, however, are characterized by more than one local construction rules that apply to different scales (Karlin & Brendel, 1993; Tsonis *et al.*, 1996). In 1993 Karlin & Brendel presented empirical and theoretical arguments that proved that such "patchiness" cannot justify power laws in spectra and in random walks generated by DNA sequences (see also Tsonis & Elsner, 1995).

0022–5193/97/010025 + 05 $25.00/0/jt960239

Given the above facts the question is posed whether or not local construction rules justify Zipf-like laws in DNA sequences.

2. Data Analysis and Computer Simulations

In order to answer this question we probed the structure of DNA sequences (both coding and non-coding) using the method of the sliding window of length n (Mantegna *et al.*, 1994). According to this approach a window of length n (i.e. n nucleotides) is considered and the different blocks of size n are obtained by shifting the window by one nucleotide at a time. Each block corresponds to a specific arrangement of the four nucleotides of length n. For a given n there exist 4^n distinct arrangements. For a completely random sequence of A, T, C, G, (where each nucleotide is selected randomly with a 25% probability) all 4^n arrangements will appear with the same frequency (assuming that the sequence is sufficiently long). If some kind of order or underlying construction rules exist, then one would expect a relation to emerge between the relative frequency (f) of an arrangement and its ranking (where the most frequent arrangement is ranked 1 and the less frequent is ranked 4^n). The sliding window approach is an effective way to reveal construction rules or the existence of periodicities. In fact such a practice can be found behind several mathematical approaches that are used to study periodicities and deterministic structure and the degree of complexity in data such as Fourier analysis, Singular Spectrum Analysis (SSA), attractor reconstruction etc. (Broomhead & King, 1986; Fraedrich, 1986; D'Alessandro & Politi, 1990).

We considered coding (intron-less) and non-coding (intron-rich) sequences and we adopted the sliding window approach thus obtaining for each sequence an f versus rank plot. Subsequently, we fitted a power law and an exponential law to the results and produced the residual ($f-\hat{f}$) between the actual (f) and the estimated ($\hat{f}$) relative frequency. We then plotted the residual against rank. Such a plot provides an effective way to determine (qualitatively and quantitatively) the goodness of fit of the considered laws. Figure 1 shows examples from two coding and two non-coding sequences. From this figure it is clear that the power law is not the appropriate model as the exponential law fits the data amazingly well in all cases. From the analysis of 20 coding and non-coding (composed of $\sim$75% to 100% introns) sequences we find that the overall significance of the exponential regression is much higher than that of the power law in all cases. Figure 2 summarizes our results by showing the distribution of the residual as estimated from all cases considered for (a) the exponential fit, and (b) the power law fit. This figure clearly shows that the power law results in a highly skewed distribution. The exponential fit on the other hand results in a rather symmetrical distribution with a mean very close to zero and a small standard deviation. As such the power law does not guarantee that the mean residual error will be zero which is a fundamental criterion for selecting an unbiased fit. Using the t-test we find that the null hypothesis that the mean of the residual distribution is zero cannot be accepted in the case of the power law at an acceptable significance level. On the contrary in the case of the exponential law the hypothesis is accepted at a confidence level of 95%. Other choices of n between 3 and 8 also do not justify power laws. It is important to note here that the estimated parameters of the exponential fits (for $n = 6$) do not seem to indicate systematic differences between coding and non-coding sequences. As such, our analysis offers no justification for structural "biases" between coding and non-coding sequences. Although the sliding window approach does not produce words as in languages, fitting power laws to the data over the range of the first 1000 most frequent arrangements (i.e. $1 < \text{rank} < 1000$) has been unjustifiably used to suggest connections between non-coding sequences and linguistics (Mantegna *et al.* 1994). However, no comparison to other laws was ever offered. Our analysis indicates that even if only that range is considered the power law does not outperform the exponential law which is still better overall. Since the power law is essential to make these connections our results suggest that while some structure exists in DNA, non-coding sequences and linguistics do not share similar dynamics.

