Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Statistical scales of order in DNA

Douglas Poland [*]

Department of Chemistry, The Johns Hopkins University, Baltimore, MD 21218, USA

ARTICLE INFO

Article history:
Received 13 January 2009
Received in revised form 3 February 2009
Accepted 3 February 2009
Available online 20 February 2009

Keywords:
DNA
Maximum entropy distributions
Persistence exponent
Fractal dimension
Power law
Mandelbrot distribution
Model sequence

ABSTRACT

In the present paper we examine the statistics of occurrence of A–T and C–G base pairs in DNA. We focus on the net base composition in blocks of base pairs of various sizes. This paper extends our previous work on randomness and order in DNA sequences and examines order on various scales. For structure on the local scale (10^0–10^1 bp) we have seen that the net base composition in given block sizes is fitted very accurately by the discrete binomial distribution for a random system. If the statistics were random for larger block sizes then the appropriate distribution would be the standard normal (Gaussian) distribution which is the continuous analog of the discrete binomial distribution. However, we have found that at the intermediate scale (10^2–10^4 bp) the composition distribution is not fit by a standard normal distribution but rather by a modified normal distribution with a standard deviation that is a marked nonrandom function of block size. In particular, the standard deviation accurately follows a power law with a characteristic exponent. This behavior can be interpreted in terms of a random walk model due to Mandelbrot that is characterized by a tendency for the walk to persist in direction. The DNA analog of the walk model is the tendency of blocks of base pairs with a given net composition to be followed by blocks of a similar composition (persistence of composition). A model based on a generating function constructed from a matrix of conditional probabilities (incorporating persistence) explains the overall order in a given genome at the intermediate scale. In the present paper we examine the block statistics in DNA using the genomes of two organisms, namely *Bacillus anthracis* and *Escherichia coli* both of which have a chain length of slightly over five million base pairs. We find that the distributions in *B. anthracis* are well fit by a Mandelbrot-like distribution. On the other hand, the distributions in *E. coli* are not so well fit by this distribution which is based on two moments. Using the maximum-entropy method we construct an improved distribution for *E. coli* based on four moments. Finally we look at the order on the scale of the entire molecule (global scale). Applying the model of a random walk to the complete DNA genome we find that the Mandelbrot distribution on an intermediate level cannot explain the global character of the random walk, there being structure to the walk with features on the scale of the total length of the molecule (10^5–10^7 bp). To understand the three scales of order (local, intermediate and global) we construct a model sequence based on the incorporation of Mandelbrot-type order on the intermediate scale in a single size block. We then find that the character of the order on the local and global scales follows naturally from this single feature. Thus all three scales of order in DNA are incorporated into our model sequence.

1. Introduction

The base sequence of the DNA for a given species contains the genetic information required to synthesize the proteins necessary for the existence of the particular species. In a series of papers [1–4] we have examined a number of DNA sequences from the point of view of the statistics of the net occurrence of A–T and C–G base pairs. What we have found is that while the local statistics are essentially the same as those for a random collection of A–T and C–G base pairs, the statistics for longer sequences of base pairs show very nonrandom behavior. Specifically, there is a nonlinear power law with a characteristic exponent that describes the widening of the base pair distributions.

In this paper we explore the nature of the statistical structure of DNA with respect to the occurrence of base pairs on three scales of length. The scales we use are: local (10^0–10^1 bp), intermediate (10^2–10^4 bp) and global (10^5–10^7 bp. In Section 2 we use the genomes of *Bacillus anthracis Ames* [5] and *Escherichia coli CFT073* [6], which are available on the web [7], to examine the local block distribution functions for these organisms. We will often abbreviate the names of these organisms to BA and EC respectively. As with the genomes we have studied previously [1,2], we find that the distribution functions for these species on the local level are essentially identical with the binomial distribution, indicating random behavior.

If one increases the block size to the intermediate range then we have shown [1,2] that the block distributions are very nonrandom. In Sections 3–5 we illustrate this behavior, again, using the genomes of *B. anthracis* and *E. coli*. If the net composition of blocks in the intermediate range had a random distribution then it would be described by a normal (Gaussian) distribution which is the continuous analog of the discrete binomial distribution we used on the local scale.

[*] Tel.: +1 410 516 7441; fax: +1 410 8420.
E-mail address: poland@jhu.edu.

0301-4622/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2009.02.003

The distribution function that is actually found in the intermediate range is similar to the normal distribution but has a much larger value of the standard deviation giving a wider distribution (reflecting persistence) than the normal distribution. The widening of the distribution function is described accurately by a power law in terms of the block size. The nonrandom behavior can be understood in terms of a fractal random walk model due to Mandelbrot [8,9] which describes the distribution of steps when there is a persistence in the current direction of the walk. The analogy of the random walk model with the power law behavior in DNA is that in DNA the tendency for walks to persist translates into a persistence of composition whereby blocks of a given net composition tend to be followed by blocks of like composition.

The distributions for *B. anthracis* are accurately described by this modified normal distribution while those of *E. coli* are not so well described, tending to be asymmetric. Using the maximum-entropy method we can construct an accurate distribution function for *E. coli* using four moments of the distribution (in contrast to the two moments required for the normal distribution). Using the notion of persistence we construct a generating function in terms of a matrix of conditional probabilities that gives an increasingly good representation of the power law behavior found at the intermediate level as the size of the matrix (reflecting the range of persistence) is increased.

In Section 6 we examine the structure on the whole molecule (global) scale. We turn the genome sequence into a random walk by treating an A–T base pair as a step in the forward direction while a C–G base pair represents a negative step. One simply starts at one end of the molecule and takes consecutive steps indicated by the base composition. We find that on the global scale there is a pattern to the overall walk that cannot be generated by the distribution function for order at the intermediate level.

Finally, in Section 7, we construct a model sequence that exhibits all types of order found in DNA: local (random), intermediate (power law) and global (whole molecule pattern). This exercise shows what features are required to reproduce the kinds of order found in DNA.

2. Local order

In this section we explore the local statistics of DNA sequences using as examples the genomes of *B. anthracis* and *E. coli*. The genomes of these bacteria both consist of a single circular chromosome and contain slightly more than five million base pairs (bp) which is about three times the size of the genomes we have previously examined. The basic statistics for these two organisms are:

$$\begin{aligned} \text{BA}: \quad & \text{number of base pairs} = 5,227,290 \\ & \text{fraction of A–T base pairs} = 0.6462 \\ & \text{fraction of C–G base pairs} = 0.3538 \end{aligned} \tag{1}$$

$$\begin{aligned} \text{EC}: \quad & \text{number of base pairs} = 5,231,428 \\ & \text{fraction of A–T base pairs} = 0.4952 \\ & \text{fraction of C–G base pairs} = 0.5048 \end{aligned} \tag{2}$$

For EC we have counted nonstandard base pairs as C–G base pairs for simplicity. We note that for BA the fraction of A–T base pairs is almost twice the fraction of C–G base pairs while for EC the fractions are almost equal.

We will use two parameters to describe the base sequence of a genome, namely

$$\begin{aligned} f_0 &= \text{fraction of A plus T} \\ f_1 &= \text{fraction of C plus G} \end{aligned} \tag{3}$$

where

$$f_0 + f_1 = 1. \tag{4}$$

We will refer to these two states as 0-states and 1-states respectively. To illustrate the behavior of the local statistics in DNA we examine the first million base pairs in BA. For this sequence we count the fractions of the different possible singlets and doublets (nearest neighbor pairs), obtaining the following results

$$f_0 = 0.635, \quad f_1 = 0.365 \tag{5}$$

$$f_{00} = 0.411, \quad f_{11} = 0.141, \quad f_{01} = f_{10} = 0.224 \tag{6}$$

Using this data we then construct the following matrix which is an indicator of the randomness of the sequence

$$\begin{pmatrix} f_0 f_0 / f_{00} & f_0 f_1 / f_{01} \\ f_1 f_0 / f_{10} & f_1 f_1 / f_{11} \end{pmatrix} = \begin{pmatrix} 0.981 & 1.035 \\ 1.035 & 0.945 \end{pmatrix} \tag{7}$$

If the sequence is random then all of the matrix elements will be equal to one. One sees that the deviation from random behavior is of the order of a few percent.

