

## Comment on “Factorial Moments Analyses Show a Characteristic Length Scale in DNA Sequences”

In a recent Letter, Mohanty and Rao [1] introduce a method to study the existence of characteristic length scales in DNA sequences. Here we show that the method used by the authors is affected by numerical artifacts that lead to spurious values of the characteristic length. In short, the procedure used in [1] works as follows: they count the number of occurrences,  $n$ , of a nucleotide (or group of nucleotides) in all possible windows of size  $\ell$  along the sequence (allowing overlapping); then they obtain the normalized distribution of these numbers (*density spectrum*)  $\rho_n(\ell)$  and repeat the procedure for each  $\ell$ . Finally, they characterize these distributions by their variance  $\sigma^2(\ell)$  or their factorial moments,  $F_q(\ell)$ . Given a normalized distribution  $\rho_n$ , its factorial moment of order  $q$  is defined as  $F_q = f_q/f_1^q$ , where

$$\begin{aligned} f_q &= \sum_{n=q}^{\infty} \rho_n n(n-1)\cdots(n-q+1) \\ &= \sum_{n=q}^{\infty} \frac{n!}{(n-q)!} \rho_n. \end{aligned} \quad (1)$$

For a random sequence, provided that  $\ell$  is *great enough*, all  $F_q$  become unity independently of  $q$ . Using this fact, Mohanty and Rao [1] determine the characteristic length,  $\ell_c$ , for DNA sequences as the length for which  $F_q$  crosses the unit value, being  $\ell_c$  almost the same for all low- $q$  factorial moments. One of the main results of the Letter is the finding of this characteristic length in DNA sequences, around which the density distribution  $\rho_n$  is nearly Poissonian and, above it, the DNA sequences show long range correlations.

The fact that  $F_q = 1 \forall q$  in a random sequence is true only if  $\ell$  is great enough. For small values of  $\ell$  the finite size of the window has to be taken into account. For a random sequence,  $\rho_n$  is a normal distribution only asymptotically. In fact,  $\rho_n$  is a binomial distribution, having factorial moments given by

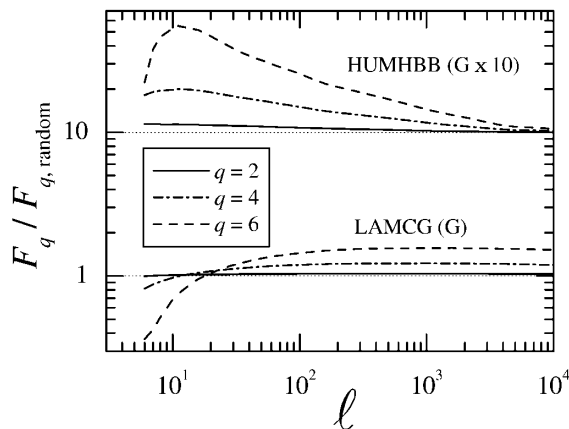


FIG. 1.  $F_q/F_{q,\text{random}}$  vs  $\ell$  for  $q = 2, 4, 6$ .

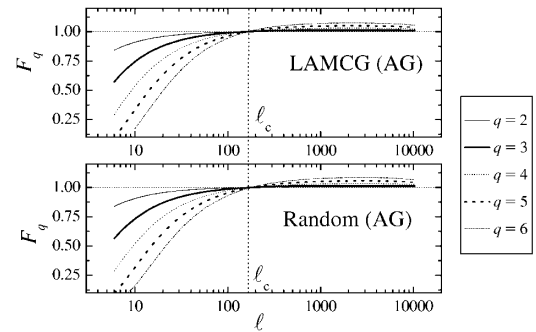


FIG. 2.  $F_q$  vs  $\ell$  for  $q = 2, 3, 4, 5$ , and  $6$  for LAMCG and an artificial sequence obtained by joining three randomly generated patches of sizes 22 000, 17 000, and 9000 bp, and probabilities of AG 0.54, 0.47, and 0.54, respectively, corresponding to the three well-known compositional domains of LAMCG.

$$F_{q,\text{random}}(\ell) = \frac{\ell!}{(\ell-q)!} \ell^{-q}. \quad (2)$$

It is easy to verify that Eq. (2) tends asymptotically to unity, but for  $\ell < 100$  it departs significantly from unity, and precisely this is the range in which Mohanty and Rao find  $\ell_c$  [e.g.,  $F_{6,\text{random}}(100) = 0.858$ ].

Therefore, to measure deviations from random behavior at a given length  $\ell$ , factorial moments must not be compared to unity (as done by Mohanty and Rao) but to Eq. (2). A reasonable way to estimate  $\ell_c$  could be to compare  $F_q(\ell)$  to  $F_{q,\text{random}}(\ell)$ . In Fig. 1 we plot  $F_q/F_{q,\text{random}}$  as a function of  $\ell$  for two of the sequences analyzed in [1]. For the viral sequence (LAMCG) we find  $\ell_c \approx 10\text{--}20$  bp, while in [1]  $\ell_c = 50$  bp was obtained. Note also that there is not a single clear point where all three moments cross unity, and, therefore, several values of  $\ell_c$  could be defined. As for the human sequence (HUMHBB), we find no  $\ell_c$  at all; i.e., all the factorial moments are above the corresponding random ones in the whole range of  $\ell$ , in contradiction with the approach in [1] which gives a clear value of  $\ell_c = 11$  bp.

To show that  $F_q$  analysis can give spurious characteristic lengths, we compare LAMCG with an artificial random sequence (Fig. 2). The plots of  $F_q$  vs  $\ell$  for both sequences are identical, leading to the same  $\ell_c$  although there cannot exist a characteristic length in the random sequence in the range 100–200 bp.

P. Bernaola-Galván and P. Carpena  
Departamento de Física Aplicada II  
E.T.S.I. de Telecomunicación  
Universidad de Málaga  
Málaga 29071, Spain

Received 10 January 2001; published 13 May 2002

DOI: 10.1103/PhysRevLett.88.219803

PACS numbers: 87.14.Gg, 05.40.-a, 72.70.+m, 87.10.+e

[1] A. K. Mohanty and A. V. S. S. Narayana Rao, Phys. Rev. Lett. **84**, 1832 (2000).