Factorial Moments Analyses Show a Characteristic Length Scale in DNA Sequences

A. K. Mohanty[1] and A. V. S. S. Narayana Rao[2]

[1]*Nuclear Physics Division, Bhabha Atomic Research Centre, Mumbai 400085, India*

[2]*Molecular Biology and Agriculture Division, Bhabha Atomic Research Centre, Mumbai 400085, India*

(Received 2 July 1999)

A unique feature of most of the DNA sequences, found through the factorial moments analysis, is the existence of a characteristic length scale around which the density distribution is nearly Poissonian. Above this point, the DNA sequences, irrespective of their intron contents, show long range correlations with a significant deviation from the Gaussian statistics, while, below this point, the DNA statistics are essentially Gaussian. The famous DNA walk representation is also shown to be a special case of the present analysis.

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

Recently, there has been considerable interest in the finding of long range correlations in DNA sequences [1–8]. Based on spectral analyses, Li *et al.* found [1] that the frequency spectrum of a DNA sequence containing mostly introns shows $1/f^{\alpha}$ behavior, which evidences the presence of long range correlations. The study of Peng *et al.* [2] on the basis of a random walk model also led to the conclusions that intron-containing DNA sequences exhibit long range correlations, whereas such a correlation was not found for any of the intronless or cDNA segments. While these observations raised interesting questions about the role of introns and intron-containing genes, other investigators have questioned whether the results are an artifact of the min-max method used by Peng *et al.* [2]. Chatzidimitriou-Dreismann and Larhammar [5], on the other hand, by doing a more careful analysis of the same data set concluded that both intron-containing and intronless DNA sequences exhibit long range correlations. A subsequent work by Prabhu and Claverie [6] also substantially corroborates these results. Alternatively, Voss [8], based on equal-symbol correlation, showed a power law behavior for the sequences studied regardless of the percent of intron contents. Therefore, despite the efforts spent, it is still an open question whether the long range correlation properties are different for intronless and intron-containing coding regions. The above controversies primarily arise due to different models used in the analyses. For example, the min-max [2] and the detrended fluctuation analysis (DFA) [9] have been suggested to eliminate the effect of local patchiness in the DNA sequence. Even the assumption of equal probability of step up or step down of the DNA walk formalism has been questioned by others [4]. The conclusion drawn from Voss's analyses, i.e., all DNA sequences show long range correlations, does not seem to be valid in general. As shown in Ref. [10], the power law spectrum is nearly flat for certain coding sequences which indicates the absence of any long range correlations. Although this cannot be generalized [11], according to Ref. [12] (based on Levy's statistics), long range correlations would imply a strong deviation from Gaussian statistics while the investigation of Ref. [13] yields an important conclusion that the DNA statistics are essentially Gaussian. Therefore, investigations based on different models seem to contradict each other, as they all look into only a certain aspect of the entire DNA sequence. Thus, it is very important to investigate the degree of correlations in a model independent way so that the interpretation of the data including the theoretical analysis becomes more meaningful. The purpose of this Letter is, therefore, to propose a new method to analyze the long range correlation from the variance analysis of the density distribution of a single or group of nucleotides. The famous DNA walk representation [2] whose statistical interpretation is still unresolved is shown to be a special case of the present formalism with a density distribution corresponding to a purine or pyrimidine group. The present method is quite general that displays the correlations at the individual nucleotide level.

We count the number of occurrences (n) of any individual nucleotide (G, A, T, or C) or any particular group between l_0 and $l_0 + l$:

$$n = \sum_{i=l_0}^{l+l_0} u_i \, , \tag{1}$$

where $u_i = 1$ if the DNA site is occupied by the above nucleotide (group) and $u_i = 0$ otherwise. Then the density spectrum ρ_n (i.e., n distribution , normalized to unity) is obtained by varying l_0 from $1, 2, 3, \ldots,$ up to $L - l$ so as to cover the entire DNA sequence of length L. Finally, we compute the variance $\sigma^2 = \langle n^2 - \bar{n}^2 \rangle$ and the factorial moments F_q. The factorial moments of order q are defined as $F_q = f_q / f_1^q$ where

$$\begin{aligned} q &= \sum_{n=q}^{\infty} \rho_n n(n-1)\cdots(n-q+1) \\ &= \sum_{n=q}^{\infty} \frac{n!}{(n-q)!} \rho_n \, . \end{aligned} \tag{2}$$

It is easy to verify that the factorial moments become unity for all order (i.e., become independent of q) when ρ_n becomes Poissonian. In this work, we have applied the above factorial moment analysis (generally used to study the fluctuations during a phase transition [14]) to study the dynamical fluctuations present in the DNA sequences.