Another statistic that illustrates deviation from random behavior is the average length of a run of 0s or 1s. We first calculate the number of different states in a sequence N bases long

$$N_0 = Nf_0, \quad N_{01} = Nf_{01}, \quad N_1 = Nf_1. \tag{8}$$

The average sequence lengths are then given by

$$\langle n_0 \rangle = \frac{N_0}{N_{01}} = \frac{f_0}{f_{01}}, \quad \langle n_1 \rangle = \frac{N_1}{N_{01}} = \frac{f_1}{f_{01}}. \tag{9}$$

If the sequence is random, with $f_{01} = f_0 f_1$, then one has

$$\langle n_0 \rangle^* = \frac{1}{f_1} \quad , \langle n_1 \rangle^* = \frac{1}{f_0}. \tag{10}$$

Taking the ratios of the quantities given in Eq. (9) to those in Eq. (10) we obtain the result

$$\frac{\langle n_0 \rangle}{\langle n_0 \rangle^*} = \frac{f_0 f_1}{f_{01}} = \frac{\langle n_1 \rangle}{\langle n_1 \rangle^*} = 1.036 \tag{11}$$

which indicates that there is a very slight tendency for the states to cluster relative to the clustering found in a random sequence. For the sample we are using from BA the sequence lengths are as follows

$$\begin{aligned} \langle n_0 \rangle &= 2.83(2.74) \\ \langle n_1 \rangle &= 1.63(1.57) \end{aligned} \tag{12}$$

where the numbers in parentheses give the respective quantities for a random sequence.

The above examples indicate that on the local level of nearest-neighbor statistics, DNA exhibits random statistics. To see if random behavior persists beyond the nearest-neighbor level we now examine the distribution function for a larger section of the molecule. To this end, we consider consecutive, non-overlapping blocks of *m* base pairs and count the number of blocks that contain [*n*] 1-states. If the *m*-block statistics are random then all possible combinations of the two states are generated by the simple product

$$\Gamma_m = (f_0 + f_1)^m. \tag{13}$$

The expansion of the generating function given in Eq. (13) gives the binomial distribution. The probability of having [*n*] 1-states in a block of *m* base pairs is then simply the appropriate binomial coefficient

$$p_n = \frac{m!}{(m-n)!n!} f_0^{m-n} \, f_1^n \tag{14}$$

where *n* can vary from zero to *m*.

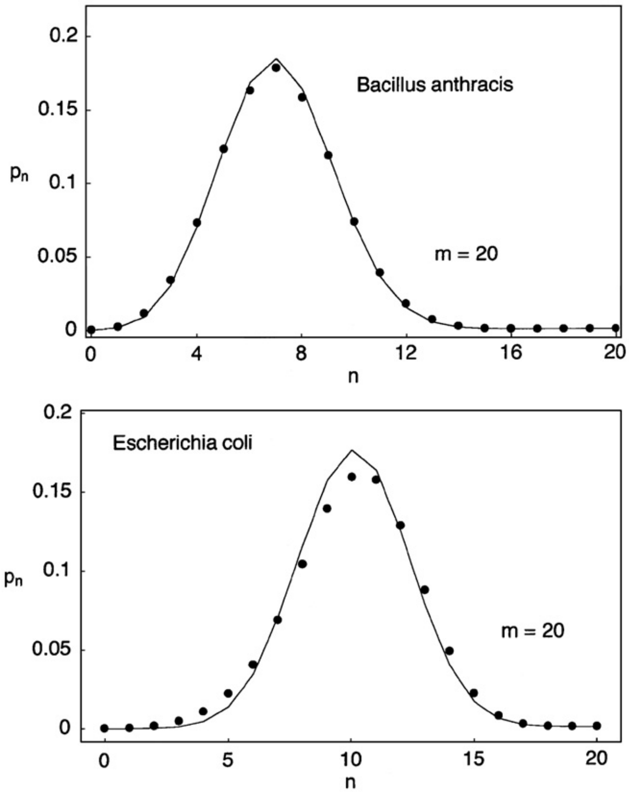

Fig. 1. Probability distributions for BA and EC giving the probability that a block of 20 base pairs will have [n] 1-states. The solid dots give the empirical distributions while the solid lines connect the points given by the random distribution of Eq. (14). The appropriate parameters for each of the species are given in Eqs. (1) and (2).

In Fig. 1 the solid dots give the empirical distributions for BA and EC. The solid lines connect the points obtained from the binomial distribution given in Eq. (14). The results given in Fig. 1 show that the empirical distributions for m-blocks with $m=20$ are essentially identical with the distribution functions for random sequences. In the next section we examine the question of how large we can make m and still obtain random behavior.

3. Intermediate order

In this section we examine the order in DNA in the range of 10^2 to 10^4 bp. In particular we want to determine whether or not the empirical m-block distribution determined by direct counting is the same as the distribution function for a random sequence as given by Eq. (14) and illustrated in Fig. 1. We have already explored this question in previous publications [1,2]. Here we review the essential points of the approach as applied to BA and EC.

The simplest way to determine the basic structure of a distribution function is through the moments of the distribution. The first two moments of the m-block distribution are given by the following standard relations

$$M_1(m) = \sum_{n=0}^{m} np_n \qquad M_2(m) = \sum_{n=0}^{m} n^2 p_n \tag{15}$$

where p_n again is the probability of having [n] 1-states in an m-block. If we know the generating function for the system then we can calculate the moments directly from that. In order to count 1s we insert a label parameter z for every 1-state in the system giving the generating function $\Gamma_m(z)$. For example, for random units we have the following modified generating function for a general m-block

$$\Gamma_m(z) = (f_0 + zf_1)^m. \tag{16}$$

Writing z in terms of an exponent

$$z = e^a \tag{17}$$

we have the following general relations for the moments

$$M_1(m) = \frac{\partial \Gamma_m}{\partial a} \qquad M_2(m) = \frac{\partial^2 \Gamma_m}{\partial a^2}. \tag{18}$$

We note that the use of Eq. (18) is not restricted to random systems. After taking the appropriate derivatives we set $a=0$ (which is equivalent to setting $z=1$).

For random systems we use Eqs. (16)–(18) and obtain the following relations giving the moments as a function of the block size (m):

$$M_1(m) = mf_1 \qquad M_2(m) = mf_1 + m(m-1)f_1^2 \tag{19}$$

We then construct the standard deviation of the distribution using the general relation

$$\sigma = \sqrt{M_2 - M_1^2}. \tag{20}$$

Using Eq. (19) in Eq. (20) gives

$$\sigma = \sqrt{mf_0 f_1}. \tag{21}$$

We then define the following function that is crucial for our understanding of long range order in DNA

$$\zeta(m) = \frac{\sigma}{\sqrt{m}}. \tag{22}$$

For random sequences we combine Eqs. (21) and (22) giving

$$\zeta = \sqrt{f_0 f_1} \tag{23}$$

which is independent of m (the block size). The deviation of zeta from the value given in Eq. (23) will then be an indication of deviation from random behavior.

To obtain $\zeta(m)$ from the BA and EC genomes we proceed as with the $m=20$ example treated in the previous section. To review, we divide the molecule up into consecutive, non-overlapping blocks each containing m base pairs. We then count the number of 1-states in each block giving the p_n, the block distribution function. Next we use the p_ns in Eq. (15) to give the moments M_1 and M_2. We then calculate σ, the standard deviation of the distribution, using Eq. (20) and, finally, we calculate $\zeta(m)$ using Eq. (22).

