**Mantegna *et al.* Reply:** Reference [1] raises the question concerning which kind of "control" sequence is the best to test the robustness of the results found in [2], suggesting that the most appropriate control symbolic sequence is the "one with power-law correlations." In [2], due to the absence of an indisputable satisfactory control, we chose to consider the simplest test and compared our results on noncoding and coding DNA with the results expected for a Bernoulli, with equal letter probability, and first-order Markovian sequences.

Our choice is in some respects unsatisfactory, because the DNA sequences are, of course, more complex than first-order Markovian symbolic sequences. The choice adapted in [1] is perhaps even more unsatisfactory for the following reasons: The frequency rank $\omega(R)$ analysis of $n$-tuples and the analysis of long-range correlations are, in general, not equivalent. Considering $\omega(R)$ analysis for power-law correlated binary strings (with equal probability of the two digits) was the subject of a study [3] that showed that the $\omega(R)$ exponent $\zeta$ and the long-range correlation exponent $\alpha$ provide information on quite different length scales; in fact, $\alpha$ may be related to $\zeta$ only if the corrections-to-scaling terms in the investigated symbolic sequences are quite small [3]. In other words, if $\alpha$ is not constant as a function of length scale $\ell$, then what is detected by the $\omega(R)$ analysis is not only some information related to $\alpha$ but mostly information related to the local correlation exponent $\alpha_\ell$ where usually $\ell \approx 10$ (see, e.g., Fig. 4 in Ref. [3]). In the control test performed in [1] $\alpha_\ell = \alpha$ is implicitly imposed by the procedure used in the test, while in real DNA sequences $\alpha_\ell \neq \alpha$ when $\ell \approx 10$. In fact, in actual noncoding DNA sequences power-law correlations are not ideal: The correlation exponent measured for length scales larger than 10 bp (base pair) differs from the approximation for shorter length scales.

To distinguish structured biological information from noise on a statistical basis would require a robust measure of complexity which is still lacking; $n$-tuple Zipf analysis is not a complexity analysis. Concerning the usefulness of the $n$-tuple $\omega(R)$ analysis discussed in Ref. [1], we note that an approximate power-law $\omega(R)$ plot is observed in natural and formal languages, but its observation in a symbolic sequence is *not sufficient* to prove the existence of an underlying language. Indeed, we presented evidence [2,4] suggesting that certain statistical properties of noncoding eukaryotic regions better resemble these properties of natural languages than do those of the coding regions; we never claimed that our results demonstrate the existence of a "language" in the noncoding DNA of eukaryotes.

Next, we present a qualitative argument showing that long-range correlations are not equivalent to a power law $\omega(R)$. Imagine a given sequence $S$ that has a long-range correlations characterized by a power-law decay with exponent $\alpha$. Divide $S$ into subsequences of $n$-tuples "words" and generate a new sequence $S'$ by randomly reordering the $n$-tuples in $S$. Certainly the long-range correlations will be affected, but the $\omega(R)$ plot as well as the entropy and redundancy function will roughly remain the same, since roughly the same $n$-tuples enter the calculations. The above argument, supported by explicit calculation (Fig. 1) confirms that long-range correlations and $\omega(R)$ analysis carry different information about the analyzed symbolic sequences when $\alpha_\ell$ is not constant.

A final reason showing that the "control" in [1] is not conclusive is the fact that languages also display long-range power law correlations [5] so that simultaneous occurrence of long-range correlations and linear behavior of an $\omega(R)$ plot in noncoding, but not in coding, DNA is not inconsistent with (but of course does not prove) the presence of a structured language in noncoding DNA.

Reference [6] claims our results [2] arise mainly from the difference in bp concentration between coding and noncoding DNA. For long coding and noncoding regions, $x_C \approx x_G$ and $x_A \approx x_T$ so we study the parameter $x_{CG} \equiv x_C + x_G$. Indeed, the CG concentration produces a strong background effect [7], but it is not sufficient to explain our findings (Fig. 2): We have analyzed DNA sequences using an AG/CT (purine-pyrimidine) *binary*