First we demonstrate the usefulness of the present formalism with two typical examples. We compute σ^2 for GA distribution (purine group, $u_i = 1$ if the site is occupied by a G or A and $u_i = 0$ otherwise) and for TC distribution (pyrimidine group, $u_i = 1$ if the site is occupied by a T or C and $u_i = 0$ otherwise). We can also have a distribution similar to that of the DNA walk model [2] with $u_i = 1$ if the site is occupied by a purine and $u_i = -1$ if occupied by a pyrimidine. However, unlike the DNA walk model of interpreting $+1$ or -1 as the probability of step up or step down which has been a subject of controversy [4], ρ_n now gives the distribution of n which represents the excess or deficit of purines over pyrimidines.

Figure 1(a) shows the plot of $\sigma(l)$ versus l in a log-log plot for the above three cases (denoted by GA, TC, and DW) for two typical DNA sequences; human β globin [HUMHBB, an intron-containing sequence of size 73 308 base pairs (bp)] and bacteriophage λ (LAMCG, an intronless sequence of size 48 502 bp). As can be seen, the curves for GA and TC distributions coincide whereas the DW curve shows a parallel shift indicating that the above three procedures result in the same slope α (where $\sigma \approx l^{\alpha}$). For example, we find $\alpha \approx 0.72$ for all three in the case of HUMHBB as found by Peng *et al.* whereas α for LAMCG shows variation with l which is also consistent with the previous findings [5,6]. The above agreement is rather remarkable because the GA or the TC distributions could be different depending on the GA or TC contents (which need not be the same), but both the distributions have similar fluctuations at all scales as that of DNA walk representation. This agreement may look obvious due to Chargaff's second parity rule according to which GA content is approximately equal to the TC contents. However, we have found this agreement even in small cDNA segments where GA and TC contents are quite different. In fact, we have noticed a very general property is that the distribution of any group of nucleotides which has a probability of occurrence p has the same (or nearly the same) variance as that of the distribution of its complementary group that has the probability of occurrence $(1 - p)$, although both have different distribution functions ρ_n [15]. This is an important observation as we can now use either GA or TC distributions as alternatives to the DNA walk representation along with the individual nucleotide distributions to study the correlations. The advantage is, since n represents a sum, unlike the DNA walk model, the entire spectrum lies in the positive half of the axis which is very much essential to compute the factorial moments. Figure 1(b) shows the factorial moments F_q (shown at $q = 2$, 4, and 6) for a random sequence and for LAMCG corresponding to typical GA and G distributions. In case of a random sequence, GA, G, A, T, or C distributions behave similarly, namely, $F_q < F_{q-1} < 1$ at small l and all moments saturate to unity asymptotically. However, in the case of a DNA sequence, $F_q < F_{q-1} < 1$ for $l < l_c$, $F_q > F_{q-1} > 1$ for $l > l_c$, and most of the (lower) moments become unity at around $l = l_c$. The presence of this critical point l_c, where the distribution function is approximately Poissonian, is a unique feature of the DNA sequences which is not found in the random one. Although shown only for the G distribution, this feature is found in A, T, and C distributions also. In fact, the deviation from the Poisson distribution for $l > l_c$ is much stronger in the case of a single nucleotide as compared to the GA distribution. It is also noticed that σ versus l in the log-log plot [see Fig. 1(a)] is not linear over a large distance, rather starts bending, i.e., the slope changes around the same point l_c where the factorial moments become unity.