Computer simulations seem to support the above conclusions. We generated artificial DNA sequences based on the simple assumption or replication of primordial blocks and simultaneous mutation. The results reported in Fig. 3 refer to the sequence generated by repeating many times the blocks CTG and AAG. More specifically we first repeated CTG 3000 times and then continued by repeating AAG 3000 times. We thus manufactured a sequence 18000 bases long whose first half dominated by CTG and the second half is dominated by AAG (CTGCTGCTG.... AAGAAGAAG....) Such a design incorporates the basic assumption of gene evolution by duplication and simultaneous mutation (Ohno, 1988; Yomo & Ohno, 1989) and the fact that DNA sequences may involve more than one local construction rules (Karlin & Brendel, 1993, Tsonis *et al.*, 1996). We then allowed the gene to mutate at

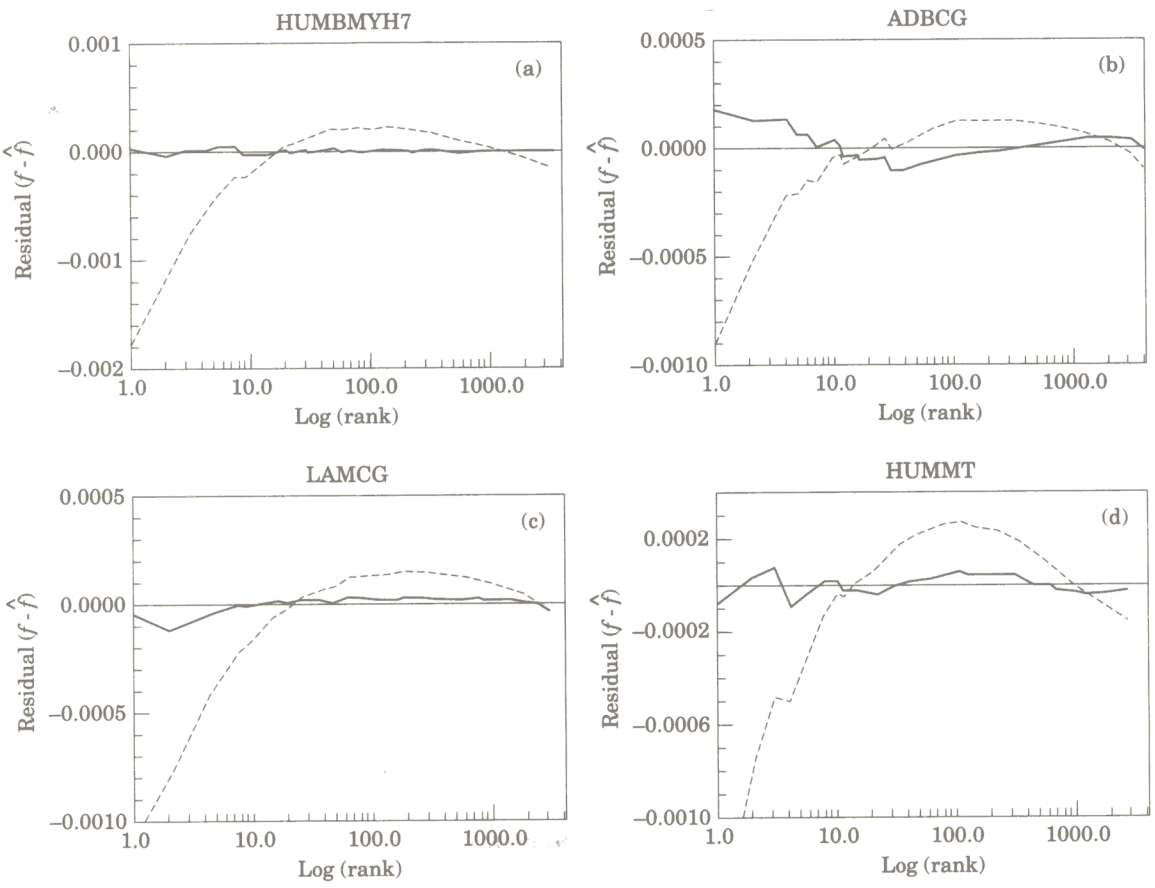

FIG. 1. Residual between the observed (f) and estimated relative frequency ($\hat{f}$) of the possible 4^n arrangements (for $n = 6$) as a function of their rank. Estimated values are obtained from a power law (broken line) and an exponential (solid line) fit to the observed frequency data. The curve providing the smaller residual values represents the better fit between the two laws. The exponential law is superior in all cases. The top two graphs correspond to the intron-rich sequences (a) human β-cardiac MHC (HUMBYMH7) and (b) adenovirus type 2 (ADBCG). The bottom two graphs correspond to the intron-less sequences (c) bacteriophage λ (LAMCG) and (d) human mitochondrion (HUMMT). Note that the exponential law applies equally well to coding and non-coding sequences.