The results of this calculation for BA and EC are shown in Fig. 2 where the solid dots give the values of $\zeta(m)$ for $m=100$ to $m=12{,}000$ in steps of 100. If the block distribution functions were random then ζ should be independent of m with the values given by Eq. (23) and shown by the dashed curves in Fig. 2. The fact that $\zeta(m)$, as illustrated in Fig. 2, is a strong function of m indicates that there is a very large deviation from randomness as m is increased. Thus on a local scale the statistics indicate a random sequence while on a large scale the distribution is very nonrandom. This feature (locally random,

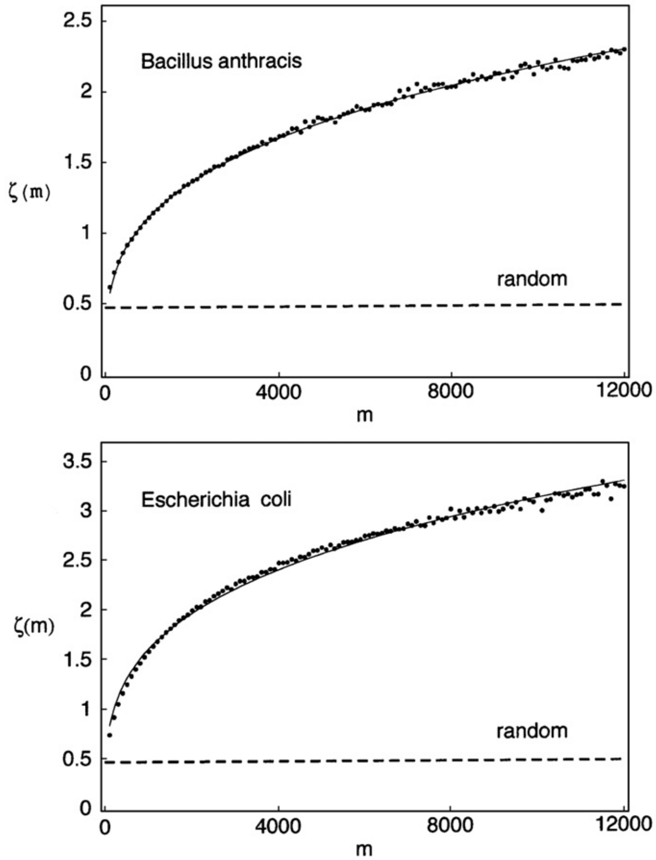

Fig. 2. Plots of the function $\zeta(m)$ defined in Eq. (22) for *Bacillus anthracis* and *Escherichia coli* where m is the block size. The solid curves are the respective power laws defined by Eq. (24) with the parameters of Eqs. (26) and (27). The dashed curves represent the behavior of ζ for random sequences given by Eq. (23).

nonrandom on an intermediate scale) is characteristic of all of the genomes that we have examined.

The behavior of $\zeta(m)$ shown in Fig. 2 suggests a power law of the following form

$$\zeta = Am^{\gamma}. \tag{24}$$

To test this hypothesis one can plot ln ζ as a function of ln m. The slope and intercept of such a plot give ln A and the exponent γ (the persistence exponent), respectively.

A further test of the power law behavior is given using the function

$$R(m) = \frac{\zeta(m)}{m^{\gamma}}. \tag{25}$$

If the power law holds, then we expect $R(m) = A$ for all m. The quantity $R(m)$ is plotted in Fig. 3 for BA and EC and one sees that it is indeed constant for all values of m (block size).

The parameters A and γ obtained in this manner for BA and EC are as follows

$$\begin{aligned}\text{BA:}\quad A &= 0.1548\\ \gamma &= 0.2866\end{aligned} \tag{26}$$

$$\begin{aligned}\text{EC:}\quad A &= 0.2282\\ \gamma &= 0.2841\end{aligned} \tag{27}$$

One notes that the values of γ are approximately the same for the two species. Previously we have obtained a similar value of $\gamma(0.290)$ for the organism *Thermoplasma volcanium* [1].

4. Distributions for intermediate order

In order to understand the behavior of the ζ function we examine the distribution function for the net base composition of blocks as a function of block size. In Eqs. (13) and (14) we have already given the distribution function for a discrete random sequence of base pairs. The analog for a random distribution in continuous space is the standard normal distribution. To distinguish the random distribution from other similar distributions we will refer to this as the Gaussian distribution. It is given by the following function

$$P_G(x) = \frac{1}{\sqrt{2\pi}\,\sigma_G}\exp\left[-(x-x_o)^2 / 2\sigma_G^2\right] \tag{28}$$

where x is the continuous analog of the index n (the number of 1-states in an m-block used in Eq. (14)). The quantity x_o is the mean value of x while the quantity σ_G is the standard deviation as given by Eq. (21) which we repeat here:

$$x_o = mf_1 = M_1(m), \quad \sigma_G = \sqrt{mf_0 f_1}. \tag{29}$$

The Gaussian distribution given in Eq. (28) works well for small values of m (e.g. $m = 20$) while it is much too sharp for large values of m (e.g. $m = 400$).

For values of m in the intermediate range we follow Mandelbrot's treatment of random walks with persistence [8,9] and use a modified normal-like distribution which differs from the Gaussian distribution in having a σ function that is constructed from $\zeta(m)$. Using

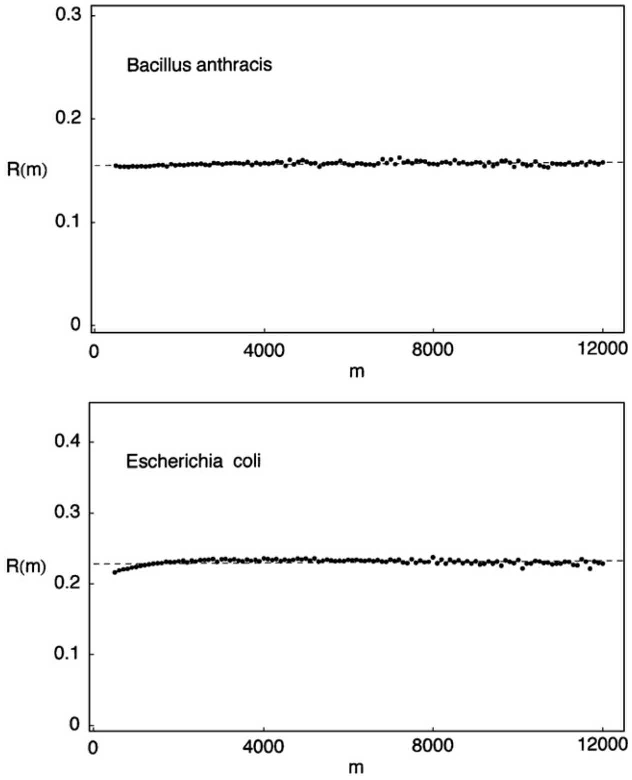

Fig. 3. Plots of the function $R(m)$ defined in Eq. (25) as a function of m based on the data given in Fig. 2. The dashed curves give the constant values of A given in Eqs. (26) and (27).

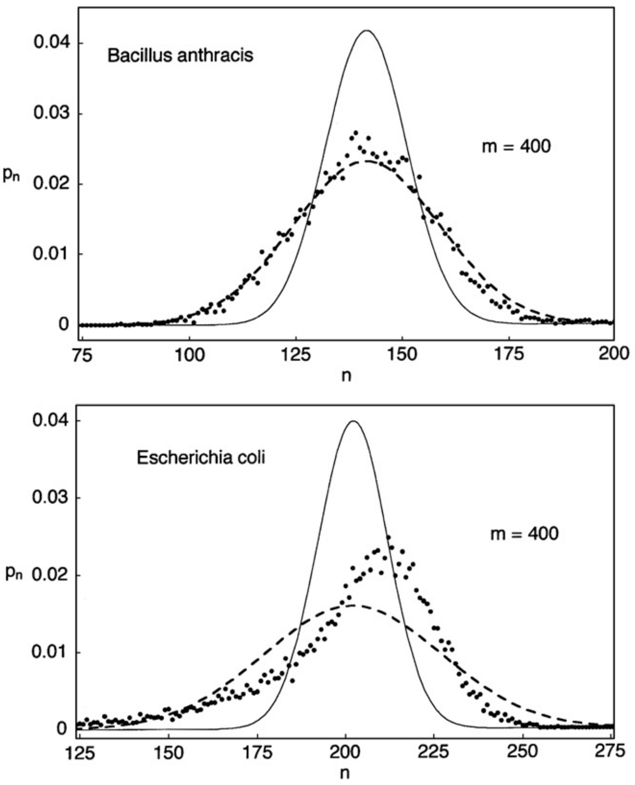

Fig. 4. Probability distributions for BA and EC giving the probability that a block of 400 base pairs will have [n] 1-states. The solid dots give the empirical distributions while the solid lines give the respective Gaussian distributions of Eq. (28). The dashed lines give the respective Mandelbrot distributions of Eq. (32).

