ELSEVIER

Biophysical Chemistry 112 (2004) 233–244

Biophysical Chemistry

www.elsevier.com/locate/bpc

The phylogeny of persistence in DNA

Douglas Poland*

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

Received 29 February 2004; accepted 1 July 2004
Available online 25 September 2004

Abstract

We continue our study, Poland [Biophysical Chemistry 110 (2004) 59–72], of the distribution of C or G (C–G for short) in the DNA of select organisms, in particular, the tendency for C–G to cluster on all scales with respect to the number of bases considered. We previously found that if we counted the number of C–G bases in consecutive, nonoverlapping boxes containing a total of m bases, then the width of the distribution function describing how many C–G bases are in a box increases with respect to m dramatically relative to the width expected for a random distribution. The relative width of the C–G composition distribution function was found to vary accurately as a power law with respect to m, the size of the box, over a very wide range of m values. We express the power law in terms of a characteristic exponent γ, that is, the relative widths of the distributions vary as m^{γ}. The enhanced relative width of the distribution functions is a direct consequence of the tendency for boxes of similar composition to follow one another. This tendency represents persistence in composition from box to box and hence we refer to γ as the persistence exponent. The occurrence of a power law means that the tendency for C–G to cluster is present on all scales of sequence length (box size) up to the total length of the chromosome which for bacteria is the entire genome. The persistence exponent γ that characterizes the power law is thus an important parameter describing the distribution of C–G on all scales from individual base pairs up to the total length of the DNA sample considered. In the present paper, we determine the characteristic exponent γ and the associated fractal dimension of DNA samples for a selection of species representing all of the major types of organism, that is, we explore the phylogeny of the exponent γ. Here we treat six prokaryotes and six eukaryotes which, together with the species we have previously treated, brings the total number of species we have examined to 15. We find the power law form for the C–G distribution for all of the species treated and hence this behavior seems to be ubiquitous. The values of the characteristic exponent γ that we find tend to cluster around the value $\gamma = 0.20$ with no obvious pattern with respect to phylogeny. The extreme values that we obtain are $\gamma = 0.057$ (yeast) and $\gamma = 0.386$ (human). We conclude by showing that the persistence of C–G clustering on the scale of the length of a chromosome is dramatically illustrated by interpreting the C–G distribution as a random walk.

Keywords: Persistence exponent; Fractal dimension; Power law; Long-range correlation; DNA

1. Introduction

In two recent [1,2] papers, we have explored long-range correlations in the base composition of DNA. In the first of these papers [1], we examined the free energy for blocks of double helix based on the known base sequence. In order to understand the nature of the correlations we found, we then treated [2] a simpler property of DNA, namely, the net distribution of A or T and C or G units in blocks of bases of a given size. In this case, we start with the given DNA base sequence and then assign a unit of zero if it is A or T (A–T) and one is it is C or G (C–G). For simplicity, we will use the designation A–T to indicate that a unit is A or T and the designation C–G to indicate that a unit is C or G. We then examine consecutive nonoverlapping blocks of m units in the chain and count how many ones there are in each block. Finally, we count how many blocks there are with no C–G units, how many with one C–G unit and so on up to m C–G units. The resulting set of numbers gives one a distribution function for the occurrence of C–G units in the molecule. If one compares the empirical C–G distribution functions

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

0301-4622/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.07.025

obtained in the manner just described with the distribution functions expected for a random distribution of C–G units then one finds that the empirical distributions are very much broader than the corresponding random distributions and that the width of the distributions relative to the width of the random distributions increases with m, the size of the block. This increase in the relative width of the distributions with block size follows a very regular pattern in that it is given very accurately over an extremely wide range of block sizes by a power law expressed in terms of a characteristic exponent.

In the present paper, we explore how general this power law behavior is over a wide range of species, that is, we explore the phylogeny of the power law phenomena. We find that the broadening of the C–G distributions is a result of the tendency for blocks with similar C–G composition to follow one another. If one interprets the A–T and C–G occurrence in the DNA chain as a random walk (A–T gives a step in the negative direction, C–G gives a step in the positive direction), then the distribution broadening is manifested by walks that tend to keep going, on average, in the same direction to a much greater extent than expected for the random occurrence of bases. Thus the power law behavior is associated with a persistence of occurrence of C–G units and the characteristic exponent for the power law can be considered a persistence exponent that measures the strength of this phenomenon. The characteristic exponent for the power law can also be used to calculate a fractal dimension for the random walk with persistence.