a rate of 67% and applied the sliding window approach. The results in Fig. 3 indicate that again the power law is not the best fit. Similar results are obtained with other simulated DNA sequences involving at least two construction rules and mutation rates between 50 and 75%. We would like to stress here that comparisons to exponential fits in this work are shown for the purpose of demonstrating that the power law is not the appropriate law and not to promote some kind of universality of the exponential law in DNA sequence. The fact that exponential laws fit the data in some cases exceptionally well, may reflect the basic assumptions behind gene evolution (Ohno, 1988; Yomo & Ohno, 1989). Accordingly, it would appear that a model describing building-up or growth processes would be more consistent with theories of gene evolution. In its simplest form such a model can be offered by the differential equation $dy/dx = ay$ whose solution is the exponential law $y = Ce^{ax}$.

3. A Biological Approach

Even though the results presented up to now show that the sequences are not strictly random, they cast doubts on possible connections to linguistics. To further investigate this possibility we thought it necessary to devise a "biological" assay. We assume that if a connection exists between DNA and linguistics, then real DNA sequences must be obtained if sentences were "transformed" to DNA sequences (as for in the case of music derived from DNA sequences (Ohno & Ohno, 1986, for example). Accordingly, we should search not necessarily for DNA "words" but for some kind of alphabet. In fact, what has made languages a powerful tool is not the words alone but the phonetics which were made by precise and selected combination of letters in every language. The letters of a given alphabet follow a given distribution. The codons in coding DNA sequences also follow a certain distribution. It, thus, appears logical to create

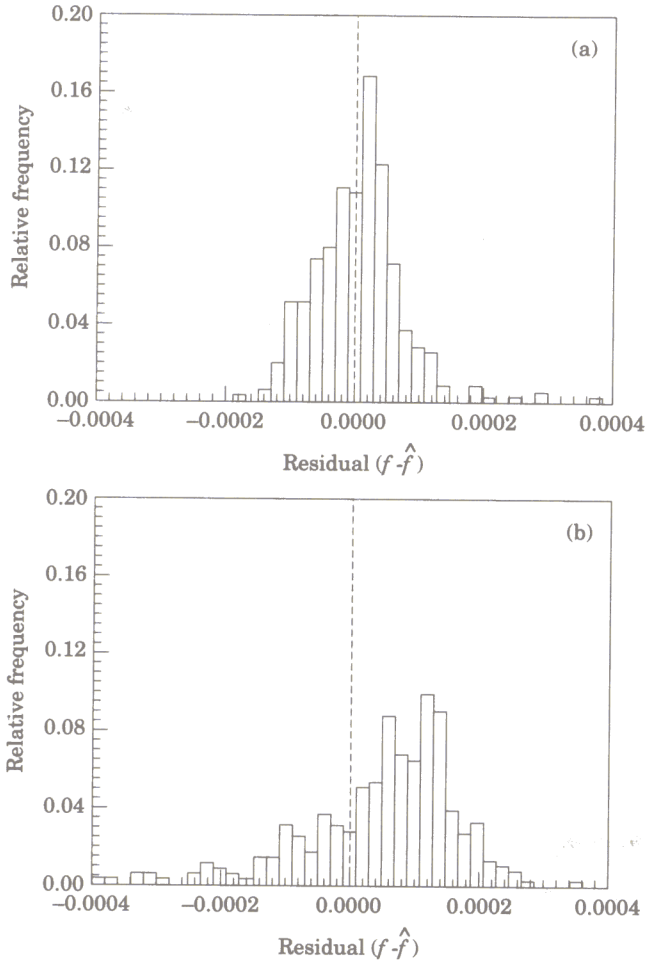

FIG. 2. Frequency distribution of the residuals from all cases considered estimated for the exponential law (a) and the power law (b). The overall superiority of the exponential law is clear (see text for details).