Eqs. (22) and (24) we solve for σ (which we designate with the subscript M for Mandelbrot)

$$\sigma_{\mathrm{M}} = \sqrt{m}\,\zeta(m) = Am^{H} \tag{30}$$

where

$$H = \gamma + 1/2. \tag{31}$$

The Mandelbrot distribution is then given by the function

$$P_{\mathrm{M}}(x) = \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{M}}} \exp\left[-(x - x_o)^2 / 2\sigma_{\mathrm{M}}^2\right] \tag{32}$$

This function looks superficially like the Gaussian distribution given in Eq. (28), but these two functions differ markedly with respect to the sigma functions given in Eqs. (29) and (30). In particular $P_{\mathrm{M}}(x)$ is a much broader distribution than $P_{\mathrm{G}}(x)$. Using the values of gamma for BA and EC given in Eqs. (26) and (27) we obtain the following values for the H parameter given in Eq. (31)

$$H_{\mathrm{BA}} = 0.787, \quad H_{\mathrm{EC}} = 0.784. \tag{33}$$

To illustrate the difference in the functions $P_{\mathrm{G}}(x)$ and $P_{\mathrm{M}}(x)$ we compare these functions with the empirical distributions obtained by direct counting of the compositions of successive m-blocks. In Fig. 4 we show these distributions for the case of $m = 400$ for BA and EC. Again the solid dots indicate the empirical results obtained by scanning the genomes of the respective species. The solid curve in both graphs gives the distribution $P_{\mathrm{G}}(x)$ as given by Eq. (28) with sigma given by Eq. (29). These functions give the distributions for random sequences and one sees in both graphs that the empirical distributions are much broader than the corresponding $P_{\mathrm{G}}(x)$ distributions. Finally, the dashed curves in both graphs give the behavior of the $P_{\mathrm{M}}(x)$ distribution of Eq. (32). For the case of BA the $P_{\mathrm{M}}(x)$ distribution is seen to fit the empirical data extremely well. However, for the case of EC the empirical data show a clearly asymmetric form (the function $P_{\mathrm{M}}(x)$ is symmetric about the mean value of x). We will construct a modified distribution that will give a good fit for the empirical data of EC shortly. But first we further illustrate how well the Mandelbrot distribution works for the BA data.

In Fig. 5 the solid dots give the empirical distributions for the cases of $m = 100$ and $m = 200$ for BA. The solid curves give the corresponding Mandelbrot distributions which are seen to give an outstanding fit to the data. We now show that we can scale the distributions so that the two become essentially identical. To do this we introduce a new variable

$$y = \frac{x - x_{\mathrm{o}}}{\sigma_{\mathrm{M}}} \tag{34}$$

with

$$\mathrm{d}y = \frac{1}{\sigma_{\mathrm{M}}}\mathrm{d}x \tag{35}$$

where σ_{M} is given by Eq. (30). Substituting these forms in Eq. (32) one obtains the simplified distribution

$$P(y) = \frac{1}{\sqrt{2\pi}} \exp\left[-y^2/2\right]. \tag{36}$$

For the discrete empirical points we make a similar change of variable.

Using the above scaling relations we can collapse the upper curves in Fig. 5 into one distribution as shown in the lower curve (where the solid line gives the distribution of Eq. (36) and the solid

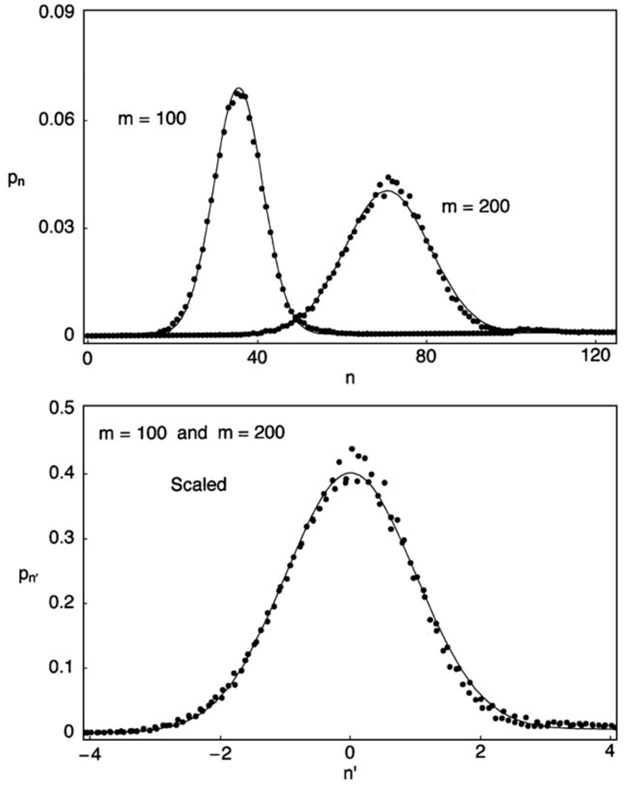

Fig. 5. The upper graph shows the empirical probability distribution (solid dots) for the cases $m = 100$ and $m = 200$. The solid curves give the respective Mandelbrot distributions of Eq. (32). The lower graphs show the same distributions as treated by the scaling relations of Eqs. (34)–(36).

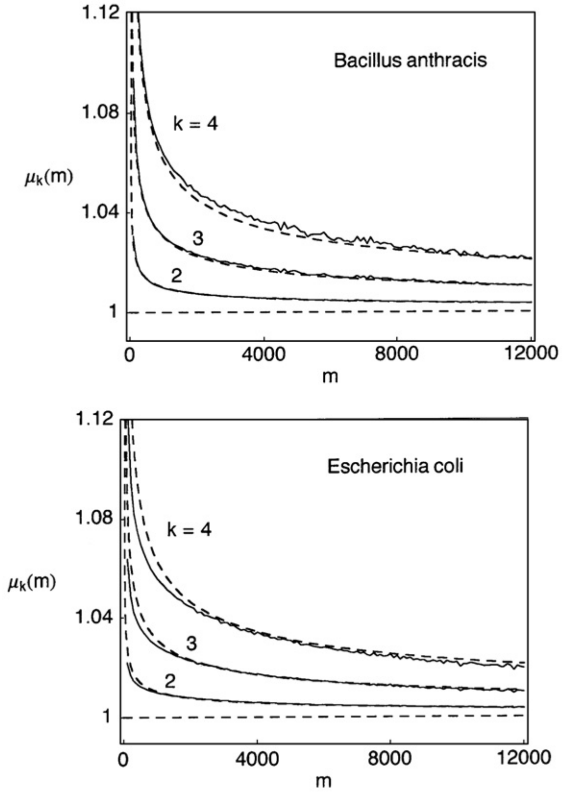

Fig. 6. The first four relative moments for BA and EC. The solid lines give the empirical results while the dashed curves give the results of using the Mandelbrot distribution.

dots give the empirical points for $m = 100$ and $m = 200$, also scaled). The results shown in Fig. 5 indicate that the block distributions for BA are well described by the Mandelbrot distribution, $P_M(x)$, as given by Eq. (32) with the width of the distribution given by Eqs. (30) and (31).

We now return to the asymmetric distribution of the EC empirical data for the case of $m = 400$ as shown in Fig. 4. In order to introduce asymmetry into the distribution we need to include the influence of higher moments. It is convenient to consider the following relative moments

$$\mu_k(m) = M_k(m) / M_1(m)^k \tag{37}$$

where (the sum replaced by an integral for continuous distributions)

$$M_k(m) = \sum_{n=0}^{m} n^k p_n. \tag{38}$$

The first four relative moments for the Mandelbrot distribution are as follows

$$\mu_1 = 1, \quad \mu_2 = 1 + r, \quad \mu_3 = 1 + 3r, \quad \mu_4 = 1 + 6r + 3r^2 \tag{39}$$

where

$$r = \left(\frac{Am^{H-1}}{f_1}\right)^2. \tag{40}$$

For the case of a random distribution one has the same relations as given in Eq. (39) with

$$\gamma = 0, \quad H = 1/2, \quad A = \sqrt{f_0 f_1}, \quad r = \frac{1}{m}\frac{f_0}{f_1}. \tag{41}$$

In Fig. 6 we compare the relative moments obtained from the empirical distributions (dashed curves) with the moments for the respective Mandelbrot distribution for BA and EC. The Mandelbrot distributions for both species show a good fit to the data.