We begin our survey of the phylogeny of the power law and the persistence exponent in DNA by referring to the schematic tree of life given in Fig. 1 which was constructed following the treatment in the book by Tudge [3]. We have already examined the power law behavior in three species: *Rickettsia prowazekii* [1], *Thermoplasma volcanium* [2], and *Homo sapiens* [2] representing, respectively, two prokaryotes (a bacteria and an archean) and one eukaryote (human). In this paper, we will treat 12 additional species that cover the major groups indicated in Fig. 1. Six of the additional species will be prokaryotes, all bacteria. Bacteria are attractive to treat since the whole genome is contained in a single chromosome and thus the statistics reflect the behavior of the entire genome. The remaining species treated will include five eukaryotes (a protist, a plant, a member of the fungi, a nematode and the fruit fly) and one virus. For the eukaryotes, we will treat part of the complete genome, usually a single chromosome, either whole or in part. In our selection of species to be examined, we have included those that have been extensively used as model organisms in biology. The protozoan we have chosen is the malaria parasite, the fungi is baker's yeast and the plant is thale crest which is a small flowering plant of the mustard family much studied in plant biology. The nematode worm is used as a model organism for studies of developmental biology while the fruit fly is a classic species used in genetic studies.

Below we list all of the species that we have examined for power law behavior. We give the proper biological name, our abbreviation for the name, the chromosome used

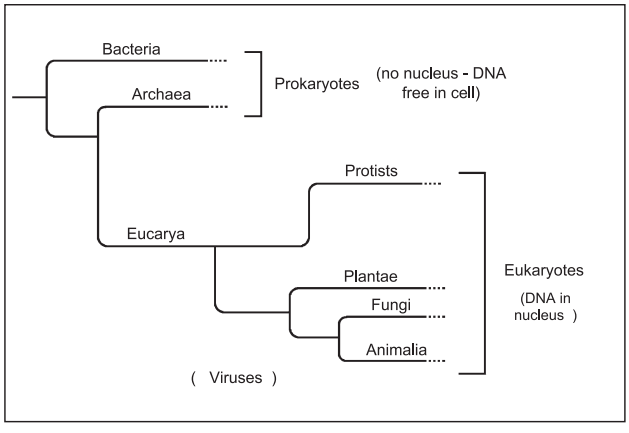

Fig. 1. Schematic diagram of the branching relationships between the major categories of life (constructed after Trudge [3]).

Species list

Bacteria
(1) *Mycoplasma pneumoniae* (Mp) (N=816,394) [4]
(2) *Treponema pallidum* (Tp) (N=1,138,012) [5]
(3) *Helicobacter pylori* (Hp) (N=1,643,831) [6]
(4) *Haemophilus influenzae* (Hi) (N=1,830,140) [7]
(5) *Streptococcus pneumoniae* (Sp) (N=2,160,837) [8]
(6) *Staphylococcus aureus* (Sa) (N=2,820,462) [9]

Protists
(7) *Plasmodium falciparum* (Pf) malaria parasite; chromosome #12 (N=2,271,916) [10]

Plantae
(8) *Arabidopsis thaliana* (At) thale cress; chromosome #4 (N=18,585,042) [11]

Fungi
(9) *Saccharomyces cerevisiae* (Sc) baker's yeast; chromosome #4 (N=1,531,912) [12]

Animalia
(10) *Caenorhabditis elegans* (Ce) nematode; chromosome #3 (N=13,783,268) [13]
(11) *Drosophila melanogaster* (Dm) fluit fly; chromosome 2 L (N=22,217,931) [14]

Virus
(12) *Bacteriophage KVP40* (Kv) T4-related bacteriophage (N=244,835) [15]

Previously Treated:
Bacteria
(13) *Rickettsia prowazekii* (Rp) (N=1,111,523) [16]
Archaea
(14) *Thermoplasma volcanium* (Tv) (N=1,584,804) [17]

Animalia
(15) *Homo sapiens* (Hs) human; chromosome #3 (N=10,000,000) [18]

(for eukaryotes) and the size, N, of the sequence treated; the reference to the original publication is also given.