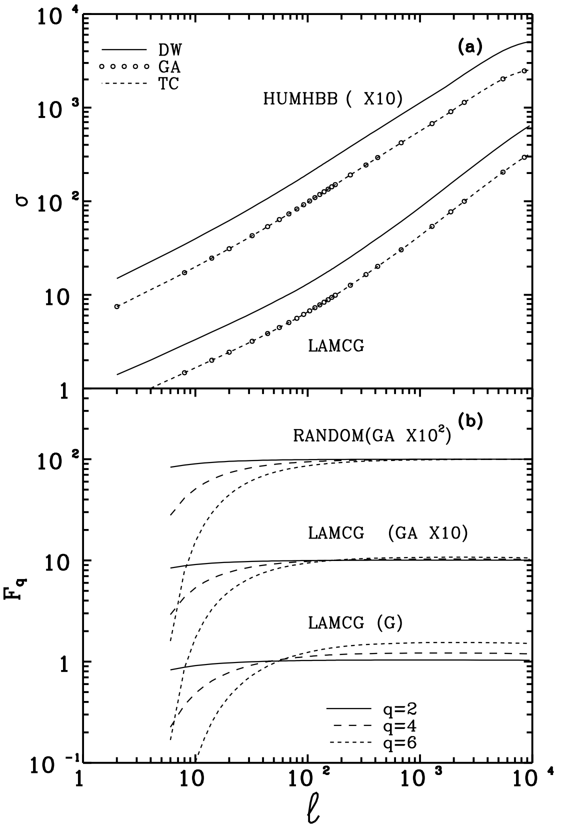

FIG. 1. (a) The variance σ versus l for human β globin (HUMHBB) and bacteriophage λ (LAMCG) with GA (circles), TC (dashed curve), and DW (solid curve) distributions. (b) The factorial moments F_q ($q = 2$, 4, and 6) for a random sequence (GA distribution) and for LAMCG (GA and G distributions).

We study the long range correlations for G, A, T, C, and GA distributions by dividing the entire sequence into two segments around l_c. Figure 2(a) shows the slope α as a function of l for GA as well as for individual nucleotide distributions in case of a typical LAMCG sequence. In

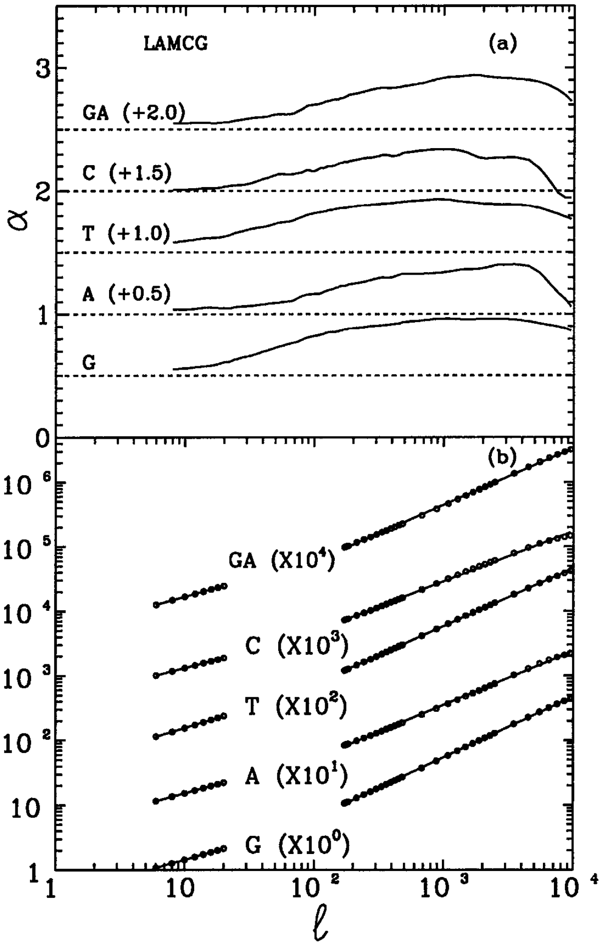

FIG. 2. (a) The slope α versus l for G, A, T, C, and GA distributions (scaled up for clarity) for a typical sequence like LAMCG. (b) The log-log plot for σ versus l.

