No Signs of Hidden Language in Noncoding DNA

Recent comparison between the statistical properties of coding and noncoding DNA sequences have been interpreted as indicating a yet-undiscovered language in noncoding DNA [1]. We argue that greater variance among nucleotide frequencies in noncoding regions explain most of the observations, which undercuts the claims in [1].

DNA sequences are long strings composed of four nucleotides (A,C,G, and T). For a statistical analysis, these strings may be split into "words" of fixed length n. Then the word frequencies, p_i, are computed. In [1] the Shannon redundancy $R(n)$, $R(n) = 1 + \sum_{i=1}^{4^n} p_i \log_2 p_i/2n$, of noncoding DNA was shown to be nonzero (as in natural languages) and significantly larger than that of coding DNA. For $n = 1$, however, this simply reflects that nucleotide frequencies are more unequal in noncoding than in coding DNA; $R(1)$ increases as the variance of the p_i distribution increases. The increase in $R(n)$ as n increases is the same for coding and noncoding DNA (see Fig. 3 in [1]) and thus does not distinguish between them. Furthermore, it can be shown that correlations of finite range r imply an increasing $R(n)$ even for $n > r$. Such local correlations may be caused by simple mutation processes or could originate from previously coding parts in noncoding DNA [2]. In short, the systematically higher values of $R(n)$ for noncoding than for coding DNA, which [1] argue to be suggestive of hidden language, arise simply because the noncoding DNA has greater variance in its p_i distribution than does coding DNA.

In a "Zipf analysis" all possible 4^n words are ranked according to their frequencies, p_i, from most to least frequent. Power-law behavior was noted in [1], visible by a linear region in a double-logarithmic plot (see Fig. 1). The slope for noncoding DNA was found to be larger than that for coding DNA, and close to that of English text, also analyzed with Zipf's method and fixed word length. This was taken as further evidence that "noncoding regions are more similar to natural languages than coding regions" [1].

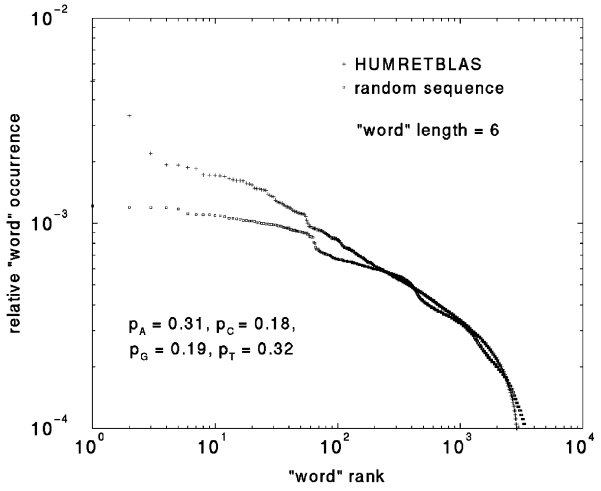

FIG. 1. Zipf plot of a random and a human DNA sequence with the same nucleotide frequencies and sequence length.

The analysis and conclusion are questionable for various reasons. Firstly, assume that noncoding regions are random strings of nucleotides, independently drawn according to the observed nucleotide frequencies. For equal frequencies, all n-tupels have equal probability 4^{-n} and the slope in a "Zipf-plot" is zero. However, as the nucleotide frequencies become more uneven, increasingly distinct plateaus appear (with statistical blurring for finite sequence length). Figure 1 illustrates this, comparing a Zipf-plot for a human DNA sequence (HUMRETBLAS, see [1], 180388 nucleotides, 98.5% noncoding) with a random sequence of identical length and nucleotide frequencies. Considering the crudeness of the approximation, these curves are strikingly similar. Local correlations in the random sequence would smooth out the remaining steps between plateaus.

Secondly, the most probable "DNA words" are very different from those of natural languages. Unlike English, where the most common words are "the," "of," "and," etc., in the present DNA example they are combinations of only the most probable letters—TTTTTT, AAAAAA, TTTTTA, etc. That these words occur more often than expected for uncorrelated random sequences (see Fig. 1), can be readily explained by unequal crossing over, which preferentially occurs in regions of short repeats [2].

Thirdly, the linguistic value of Zipf's approach has been doubted for a long time: Even randomly generated "text" (with words of different length) exhibits power-law behavior with an exponent close to that of natural languages [3].

We have thus shown that most of the observations in [1] may be simple consequences of unequal nucleotide frequencies. Our explanation does not exclude the existence of an undeciphered language in noncoding DNA, but it does undercut speculative arguments based on Zipf's Law or Shannon redundancy [4]. There remains, however, the very interesting question implicit in [1]: Why are there differences in nucleotide frequencies between coding and noncoding DNA?

Sebastian Bonhoeffer, Andreas V. M. Herz, Maarten C. Boerlijst, Sean Nee, Martin A. Nowak, and Robert M. May
Department of Zoology, University of Oxford
Oxford OX1 3PS, United Kingdom

Received 10 February 1995
PACS numbers: 87.10.+e, 02.50.−r, 05.40.+j, 72.70.+m

[1] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169–3172 (1994).
[2] W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution* (Sinauer Sunderland, MA, 1991).
[3] B. Mandelbrot, in *Structures of Language*, edited by R. Jacobson (AMS, New York, 1961).
[4] F. Flam, Science **266**, 1320 (1994).

