Long-Range Power-Law Correlations in DNA

Voss [1] has recently proposed that *coding* as well as noncoding DNA sequences display long-range power-law correlations in their base position (bp) sequences. This finding disagreed with our earlier analysis [2], claiming that coding DNA sequences do not display power-law correlations. However, the discrepancy between [1] and [2] could have arisen because the analysis in [2] was based on partitioning the entire coding sequence into a few large subsequences of constant overall compositional bias. It is important to resolve this discrepancy, since Voss based his scientific conclusion ("immunity to errors on all scales") on his claim of power-law correlations in *coding* sequences.

We prove here that the Voss proposal does not hold generally. Specifically, we present two counterexamples that clearly display *no* long-range correlations *when directly analyzed* (without partitioning into subsequences): (i) the complete genome of T7 bacteriophage (39 936 bp), which contains *only* coding regions, and (ii) the Ti plasmid fragment (24 595 bp), which is believed to consist almost entirely of coding regions.

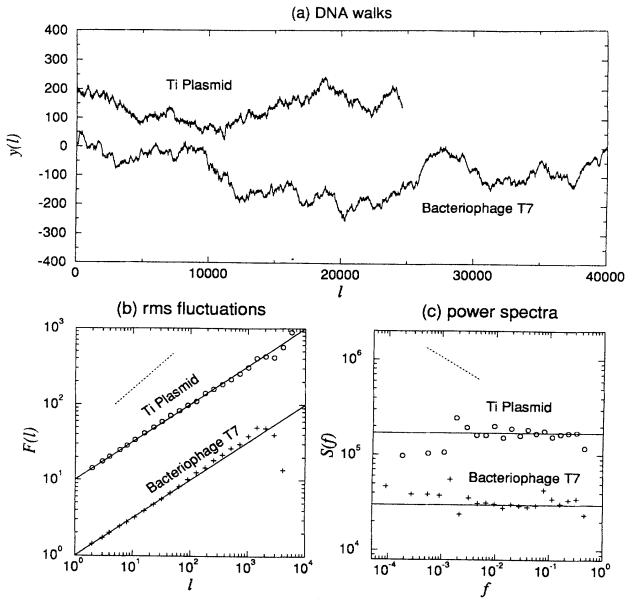

FIG. 1. (a) "DNA walks" [2] of nucleotide sequences; a random walker moves either up or down depending on whether the nucleotide at position ℓ is a pyrimidine or purine. (b) The rms fluctuation, $F(\ell)$, of the DNA walk displacement $y(\ell)$. (c) Power spectrum, $S(f)$, of the binary nucleotide sequences (pyrimidine= 1, purine= -1). Shown are two examples that display *no* long-range correlations: (i) the complete genome of T7 bacteriophage (GenBank name PODOT7), which contains *only* coding regions, and (ii) the Ti plasmid fragment (ATACH5), which is believed to consist almost entirely of coding regions. The solid lines in (b) and (c) have slopes $\alpha = 1/2$ and $\beta = 0$, respectively; for comparison, dashed lines of slope 2/3 and $-1/3$ are also shown, corresponding to the typical behavior found for sequences containing noncoding regions [2]. The data for Ti (○) shifted on the plots for visual comparison.

Figure 1(a) shows the DNA walks for (i) and (ii). Figure 1(b) shows $F(\ell)$, the fluctuation in rms amplitude; the slopes of the log-log plots, *fit over 3 decades*, are 0.53 and 0.49, indicating the absence of long-range correlation for both cases [3]. Figure 1(c) shows the power spectrum $S(f)$, which is almost perfectly flat (indicative of no correlation or "white noise"). The scaling behavior in Figs. 1(b) and 1(c) is markedly different than that found for genomic sequences containing substantial noncoding subregions, for which $F(\ell) \sim \ell^{\alpha}$ with $\alpha \approx 2/3$ and $S(f) \sim 1/f^{\beta}$ with $\beta \approx 1/3$ [2].

So why did Voss find $\beta = 1.02$ and 1.16, respectively, for phage and bacteria (which contain mostly coding regions)? A clue is apparent from comparing Voss' analysis for these cases (Fig. 3 of [1]) with his fits for the noncoding segments. We see that all except a few small-f data points are well fit by a horizontal line, corresponding to $\beta = 0$ (no long-range correlation). The departure at small f likely corresponds to the fact that most coding sequences contain uncorrelated subsegments—*with a characteristic length*—of alternating compositional bias. The DNA walks, therefore, resemble a spliced together string of *uncorrelated* but *biased* random walks. We confirmed this likely source of spurious low-f behavior by calculations on artificial "control" sequences.

S. V. Buldyrev,[1] A. L. Goldberger,[2] S. Havlin,[1]
C.-K. Peng,[1] M. Simons,[2,3] F. Sciortino,[1]
and H. E. Stanley[1]
[1]Physics Department, Boston University
Boston, Massachusetts 02215
[2]Cardiovascular Division, Beth Israel Hospital
Harvard Medical School
Boston, Massachusetts 02215
[3]Biology Department
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Received 23 October 1992
PACS numbers: 87.10.+e, 05.40.+j, 06.50.–x, 72.70.+m

[1] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[2] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[3] One advantage of the method of [2] is that to find the exponent characterizing the long-range correlation one need not correct the data by subtracting the white noise, $S(\infty)$ [1]. Since there is no unambiguous method of estimating $S(\infty)$, this need to correct the data introduces an uncontrollable source of uncertainty. In the power spectrum analysis for those sequences containing noncoding regions, subtracting of the white noise, $S(\infty)$, as performed in [1], gives more weight to the noncoding segments (correlated) than the coding segments (uncorrelated). See H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, and M. Simons, Physica (Amsterdam) **199A**, 3 (1993), and references therein.