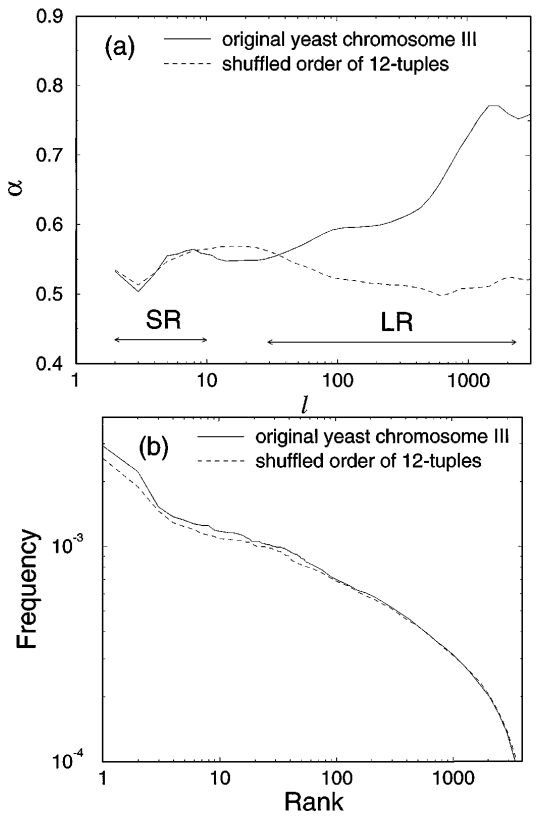


FIG. 1. Example showing, for yeast chromosome III (315 000 bp), that long-range correlation analyses are in general not equivalent to a power-law $\omega(R)$ plot. Shown is the original sequence before and after shuffling the order of 12-tuple binary "word". (a) Dependence of the effective value $\alpha_\ell$ of the long-range correlation exponent $\alpha$ on length scale $\ell$, (b) the $\omega(R)$ plot. Shuffling "word" order destroys long-range correlation but has a markedly smaller effect on the $\omega(R)$ plot.

  

rule in which the CG effect is eliminated, and find that the $n$-gram redundancy $R(n)$ for all lengthy vertebrate and invertebrate DNA is larger for noncoding than coding.

To compensate for the CG effect, we compare $\omega(R)$ plots observed for coding and noncoding regions having the same $x_{\mathrm{CG}}$ (within a 1% tolerance) by analyzing all the coding and noncoding regions of the set investigated in [2] (longer than $10 \times 4^6$). Such long coding and noncoding sequences having identical CG content and *also* representing the same organism do not exist, so we compare different organisms. In spite of this limitation, the test might be useful. We are able to identify 15 pairs of sequences having almost the same value of $x_{\mathrm{CG}}$; in the majority of cases, $\omega(R)$ is roughly a power law for the noncoding regions and a logarithmic function for the coding regions (cf. Fig. 3). We carry out a parallel analysis for $R(n)$, and find that the noncoding DNA in eukaryotes has larger $R(n)$ than the coding for all but two cases (the viruses VACCG and HSGEND).

Reference [8] makes several claims: (a) The first is that, even if correct, Zipf analysis provides no useful information about natural languages. In fact, our work concerns the analysis of *n-tuples*, not traditional Zipf analysis of words *of different lengths* (for which the claim of [8] is known to be correct). The $n$-tuple $\omega(R)$ analysis we perform—unlike the older Zipf analysis—*is* a current linguistic tool [9].

(b) Figure 1(a) of [8] is incorrectly interpreted to imply that we did not take into account the known fact that $\omega(R)$ has a nonzero slope for a randomized control of *finite* length. Finite-size effects exist, but Fig. 1(a) of [8] shows that for the sequence lengths we analyzed, the finite-size contributions to $\zeta$ are smaller than the values of $\zeta$: The top curves produce values for exponent $\zeta$ that are of order 0.1 while we found typical values of $\zeta \approx 0.3$. Moreover, for the majority of cases, the coding is shorter than the noncoding part, so the contribution from the finite-size effect would lead to spuriously *larger* values for coding DNA—the opposite of what is found [2,4].

(c) Figure 1(b) of [8] considers the effect of nonuniform bp concentrations, a known feature of both coding and noncoding DNA. The magnitude of this effect is misrepresented (the values 0.6, 0.25, 0.10, and 0.05 differ greatly from the known nonuniformity [7]).

(d) Figure 1(c) of [8] is incorrectly interpreted to imply that trivial local correlations in *noncoding* DNA may explain the $\omega(R)$ behavior of [2]. The trivial correlations used in [8] imply a probability distribution of simple repeats of $n$ identical bp's decaying exponentially with $n$, but in all sufficiently long sequences we studied, we find such exponential distributions only in *coding* regions, while in noncoding regions the length distribution of repeats is better approximated by a power law (with exponent $\mu > 3$) than by an exponential (cf. Fig. 4) [10]. Moreover, the process suggested in [8] is in fact a first-order Markov process, but Fig. 2 clearly shows that finite-order Markov processes cannot explain the $R(n)$ behavior reported in [2].