The genomes were obtained from The Institute for Genomic Research [19] for species 1–6 and 13 and from the National Center for BioTechnology Informatics [20] for species 7–12 and 14.

2. Composition distribution functions

As indicated in the Introduction, we study the distribution of DNA base composition in consecutive, nonoverlapping boxes containing m bases where we examine the distributions for a wide range of m values. In a given block of m bases, we replace the designators of the actual sequence (A, T, C or G) by a zero if a particular base is A or T and by a one if the base is C or G and we then count how many ones there are in a given block. An example of this process is given below for m=6:

$$\begin{aligned}&\ldots(\mathrm{AACTGC})(\mathrm{GGACGA})(\mathrm{GTACTT})\ldots\\ &\quad\rightarrow\ldots(0\ 0\ 1\ 0\ 1\ 1)(1\ 1\ 0\ 1\ 1\ 0)(1\ 0\ 0\ 1\ 0\ 0)\ldots\\ &\quad\rightarrow\ldots(3)(4)(2)\ldots\end{aligned} \tag{1}$$

One then simply tabulates how many boxes contain a given number, n, of C or G units where n can vary from zero to m. The resulting distribution is roughly a bell-shaped curve with a maximum at approximately the value determined by the average C–G composition of the DNA.

As an example of the construction of the composition distribution function, we will use a piece of DNA from At (thale cress, species number 8 in our list). We will take statistics from a sequence of bases from chromosome number 4 which contains 18,585,042 bp. The specific sequence we will use is a 30,000 base sample on chromosome #4 beginning at base number 9,050,000 which contains no undetermined bases. We will look at consecutive nonoverlapping blocks of m=100 and thus in this sequence we have 300 such blocks.

Before we give the statistics of C–G composition for these 300 blocks, we review some basic notions about distributions. The most important overall statistic for a given sequence is the fraction of A or T units and the fraction of C or G units which we designate as follows

$$\begin{aligned}f_{at} &= \text{ fraction A or T}\\ f_{cg} &= \text{ fraction C or G}\end{aligned} \tag{2}$$

where these fractions sum to one

$$f_{at}+f_{cg}=1 \tag{3}$$

If the distribution of A–T and C–G units is random then the probability of having n C–G units in a box of m units is given by the binomial distribution

$$P(n)=\left(\frac{m!}{n!(m-n!)}\right)f_{at}^{m-n}f_{cg}^{n} \tag{4}$$

For any distribution, we can obtain a measure of the width of that distribution by constructing the root-mean-square variation

$$\sigma_m=\sqrt{\mu_2(m)-\mu_1(m)^2} \tag{5}$$

where μ_1 and μ_2 are in general the first and second moments of the distribution. For the random distribution given in Eq. (4), one has

$$\sigma_m(\text{random distribution})=\sqrt{m}\sqrt{f_{cg}-f_{cg}^2} \tag{6}$$

For our sample of At, we have we have the following values for the base fractions given in Eq. (2)

$$\begin{aligned}f_{cg}&=0.3566\\ f_{at}&=0.6434\end{aligned} \tag{7}$$

Using these values in Eq. (6) together with the block size m=100, we obtain the following distribution width for the random distribution

$$\sigma_m(\text{random distribution})=\sqrt{m}(0.479)=4.79 \tag{8}$$

We let n_{b} designate the position number of a given block in the sequence where n_{b} can vary from 1 to 300 for our example. We define $B(n_{\mathrm{b}})$ as the number of C or G bases in a given block. The upper graph in Fig. 2 (marked At sequence) gives the values of $B(n_{\mathrm{b}})$ as a function of n_{b} for our chosen base sequence. The lower curve gives an example of the same function for a random distribution of bases with the overall base frequencies given in Eq. (7). Based on the net averages given in Eq. (7), the average number of C or G bases in a block of m=100 would be

$$\langle B\rangle=100f_{cg}=35.66 \tag{9}$$

If the distribution were random then most of the distribution would be contained between the following two values (using the numbers given in Eqs. (8) and (9))

$$\langle B\rangle-\sigma_m/2=33.27 \text{ and } \langle B\rangle+\sigma_m/2=38.05 \tag{10}$$

These values are indicated in both graphs in Fig. 2 by the two dashed lines. In the upper graph, representing the statistics obtained from the actual At sequence, one sees that the values of B fluctuate greatly outside the most probable bounds given by Eq. (10). In contrast, the variation of B for a random sequence shown in the lower graph is mainly contained between the most probable bounds. One clearly sees from this example that the variation in the C–G distribution is much broader in an actual DNA sequence than for a corresponding random sequence having the same overall C–G fraction.