Given the first four moments of the distribution we can now construct a higher order distribution. To do this we apply the maximum-entropy method which has been used [10–14] to construct distribution functions for proteins and other systems. The maximum-entropy distribution, $P(x)$, is the exponential of a finite polynomial in x

$$P(x) = Exp\left[-\left(a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4\right)\right]. \tag{42}$$

One has as many powers of x in the exponential as one has moments (the term a_0 is determined by requiring that the distribution is normalized to unity). Thus given four moments one has a quartic in x in the exponential. If one had only two moments then one would have a quadratic in the exponential and this would be exactly the same as the standard normal distribution. Thus one trades the numerical values of a set of moments for the numerical values of a set of parameters in the distribution function. The calculation of the coefficients requires a simple iterative procedure.

In Fig. 7 we reproduce the empirical data (solid dots) of EC for the case of $m = 400$. The dashed curve gives the Mandelbrot distribution of Eq. (32) based on two moments while the solid curve gives the maximum-entropy distribution of Eq. (42) based on four moments. The latter distribution now gives a good fit to the data, reproducing the asymmetric form of the empirical distribution quite well.

5. Persistence and intermediate order

We have seen in the previous section that the function $\zeta(m)$ shows strong nonrandom behavior and that this can be related to a broadening of the m-block distributions as shown in Fig. 4. In turn this can be related to the standard deviation of the Mandelbrot distribution as shown in Eq. (32). Following Mandelbrot's fractal model of a random walk with persistence (i.e. the walk tends to continue in the same direction for more steps than a purely random walk) the broadening of the $P_M(x)$ function can be attributed to the tendency for blocks of like or near-like content of one-states to follow one another (hence the origin of persistence).

As we have shown previously [1] we can graphically show the presence of persistence in DNA by examining the base composition of consecutive m-blocks. We record how many blocks contain [i] 1-states (singlet distribution) and how many blocks containing [i] 1-states are

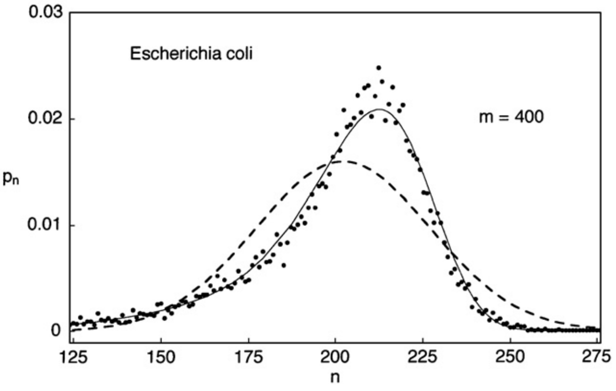

Fig. 7. The probability distribution for $m = 400$ for EC. The solid dots give the empirical results while the dashed curve gives the result of the Mandelbrot distribution based on two moments. The solid curve gives the maximum-entropy distribution of Eq. (42) based on four moments.

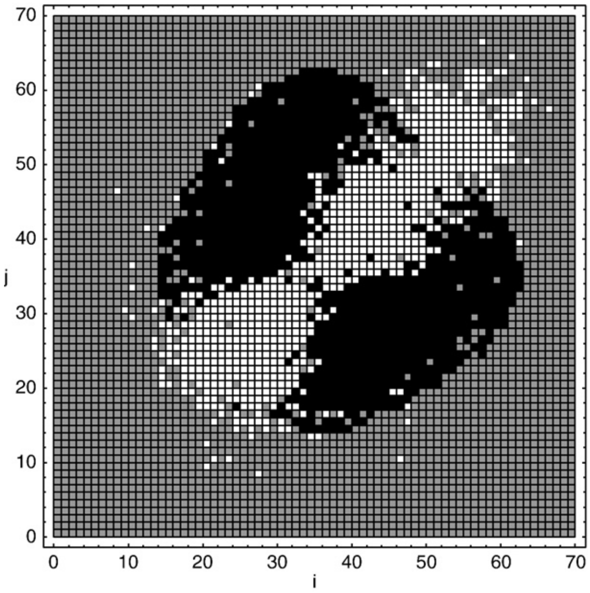

Fig. 8. The difference function defined by Eq. (43) for BA with $m = 100$. The color code is given in Eqs. (44) and (45). The white blocks indicate that like blocks tend to follow one another while the black blocks indicate that unlike blocks tend to follow one another thus illustrating the phenomena of persistence in this DNA.

followed by a block containing [j] 1-states (doublet distribution). The number of singlets and doublets are expressed as the frequencies of the various possible states, namely, f_i and f_{ij}. We then construct an $(m+1)\times(m+1)$ table that gives the difference between the empirical doublet frequencies and the doublet frequencies for a random placement of m-blocks where

$$w_{ij} = f_{ij} - f_i f_j. \tag{43}$$

We convert the numbers w_{ij} into color codes as indicated below

$$w_{ij} > 0(\text{white}), \quad w_{ij} = 0(\text{gray}), \quad w_{ij} < 0(\text{black}). \tag{44}$$

An $(m+1)\times(m+1)$ grid of colors constructed as outlined in Eq. (44) for the case of $m = 100$ for BA is shown in Fig. 8. The color codes have the following simple interpretation

$$\begin{array}{l} w_{ij} > 0(\text{white—like blocks follow one another}) \\ w_{ij} < 0(\text{black—unlike blocks follow one another}). \end{array} \tag{45}$$

One sees clearly that the white states (positive correlation of like states) tend to occur along the $j = i$ axis while the black states (negative correlation of like states) tend to occur along the $j = 70 - i$ axis (perpendicular to the other axis). Thus the results shown in Fig. 8 clearly indicate the presence of persistence in the BA genome.

The results shown in Fig. 8 give dramatic qualitative evidence of persistence in DNA sequences. We now will incorporate this effect into a quantitative procedure. In Eq. (13) we used a generating function for a random sequence to obtain all possible random configurations. Here we will pursue a similar approach, but in this case we construct a generating function that is a product of matrices that contain information about the correlation of consecutive blocks. We have used this approach previously for small genomes [1,2].

We begin by picking a reference block size which we will designate as m_0. We then collect data on the occurrence of singlet and doublet frequencies in blocks of size m_0. Thus we have f_i and f_{ij} for the range $i = 0$ to m_0. We then construct the following $(1+m_0)\times(1+m_0)$ matrix

$$\mathrm{P} = \left[\frac{f_{ij}}{f_i} z^{j-1}\right] \tag{46}$$

where

$$\frac{f_{ij}}{f_i} = P(i|j) \tag{47}$$

is the conditional probability that given i, j follows. If the sequence is random (independent) then $f_{ij} = f_i f_j$ and $P(i|j) = f_j$. In Eq. (46) z is a label to count 1-states just as in Eq. (16).

We next construct two vectors. The first is a vector of the a priori probabilities of the possible given compositions (number of 1-states) in an m_0-block

$$\boldsymbol{p} = \left[f_j z^{j-1}\right]. \tag{48}$$

The second is simply a vector of 1s

$$\boldsymbol{v} = [1]. \tag{49}$$

We then consider n consecutive blocks of size m_0 ($n = 1, 2,\ldots$) with

$$m = m_0 n. \tag{50}$$

Finally we construct the generating function for moments as a matrix product

$$\Gamma_m = \boldsymbol{p}\,\mathrm{P}^{n-1}\boldsymbol{v}^{+} \tag{51}$$

where $\boldsymbol{v}^{+}$ is the transpose of the vector $\boldsymbol{v}$ to give a column vector. We then use Eqs. (17)–(22) to obtain $\zeta(m)$. We consider m values in integer blocks of m_0 with a maximum value of $m = 12{,}000$. The values of n that we use in Eq. (50) vary in the range of $n = 1$ to $n = m_{\max/}m_0$ with $m_0 = 200$, 500, 1000, 2000.