general, the slope α depends on l [5,6]. However, the above dependence is rather weak as it spreads over a distance of several nucleotides except for the region around l_c, where the effect of slope change is significant. Therefore, the log-log plot as shown in Fig. 2(b) is highly linear with slope α_1 for $l < l_{c1}$ and slope α_2 for $l > l_{c2}$ where l_{c1} and l_{c2} are the minimum and maximum values of all of the l_c of G, A, T, C, and GA distributions of a given sequence. We exclude the region $l_{c1} < l < l_{c2}$ for better linearity. We also choose $l_{\max} \approx L/20$, i.e., at least we have 20 independent data sets so that the statistical analysis becomes meaningful. We have estimated l_c and α values for several cases covering both intronless and intron-containing sequences. A general observation is that for $l < l_c$, the correlation is rather weak, even in some cases, the α_1 values are quite close to 0.5. However, for $l > l_c$, all the sequences irrespective of their intron contents show long range correlations having large α_2 values. These findings for $l > l_c$ are also in agreement with that of Voss [8]. Although they cannot be generalized, the l_c values for intron-containing sequences like HUMHBB are found only a few nucleotides away. Therefore, the nonlinearity in the log-log plot is not very prominent and the average slope obtained without a lower bound will not differ from α_2 significantly [which is around 0.72 as shown in Fig. 1(a)]. However, for the intronless sequence like LAMCG, the l_c value for GA distribution is quite large (as compared to l_c of individual nucleotides) and the sequence shows no or weak correlations up to a distance of several nucleotides in the DNA walk representation even though the correlations in G, A, T, and C still exist. We have also analyzed the *E.coli* sequence both for inter-ORF (Open Reading Frame) and coding regions. The correlations found in the inter-ORF regions are always higher than the values found in coding regions. An interesting aspect is that the slope (α_2) obtained from the GA distribution for the coding region is quite close to 0.5 even for $l > l_c$ although α_2 values for G, A, T, and C are still large (0.65, 0.66, 0.66, and 0.72). Therefore, the effect of correlations may get washed out (or reduced) in any OR spectrum (*GA* spectrum is a OR spectrum where we look for the occurrence of either of the nucleotides G or A) or in any such combinations. For example, the detrended fluctuation analysis of Peng *et al.* [9], which is an improvement over min-max partitioning [2], predicts no correlations up to a length scale of the order of a thousand nucleotides in the coding sequence. This could be sometime misleading as a particular way of fluctuation analysis (i.e., using a specific model like DFA) may not show any long range correlations although it exists for individual nucleotides. Therefore, the correlation analyses for individual nucleotides are more meaningful than for any group or combinations. It has also been noticed that the factorial moments for many sequences start decreasing towards unity at large distances indicating the existence of an upper limit above which the sequence may behave randomly.

As mentioned before, the long range correlations, according to Ref . 12], would imply a strong deviation from Gaussian statistics. However, the investigation of Arneodo *et al.* [13], based on wavelet analysis, yields an important conclusion that the DNA statistics are essentially Gaussian. To resolve this controversy, we study the density distribution ρ_n at various lengths. Figure 3 shows the ρ_n for a cytomegalovirus (AD169, 229 354 bp) at $l = 100$, 200, 400, and 600 for a typical G distribution. The corresponding distributions expected from a random sequence having the same G, A, T, and C contents as that of AD169 are also shown (dotted curves) for comparison. As can be seen, the distributions are nearly Gaussian as that of the random one at small l where it starts deviating from the Gaussian nature at large distances. In case of AD169, l_c for G distribution is about 36 nucleotides away which is much smaller than the l_c of the GA distribution of about 148 nucleotides away. Therefore, the deviation in the G distribution sets in at a relatively smaller distance and also the deviation is stronger as compared to the GA spectrum. The above study