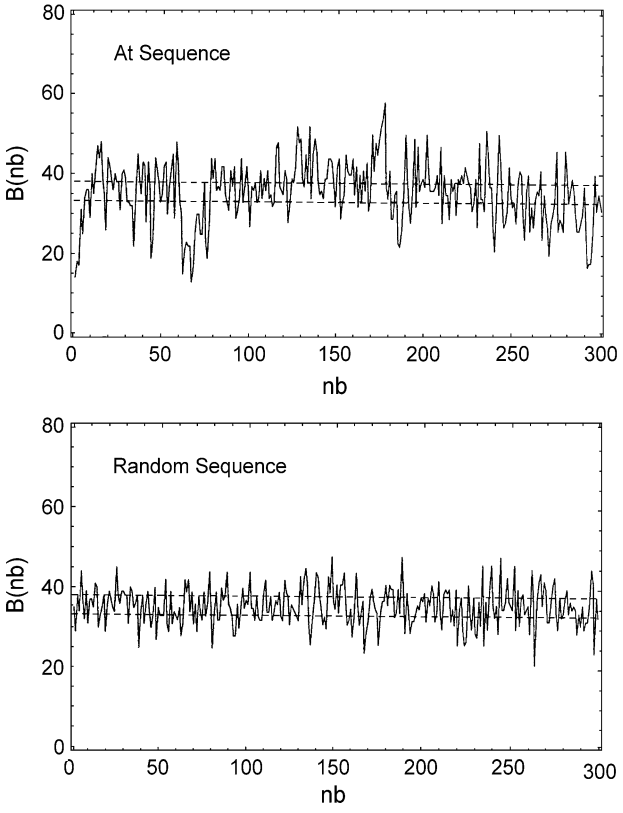

Fig. 2. The upper graph shows the number of C–G units in consecutive, nonoverlapping blocks of m=100 bp as a function of location, n_b, in the DNA sequence. The particular sequence illustrated is a 30,000 bp piece of DNA beginning at base 9,050,000 in chromosome 4 from the plant At. The lower graph shows the same quantity for a random sequence of bases having the same overall base composition as the specific sequence treated in the upper graph. The dashed lines in both graphs give the bounds given in Eq. (10) for a random sequence. One sees that the values of B for the At sequence vary significantly outside the most probable bounds for a corresponding random sequence.

We obtain the distribution of C–G composition by using data like the $B(n_b)$ values shown in the upper graph in Fig. 2 and counting how many of the boxes (each with m=100) contain n C–G bases. To get better statistics, we enlarge the sample sequence treated to 9,500,000 bp, still on chromosome #4. The results of this tabulation are shown in Fig. 3 where the solid dots give the empirical data indicating what fraction of the boxes contain n C–G units. The smooth curve is a plot of Eq. (4) giving the distribution for a random sequence with the same overall C–G composition as the actual sequence. One now sees clearly that the empirical C–G distribution is significantly broader than that the corresponding random distribution. The core of our story is that this difference in width between the empirical distribution and the corresponding random distribution increases dramatically as m is increased and that the ratio of the widths of the two functions obeys a power law

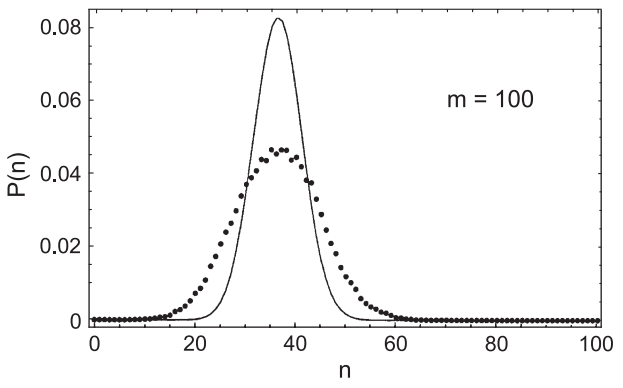

Fig. 3. The C–G distribution function for blocks with m=100 for the plant At based on a 9,500,000 bp sequence from chromosome 4. The solid dots give the distribution based on the At sequence while the solid curve is the random distribution of Eq. (4) based on the same net composition as the actual sequence.