(e) Figure 2(f) of [8] gives the impression that the differences in $\omega(R)$ plots for coding and noncoding sequences for lengthy primates is negligible. The observation that this difference is small was shown in Table I of [2]. Important is not only the slope of the power-law approximation but also the functional form of the coding and noncoding regions, which cannot be seen on the scale of Fig. 2 of [8]; we found that coding DNA is better fit by a logarithmic function than by a power law [4]. Incorrectly interpreted is Fig. 2(e) of [8], which concerns the longest human sequence HUMTCRB (unavailable when [2] was written): While the noncoding DNA lies below the coding DNA on a $\omega(R)$ plot, what is relevant is not the absolute

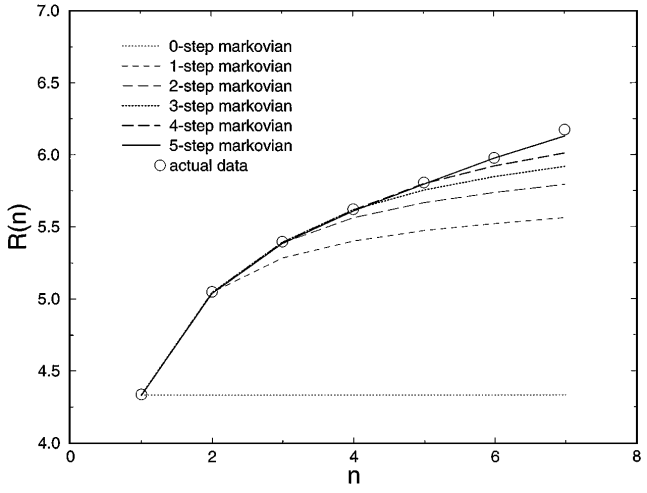


FIG. 2. $R(n)$ for C-elagans chromosome. Circles are the actual data, while lines are successive Markovian approximations. The zero-order Markovian generates the uncorrelated sequence with the correct CG concentration. The first-order Markovian generates the symbol repeats suggested in [8], and clearly cannot account for $R(n)$ for $n > 1$.

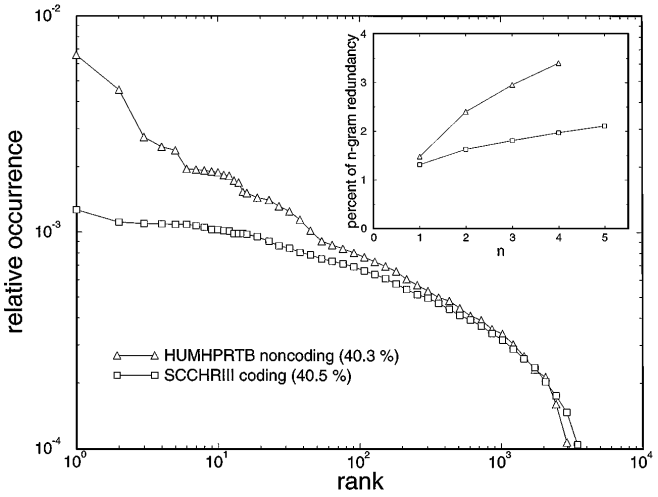


FIG. 3. Comparison of the noncoding regions of HUMHPRTB ($L = 56080$, $x_{\mathrm{CG}} = 0.403$) and the coding regions of SCCHRIII ($L = 211091$, $x_{\mathrm{CG}} = 0.405$). $\omega(R)$ is logarithmic for coding regions and power law for noncoding regions. The inset shows that $R(n)$ is larger for the noncoding regions.

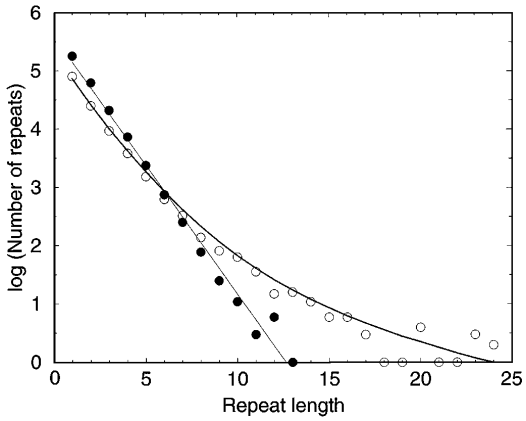


FIG. 4. Frequency of repetitions of $n$ identical bp's in the four yeast chromosomes III, VI, IX, and XI. Coding data (●) follow an exponential distribution, while noncoding data (○) follow a power law—in disagreement with the arguments of [8].

value but rather the functional form. We find that the coding data are less linear on a log-log plot (Fig. 5).

(f) Figure 2(g) of [8] shows that for bacteria the difference between coding and noncoding—although it does clearly exist—is small. However, for bacteria the noncoding regions are short [11] and have different biological functions, which is why, in our studies [2,4], we limited ourselves to the noncoding regions of eukaryotes.