The results of these calculations using BA data are shown in Fig. 9 where the solid dots give the values of the $\zeta(m)$ for the values of m and m_0 used. The cases of a random sequence, using the values given by Eq. (23) and the power law of Eq. (24) for BA are shown for reference by the dashed curves. One sees that as m_0 is increased the matrix product gives a better and better fit to the power law behavior of $\zeta(m)$. This result indicates that the power law-like behavior is a consequence of the persistence correlation between neighboring blocks. For large values of n, the generating function given in Eq. (51) will be asymptotic to the largest eigenvalue of $\mathbf{P}$ raised to the nth power. Using Eqs. (18), (20), (22) one finds that in the large-n limit, $\zeta(m)$ is asymptotic to a constant. Thus Eq. (51) can never truly give a power law for large m-values. However in

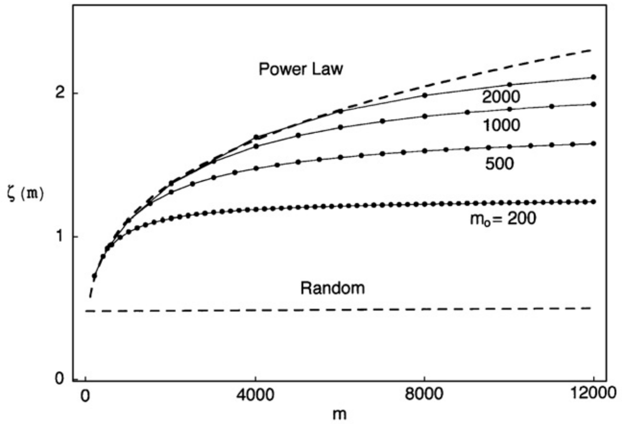

Fig. 9. The behavior of the function $\zeta(m)$ obtained from the matrix of conditional probabilities given by Eq. (46) using the empirical doublet frequencies obtained from the sequence of BA for varying size reference blocks, m_0. The solid curve is the power law defined by Eq. (24) with the parameters of Eq. (26) while the dashed curve represents the behavior of ζ for random sequences given by Eq. (23).

the intermediate-*m* range the matrix product for the generating function gives a good approximation to the empirical variation of $\zeta(m)$.

6. Global order

In this section we look at order in DNA on the scale of the entire genome (10^5–10^7 bp) using the model of a random walk. As Mandelbrot points out, a random walk generates a fractal structure (similar behavior on all scales) with a characteristic fractal dimension. The standard random walk in one dimension describes the path a growing chain takes when at each interval the new unit in the chain can take a step in the forward direction (+1) or in the reverse direction (−1) at random. The analogy with a DNA sequence is based on the fact that there are two possible states per unit (one for C–G pairs and zero for A–T pairs.) These two states can be mapped onto the parameters for a random walk as follows

$$\begin{aligned}\alpha_i &= -1(\text{for } C \text{ or } G)\\ \alpha_i &= +1(\text{for } A \text{ or } T).\end{aligned} \tag{52}$$

After *n* steps (*n* base pairs) the locus of the *n*th unit is then given by summing the appropriate values of the αs. One has simply

$$L(n) = \sum_{i=1}^{n} \alpha_i - n\Delta f \tag{53}$$

where

$$\Delta f = f_0 - f_1. \tag{54}$$

The inclusion of the Δ*f* term in Eq. (53) gives the walk relative to the average composition. Its presence guarantees that $L(N) = 0$ for the last unit in the walk.

We first calculate sample walks for systems obeying the Gaussian (Eq. (28)) and the Mandelbrot (Eq. (32)) distributions. We will construct the walks based on blocks containing m_o base pairs. For a given block number *j* we select a value of $x = x_j$ at random but distributed according to the distribution used. The number of 1-states and 0-states in the *j*th block are then given by

$$\begin{aligned}n_1 &= x_1\\ n_0 &= m_o - n_1\\ (n_0 - n_1) &= m_o - 2n_1.\end{aligned} \tag{55}$$

The contribution that this block has to the overall walk is given by

$$\begin{aligned}\Delta L_j(m_o) &= (n_0 - n_1) - m_o\Delta f\\ &= \left(m_o - 2x_j\right) - m_o\Delta f.\end{aligned} \tag{56}$$

The net walk is then given by counting up the m_o blocks

$$L(n) = \sum_{j=1}^{N/m_o} \Delta L_j\left(x_j\right) \tag{57}$$

which gives the value of $L(n)$ at every unit that is a multiple of m_o. Again, the x_j are calculated at random but distributed according to either $P_G(x)$ or $P_M(x)$. We note that the basic parameters for the two distributions are as follows

$$\begin{aligned}&\text{Gaussian: } x_0 = m_o f_1, \quad \sigma_G = \sqrt{m_o f_o f_1}\\ &\text{Mandelbrot: } x_o = m_o f_1, \quad \sigma_M = A m_o^H.\end{aligned} \tag{58}$$

As pointed out by Mandelbrot [8,9], every fractal has an associated fractal dimension, *D*. For our random walk model *D* is related to γ, the persistence exponent is as follows

$$D = 2 - H = 3/2 - \gamma. \tag{59}$$

For our two distributions this gives

$$\begin{aligned}&\text{Gaussian: } (\gamma = 0) \quad D = 1.50\\ &\text{Mandelbrot: } (\gamma = 0.287, \text{ BA}) \quad D = 1.21.\end{aligned} \tag{60}$$

Sample Gaussian ($D = 1.50$) and Mandelbrot ($D = 1.21$) fractal walks based on the above parameters are shown in Fig. 10. The parameters used are those from BA. The curves show the characteristic fractal nature of random walks. We note that the fractal walk based on the Mandelbrot distribution is broader than the walk based on the Gaussian distribution, indicating that for the Mandelbrot distribution the walk increments tend to keep going in the same direction longer than for the Gaussian walk, indicating that the Mandelbrot walk has greater persistence. The total length of the sequences used is 5, 200, 000 (approximately the length of both BA and EC. We note that it is an important characteristic of random walks that the value of $L(n)$ will often be on a given side of the $L(n) = 0$ line for thousands and even millions of steps.

The jagged curves in Fig. 11 show the fractal walks based on the actual base sequences for BA and EC. For these walks the values of the x_j in Eq. (57) are taken from the actual sequence. We note the difference in the vertical scales of the curves given in Fig. 10 and those shown in Fig. 11 This indicates that the walks based on the base sequences have a structure that is on a scale much larger than that for the fractal walks shown in Fig. 10 and indicates a level of order that encompasses the entire sequence.

We can determine this global structure by fitting the jagged curves in Fig. 11 to a high order polynomial in *n* (here we have used the eleventh order). The results of this calculation are shown by the smooth curves in Fig. 11 which give the overall structure of $L(n)$ as a function of *n*. We then subtract the global curve from the overall curve as follows

$$DL(n) = L_{\text{overall}}(n) - L_{\text{global}}(n) \tag{61}$$

to obtain $DL(n)$, the local variation of $L(n)$. This is shown in Fig. 12 which now has the character of the $D = 1.21$ (Mandelbrot) walk shown in Fig. 10.

7. Model sequence with three levels of order

In the previous sections we have examined the occurrence statistics of A–T and C–G base pairs in DNA, using the genomes of BA and EC as examples. In Section 2 we focused on local statistics in the range of 10^0–10^1 bp. In this range we found that with respect to nearest-neighbor probabilities and block distribution functions the statistics was that of essentially random units. In Section 3 we

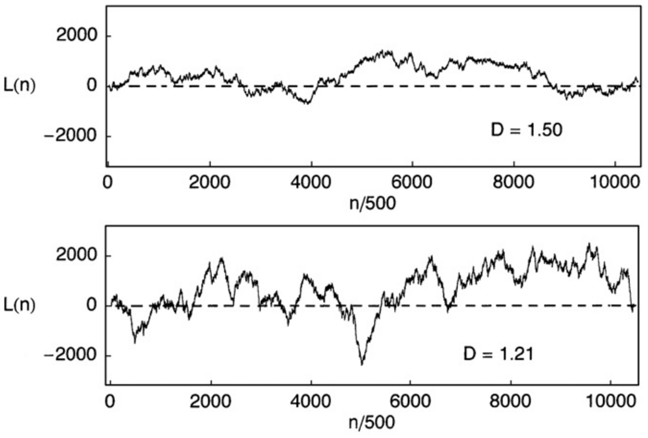

Fig. 10. Sample random walks given by the function $L(n)$ defined in Eq. (53) for a Gaussian random walk ($D = 1.50$) and a Mandelbrot random walk ($D = 1.21$). The walks are constructed using the parameters for BA. Every 500th point is shown.