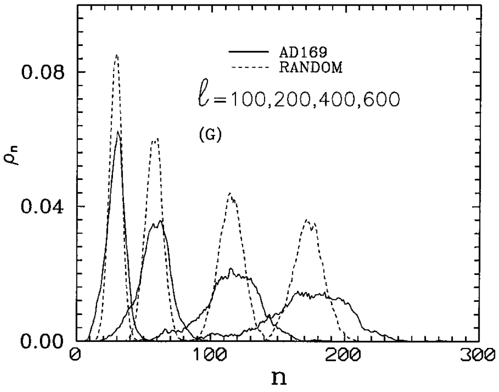

FIG. 3. The density distribution ρ_n for a typical G distribution at different lengths for human cytomegalovirus strain AD169 (chosen for better statistics).

suggests that the DNA statistics is Gaussian for $l < l_c$ and starts deviating from Gaussian nature for $l > l_c$.

In conclusion, we have suggested a new method to study the long range correlations in DNA sequences using the variance analysis of the density distribution of a single or group of nucleotides. The individual nucleotides provide a more fundamental measure of the correlations than any combination or group (like the DNA walk representation) where the effects may get reduced or washed out. The novelty of the present method is the factorial moment analysis of the density distribution which shows the existence of a characteristic length scale around which most of the (lower) moments become unity and also cross the axis showing strong deviation from the Gaussian behaviors. This unique feature which is found for DNA sequences may have important biological significance that needs to be studied further. However, based on this scale we have divided the DNA sequence into two segments to study the long range correlations. It is found that below this scale, the correlation is weak and the DNA statistics is essentially Gaussian, while above this all DNA sequences show strong long range correlations irrespective of their intron contents with a significant deviation from the Gaussian behavior. These findings are consistent with the results obtained based on spectral analyses [8,16] where different correlations are found in different frequency domains. The present analyses, however, are able to provide a quantitative measure of a length scale above which the correlation is significant. Further, the extracted value of α in $1/f^{\alpha}$ to some extent depends on the methodology adopted [8,16]. The present procedure, on the other hand, is simple and free from any ambiguity. This work is also able to resolve many of the controversies that primarily arise due to a specific model. In this context, our analysis is model independent as it involves only the counting of an individual or group of nucleotides in a given length to build the density distribution. Interestingly, the famous model of DNA walk representation is found to be a special case of the present analysis with a density distribution corresponding to either a purine or a pyrimidine group. The present analysis seems to support the hypothesis of the joint action of a deterministic and random process as a possible mechanism of generating the sequences [17] with a few exceptions [15]. However, in this work, we do not advocate for any specific model, rather provide an elegant tool to measure the degree of correlations unambiguously so that the interpretation of data including theoretical analysis will become more meaningful. This work will provide further impetus to develop models for the understanding of the DNA dynamics.

We are thankful to Dr. S. K. Mahajan and S. K. Bhattacharjee for many illuminating discussions during the course of this work.

[1] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992); W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); W. Li, T. Marr, and K. Kaneko, Physica (Amsterdam) **75D**, 392 (1994).

[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] J. Maddox, Nature (London) **358**, 103 (1992).

[4] S. Nee, Nature (London) **357**, 450 (1992).

[5] C. A. Chatzidimitriou-Dreismann and D. Larhammar, Nature (London) **361**, 212 (1993).

[6] V. V. Prabhu and J. M. Claverie, Nature (London) **359**, 782 (1992).

[7] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[8] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[9] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[10] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, F. Sciortino, and H. E. Stanley, Phys. Rev. Lett. **71**, 1776 (1993).

[11] The long range correlation does not necessarily imply a deviation from Gaussianity. For example, the fractional Brownian motion which has Gaussian statistics shows an inverse power law spectrum.

[12] P. Allegrini, P. Grigolini, and B. J. West, Phys. Rev. E **54**, 4760 (1996).

[13] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[14] A. K. Mohanty and S. K. Kataria, Phys. Rev. Lett. **73**, 2672 (1994); **75**, 2449 (1995); Phys. Rev. C **53**, 887 (1996).

[15] A. K. Mohanty and A. V. S. S. Narayana Rao (to be published).

[16] M. de Vieira, Phys. Rev. E **60**, 5932 (1999).

[17] P. Allegrini, M. Barbi, P. Grigolini, and B. J. West, Phys. Rev. E **52**, 5281 (1995).