some kind of equivalence between letters and codons. In the English language there are 26 letters and in coding sequences we find 64 codons. The 26 most frequent codons make up 70% of the codons found in coding sequences. Accordingly, a good and simple transformation would be to associate the most frequent letter to most frequent codon and so on down to the least frequent letter and the 26th most frequent codon. A similar transformation can be found for the non-coding sequences. Interestingly, non-coding sequences exhibit a certain distribution of triplets as well. This is a different distribution than that of the codon triplets but it appears to be very consistent among introns. For coding sequences the 26 most frequent codons (with their relative frequency in percentage given by the value in the parenthesis) are (Watson *et al.*, 1987): AAG (4.9), CTG (4.7), GCC (3.8), GAG (3.4), GTG (3.3), GGC (3.2), GCT (2.8), ACC (2.8), TTC (2.8), CAG (2.8), AAC (2.8), CTC (2.7), GAC (2.4) ATC (2.4), TAT (2.3), GGT (2.2), AGC (2.1), GTC (2.1), GAA (2.1), CAC (2.1), AAA (1.9), TCC (1.8), CCC (1.7), TCT (1.6), GAT (1.6) and ATG (1.6). From 20 different genes with a total of 10868 triplets we find that the following are the 26 most frequent triplets in non-coding sequences: AAA(4.8), TTT (4.6), TTA (3.4), TAA (3.0), ATT (2.8), ATA (2.7), AAT (2.7), TAT (2.5), AAG (2.3), GAA (2.0), CAA (2.0), TCT (1.9), TTC (1.9), TGA (1.9), ATG (1.8), TTG (1.8), GAG (1.8), TGG (1.8), AGA (1.8), TGT (1.7), CTA (1.7), GAT (1.7), GCA (1.6), GTA (1.6), AGG (1.6), and TAG (1.6). It is interesting to note here that even though the most frequent triplets are different in coding and in non-coding sequences, their frequency distributions (see numbers in parenthesis) are very similar to each other. Having those transformations we can then consider a sentence and create two DNA sequences one being a "coding" and one being a "non-coding" sequence. Two different types of sentences were employed. Type 1 involved usual sentences that obey the grammatical rules (syntax). Type 2 involved strings of random words that obeyed no rules whatsoever. Subsequently, the constructed coding and non-coding DNA sequences were compared with real DNA sequences in the Genebank for homologies. We found that both type 1 and type 2 sentences result in DNA sequences that score similar homologies with real DNA sequences. This result would indicate that "linguistically" speaking DNA has no syntax but it may have words or phonetics-like structure. In order to test this we shuffled the letters in the original sentences and repeated the analysis. The shuffling

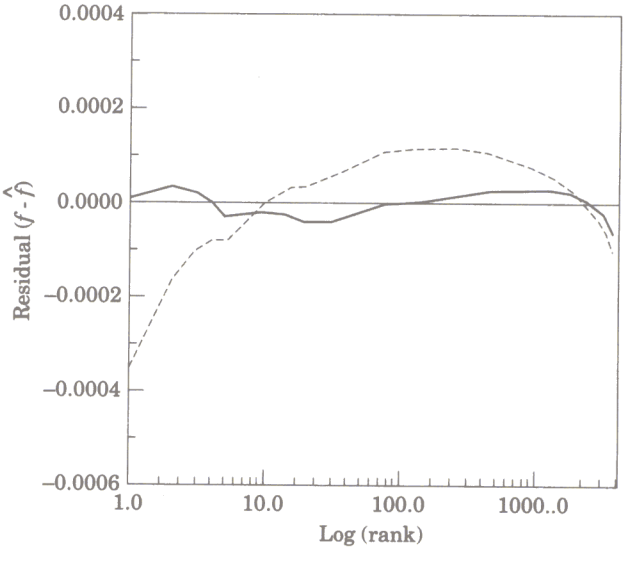

FIG. 3. Same as Fig. 1 but for an artificial DNA sequence (see text for details). Key: ——, exponential law; – – –, power law.