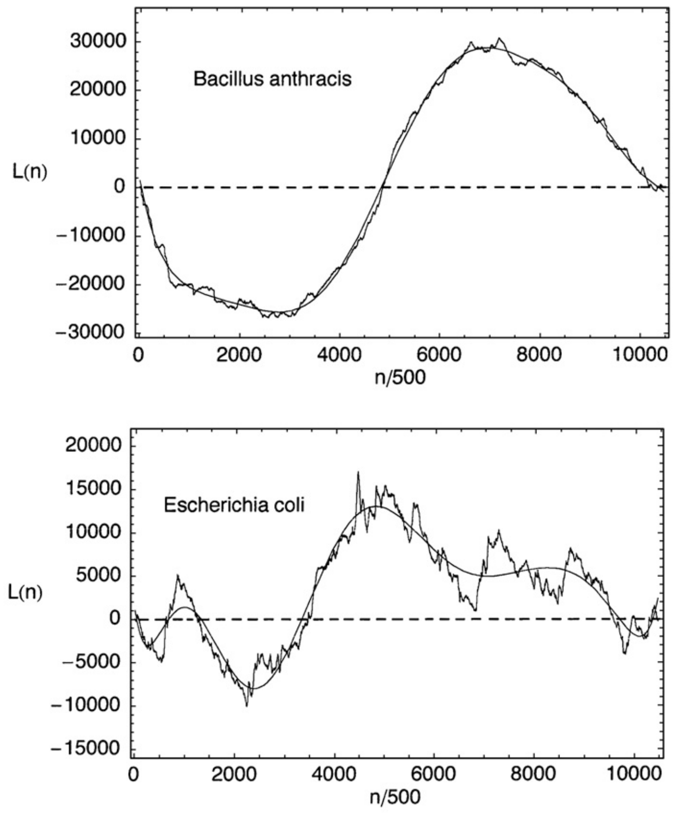

Fig. 11. Comparison of the background functions (smooth curves) with the total walks (jagged curves) of BA and EC.

constructed the function $\zeta(m)$ that measures the width (standard deviation) of the probability of finding m-blocks with a given number of 1-states and reflects order in the range of 10^2–10^4 bp. Here we find very large deviations from random behavior exemplified by the power law form of $\zeta(m)$. We found that we could understand this order in terms of the Mandelbrot distribution that incorporates persistence (the tendency of sequences to be followed by sequences of similar net composition). Finally, in the previous section we saw that there was a global (10^5–10^7 bp) structure to the sequence walks that could not be derived from the intermediate statistics as a consequence of persistence.

In this final section we explore the construction of a synthetic specific sequence that has all three levels of order found in DNA. Our purpose is to see what the requirements are to get all three levels of order in one sequence.

We begin by taking a chain of five million base pairs which is the approximate chain length for BA as given in Eq. (1). We then divide the chain up into large boxes of size $m_{\rm o}$ where we take $m_0 = 10{,}000$ which is close to the upper limit of m in the graphs of $\zeta(m)$ given in Fig. 2. We then assign a number, N_1, to each box giving the number of 1-states in the given box. The values of N_1 are picked at random according to a particular distribution function that we take as a modified version of the Mandelbrot distribution given in Eq. (32)

$$P_{\rm S}(x) = \frac{1}{\sqrt{2\pi}\,\sigma_{\rm S}} \exp\left[-(x-x_{\rm o})^2 / 2\sigma_{\rm S}^2\right] \tag{62}$$

where $\sigma_{\rm S}$ is a slight modification of the parameter given in Eq. (30), namely

$$\sigma_{\rm S} = b\sqrt{m}\,\zeta(m) = bA\,m^H \tag{63}$$

where b is an adjustable parameter. The parameter H remains the same as given in Eq. (31).

We have now generated a sequence of N_1 values, one for each $m_{\rm o}$-block, with the N_1 values distributed randomly but with frequencies given by the distribution function given in Eq. (62). Thus for blocks with $m_0 = 10{,}000$ we can generate any value for the function $\zeta(m = m_0)$ by varying the parameter b in Eq. (63). This then allows us to control the value of $\zeta(m)$ for large m.

The next step is to generate specific sequences of 0s and 1s in each m_0-block, consistent with the value of N_1 for that particular block. We will generate the sequences using a random distribution and thus the overall sequence will be basically random. A given block will have the following parameters:

$$\begin{aligned} \mathrm{m_0} &= \text{length of box} \\ \mathrm{N_1} &= \text{number of } 1'\text{s} \\ \mathrm{N_0} &= \text{number of } 0'\text{s} = \mathrm{m_0} - \mathrm{N_1} \end{aligned} \tag{64}$$

To incorporate the constraints of Eq. (64) we use the following procedure. We start with a vector of the successive integers from 1 to m_0 where the numbers indicate the locus of a site in the m_0-block. As an example, we take the following parameters: $m_0 = 6$, $N_1 = 4$ and $N_0 = 2$ giving

$$\boldsymbol{V}_{\rm o} = (1,2,3,4,5,6). \tag{65}$$

We then pick an integer from 1 to 6 at random and remove that element from the vector, for example, the integer 3

$$\boldsymbol{V}_1 = (1,2,4,5,6). \tag{66}$$

We repeat this procedure by picking an element from 1 to 5 at random and removing it (for example, the integer 4; note that we remove the 4th element in the vector and not the number 4)

$$\boldsymbol{V}_2 = (1,2,4,6). \tag{67}$$

Next we take a vector containing m_0 0-states as elements

$$\boldsymbol{U}_0 = (0,0,0,0,0,0). \tag{68}$$

Finally we set all of the vector elements in $\boldsymbol{U}_0$ corresponding to the elements in $\boldsymbol{V}_2$ equal to 1 in $\boldsymbol{U}_0$ giving

$$\boldsymbol{U} = (1,1,0,1,0,1) \tag{69}$$

which is a random sequence that has four 1s and two 0s. This method is very fast and efficient.

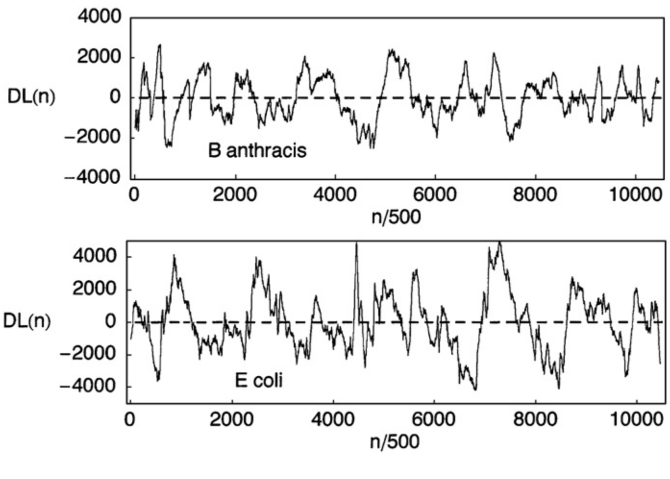

Fig. 12. The difference walks obtained by taking the difference between the background functions and the total walks given in Fig. 11 for BA and EC. Every 500th point is shown.

We thus have an algorithm that, first, constructs a specific sequence in which we have built in long range correlations according to a distribution of choice (such as that in Eq. (62)) for very large blocks ($m \approx 10{,}000$), thus giving order on the intermediate scale, and, second, fills out the sequence with a random selection of 0s and 1s as illustrated in Eqs. (65)–(69). We generate the sequence using random choice, but once the choice is made the sequence is specific. We now want to see what this sequence looks like with respect to short, intermediate and global order.