(g) In general, Fig. 2 of [8] is interpreted to demonstrate that the differences between noncoding and coding DNA vanish when the entire GenBank database is analyzed. The GenBank is not an unbiased sample; e.g., multiple copies of DNA sequences from the same gene (either representing different clones or different organisms) are present in it [12]. Indiscriminate use of the GenBank does not increase the robustness of one's analysis and may, in fact, lead to erroneous conclusions. The set of sequences of the GenBank database is an artificially redundant set for which a reliable frequency analysis of $n$-tuples cannot be performed. Moreover, the longest

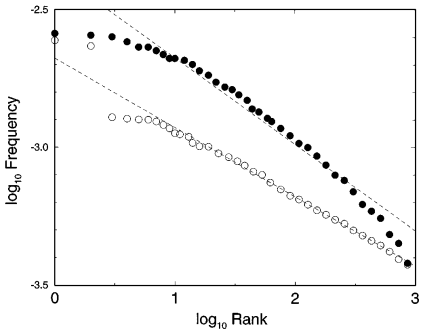


FIG. 5. $\omega(R)$ plot for the longest sequence studied in [8]. Note that over the range $10 < R < 1000$, the noncoding regions (○) are well fit by a power law ($R = 0.99$), but the coding regions (●) cannot be ($R = 0.96$). This difference in scaling behavior is consistent with the analysis in Ref. [4].

mammalian sequences analyzed in [2,4,8] are probably not typical mammalian sequences selected at random—usually they represent regions that contain multiple copies of genes in a family of closely related genes (as in the case of HUMTCRB) [12]. The natural redundancy of the particular sequences chosen could be significantly different from that of typical *randomly chosen* sequences of the mammalian genome. For this reason, we chose to analyze complete chromosome sequences (yeast III, XI) as well as the 2.2 Mb sequence from *C. elegans.* As more complete chromosomal sequences become available, the entire question of whether there is a genuine difference in $\omega(R)$ and $R(n)$ for coding and noncoding DNA will be resolved.

R. N. Mantegna,[1] S. V. Buldyrev,[2] A. L. Goldberger,[3]
S. Havlin,[2] C.-K. Peng,[2] M. Simons,[3] and H. E. Stanley[2]
[1]Dipartimento di Energetica ed Applicazioni di Fisica
Universtà di Palermo, Palermo, I-90128, Italy
[2]Physics Department, Boston University
Boston, Massachusetts 02215
[3]Harvard Medical School, Boston, Massachusetts 02215

Received 24 October 1995
PACS numbers: 87.10.+e, 02.50.−r, 05.40.+j, 72.70.+m

[1] N. E. Israeloff *et al.*, preceding Comment, Phys. Rev. Lett. **76**, 1976 (1996).
[2] R. N. Mantegna *et al.,* Phys. Rev. Lett. **73**, 3169 (1994).
[3] A. Czirók, R. N. Mantegna, S. Havlin, and H. E. Stanley, Phys. Rev. E **52**, 446 (1995).
[4] R. N. Mantegna *et al.*, Phys. Rev. E **52**, 2939 (1995).
[5] A. Schenkel, J. Zhang, and Y-C. Zhang, Fractals **1**, 47 (1993); M. Amit *et al.*, Fractals **2**, 7 (1994); W. Ebeling and A. Neiman, Physica (Amsterdam) **215A**, 233 (1995).
[6] S. Bonhoeffer *et al.*, preceding Comment, Phys. Rev. Lett. **76**, 1977 (1996).
[7] S. V. Buldyrev *et al.* (unpublished).
[8] R. F. Voss, preceding Comment, Phys. Rev. Lett. **76**, 1978 (1996).
[9] See, e.g., M. Damashek, Science **267**, 843 (1995).
[10] We argue that artificial sequences with power-law distribution of simple repeats can mimic the behavior of both the $\omega(R)$ and $R(n)$ plots for noncoding regions much better than low-order Markovian processes [S. V. Buldyrev (unpublished)]. Repeats with $\mu \geq 3$ cannot explain long-range correlations within the framework of the generalized Lévy-walk model [S. V. Buldyrev *et al.*, Phys. Rev. E **47**, 4514 (1993)].
[11] Reference [8] overlooks the fact that if two sequences of length $L_1$ and $L_2$ are characterized by $L_1 \ll L_2$ and if $L_1$ is not sufficiently long for good statistics (longer than $10 \times 4^n$), then $\zeta_1$ is overestimated [H. Herzel *et al.*, Chaos, Solitons Fractals **4**, 97 (1994)]. The coding regions are, in some cases, much shorter than the noncoding regions, so the conclusions drawn in [8] are inconclusive.
[12] R. Guigó and J. W. Fickett, J. Mol. Biol. **253**, 51 (1995).