destroys every possible structure within a word and provides an appropriate negative control or surrogate data. We find that these new constructed DNA sequences score homologies with real DNA sequences that are identical to the scores of DNA sequences constructed from the original sentences. Thus, DNA sequences from actual sentences, from strings of random words, and from their shuffled surrogates result in similar homology scores with real DNA sequences. The same procedure was applied to sentences from another language (Greek) with identical results. The inescapable conclusion is clear: DNA sequences show no linguistic properties.

4. Conclusion

The application of mathematical or statistical techniques to DNA sequence is in its beginning stages. If the evolution or function of genes, proteins etc. obeys certain deterministic rules, then mathematics will be able to tell us exactly what these rules are.The benefits of such discoveries would be enormous as they may eventually assist experimental biologists in designing experiments and interpreting results. However, the search for determinism in DNA sequences cannot evolve unless we discover the very basic and guiding models or principles. In that respect it is imperative that we proceed with caution. Our work is a step toward that direction. We believe that developments in this area will be forthcoming and that many insights about DNA sequences will be revealed.

REFERENCES

1. Broomhead, D. S. & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica* **D20,** 217–236.
2. Casti, J. L. (1995). Bell curves and monkey languages. *Complexity* **1 (1)** 12–15.
3. D'Alessandro, G. & Politi, A. (1990). Hierarchical approach to complexity with applications to dynamical systems. *Phys. Rev. Lett.* **64,** 1609–1612.
4. Fraedrich, K. (1986). Estimating the dimensions of weather and climate attractors. *J. atmos. Sci.* **43,** 419–432.
5. Karin, S. & Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science* **259,** 677–680.
6. Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M. et al. (1994). Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* **73,** 3169–3172.
7. Mandelbrot. B. B. (1983). *The Fractal Geometry of Nature* New York: Freeman (1983).
8. Nee. S. (1992). Uncorrelated DNA walks. *Nature*, **357,** 450.
9. Ohno, S. & Ohno, M. (1986). The all pervasive principle of repetitions recurrence governs not only coding sequence construction but also human endeavour in musical composition. *Immunogenetics* **24,** 71–78.
10. Ohno, S. (1988). Codon preference is but an illusion created by the construction principle of coding sequences. *Proc. Natl. Acad. Sci. USA* **85,** 4378–4382.
11. Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Scortino, F., Simons, M. & Stanley. H. E. (1992). Long-range correlations in nucleotide sequences. *Nature* **356,** 168–170.
12. Prabhu, V. V. & Claverie, J.-M. (1992). Correlations in intronless DNA. *Nature*, **359,** 782.
13. Shepherd, J. C. (1982). From primeval message to present day gene. *CSH Symp. Quant. Biol.* **47,** 1099–1108.
14. Tsonis, A. A., Elsner, J. B. & Tsonis, P. A. (1991). Periodicity in DNA sequences: Implications in gene evolution. *J. theor. Bio.* **151,** 323–331.
15. Tsonis, A. A., Elsner, J. B. & Tsonis, P. A. (1993). On the existence of scaling in DNA sequences. *Biochem. Biophys. Res. Commun.* **197,** 1288–1295.
16. Tsonis, A. A. & Elsner, J. B. (1995). Testing for scaling in natural forms and observables. *J. Stat. Phys.* **81,** 869–880.
17. Tsonis,. A. A., Kumar, P., Elsner, J. B. & Tsonis, P. A. (1996). Wavelet analysis of DNA sequences. *Phys. Rev. E* **53,** 1828–1834.
18. Voss, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **68,** 3805–3808.
19. Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A. & Weiner, A. M. (1987). *Molecular Biology of the Gene*. Menlow Park: Benjamin-Cummings.
20. Yomo, T. & Ohno, S. (1989). Concordant evolution and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. *Proc. Natl. Acad. Sci. US* **86,** 8452–8456
21. Zipf, G. K. (1949). *Human Behavior and the principle of Least Effort* Cambridge, MA: Addison-Wesley.