We begin by examining the short-range order. To this end we repeat the calculation used for BA and EC as illustrated in Fig. 1. In that case we counted the number of 1-states in blocks with $m=20$. We then compared the resulting distribution function with the binomial distribution of Eq. (14). The results of this calculation for our model sequence are shown in Fig. 13 where the solid dots are the results obtained by direct counting and the solid lines connect the corresponding points given by the binomial distribution. One sees that there is almost exact agreement between the empirical count and the random distribution. Thus on a local scale the model sequence behaves as a purely random system. This is not surprising since the local structure was generated randomly. However, the sequence is random subject to the constraint that the total numbers of 1-states in each large m_o-block are generated by a non-random distribution function.

Next we examine the function $\zeta(m)$ for the model sequence and compare it with the power law function for BA. The results of this calculation are shown in Fig. 14 where the solid dots are the results obtained by explicit counting for the model sequence and the upper dashed line is the power law form for BA. The model sequence was calculated using the distribution of Eq. (62) with $b=1.2$, this value of b being found to give the best overall fit to $\zeta(m)$ for BA. The results given in Fig. 14 indicate that if one adjusts the value of ζ for large m (here, $m=10{,}000$) using a single block size, then the overall pattern of $\zeta(m)$ follows. We note that the bump in the $\zeta(m)$ data at $m=10{,}000$ is a result of using only one box size. By using more box sizes it could be eliminated. Here we have chosen to pick the simplest approach using only one box size.

Finally we examine the global behavior of the model sequence using the walks as described in the previous section. The results are shown in the upper graph in Fig. 15 where the jagged curve gives the net function $L(n)$ for the model sequence. As with BA and EC there is a long-range structure to this function. We extract this behavior by fitting $L(n)$ to a polynomial in powers of n (11th order). The global function found in this manner is shown by the smoothly varying curve in Fig. 15. Next we tabulate the difference between the total $L(n)$ walk function and the global curve as indicated in Eq. (61). This result is shown in the lower graph in Fig. 15 and is seen to be very similar to the Mandelbrot walk curve given in Fig. 10 ($D=1.21$).

These calculations have shown that the model sequence has all of the order properties of real DNA: random locally, persistence power law in the intermediate range and a smooth variation at the global level. All of these features are the result of generating the number of 1-states in $m=10{,}000$ blocks using the distribution function of Eq. (62) and then filling in with a random sequence that obeys the overall constraints. Thus given the intermediate tendency for persistence of composition, all of the other features follow naturally. Of course in real DNA the intermediate structure is determined by the net composition of the genes.

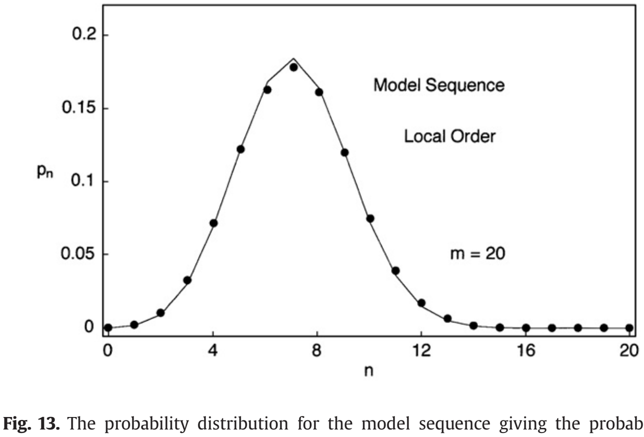

Fig. 13. The probability distribution for the model sequence giving the probability that a block of 20 bp will have [n] 1-states. The solid dots give the empirical distribution while the solid lines connect the points given by the binomial distribution of Eq. (14).

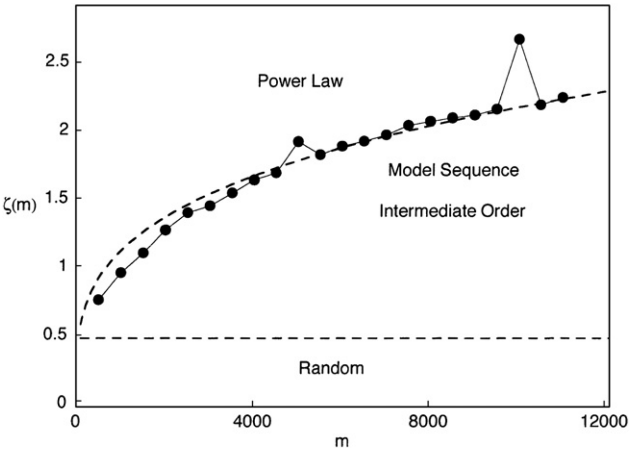

Fig. 14. The solid dots give the values of the $\zeta(m)$ function for the model sequence. The upper dashed curve gives the power law for BA.

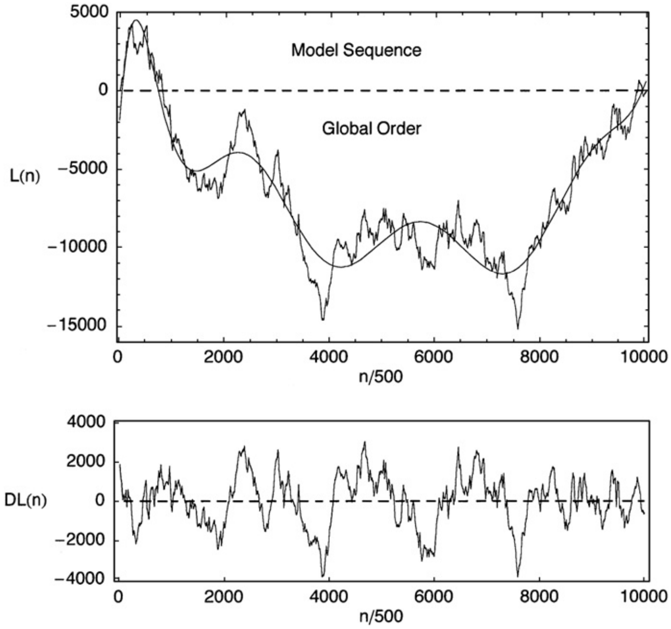

Fig. 15. The walk function $L(n)$ for the model sequence. The upper graph shows the net $L(n)$ function (jagged curve) while the smooth curve is the background global variation. The lower graph gives the difference between the two curves shown in the upper graph.

References

[1] D. Poland, The persistence exponent of DNA, Biophys. Chemist. 110 (2004) 59–72.

[2] D. Poland, The phylogeny of persistence in DNA, Biophys. Chemist. 112 (2004) 233–244.

[3] D. Poland, Universal scaling of the C–G distribution of genes, Biophys. Chemist. 117 (2005) 87–95.

[4] D. Poland, DNA probability profiles: examples from the *Treponema pallidum* genome, Biophys. Chemist. 104 (2003) 279–289.

[5] T.D. Read, et al., The genome sequence of *Bacillus anthracis Ames* and comparison to closely related bacteria, Nature 423 (2003) 81–86.

[6] R.A. Welch, et al., Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*, Proc. Natl. Acad. Sci. U.S.A. (2002) 17020–17024.

[7] The Institute for Genomic Research (TIGR) is now merged with the J. Craig Venter Institute (JCVI). Either of the web addresses www.tigr.org or www.jcvi.org leads to the same web page.

[8] B.B. Mandelbrot, The Fractal Geometry of Nature, W.H. Freeman and Company, New York, 1982.

[9] J. Feder, Fractals, Plenum Press, New York, 1989.

[10] L.R. Mead, N. Papanicolaou, Maximum entropy in the problem of moments, J. Math. Phys. 25 (1984) 2404–2417.

[11] D. Poland, Distribution functions from moments and the maximum entropy method, Methods Enzymol. 383 (2004) 427–465.

[12] D. Poland, Maximum-entropy calculation of energy distributions, J. Chem. Phys. 112 (2000) 6554–6562.

[13] D. Poland, Ligand binding distributions in nucleic acids, Biopolymers 58 (2001) 477–490.

[14] D. Poland, Maximum-entropy calculation of free energy distributions in tRNAs, Biophys. Chemist. 101-102C (2003) 485–495.