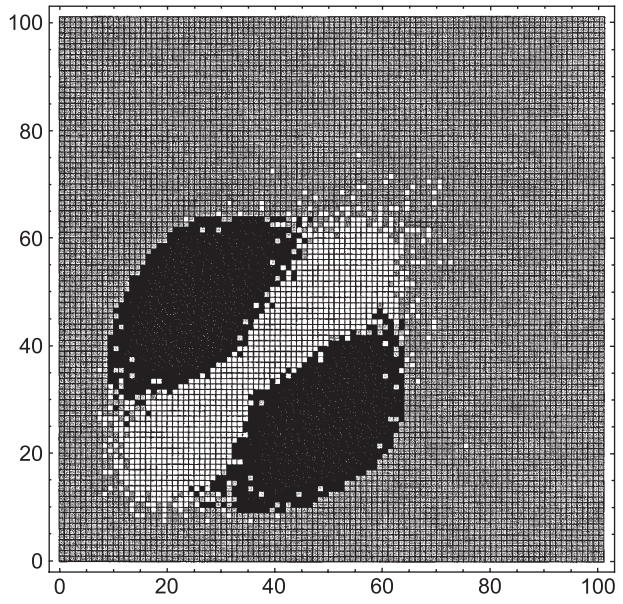

Fig. 4. The plus/minus map, based on Eq. (11), for the At sequence used in Fig. 3. The graph shows the statistics obtained from blocks with m=100. The axes index the number of C–G units in consecutive blocks. The white squares indicate that state i tends to be followed by state j more often than random occurrence, while the black squares indicate just the opposite.

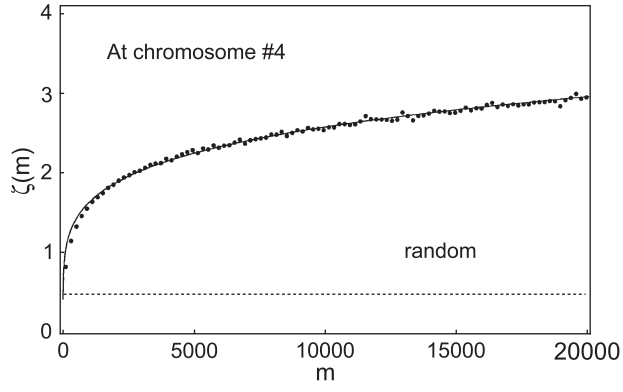

Fig. 5. A plot of the relative width function $\zeta(m)$ defined in Eq. (13) for the At sequence used in Fig. 3. The solid dots are the empirical points for values of m that are multiples of 100 for m=100 up to 20,000. The solid curve is a plot of the power law given in Eq. (14) using the parameters given in Eq. (15). The dashed curve gives $\zeta(m)$ for a random sequence.

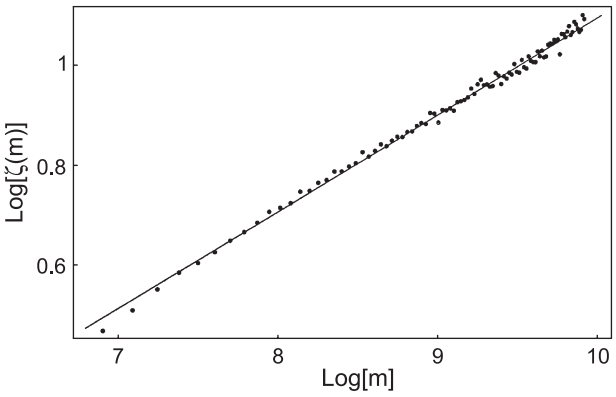

Fig. 6. A plot of the natural logarithm of $\zeta(m)$ as a function of the natural logarithm of m using the empirical data points for At given in Fig. 5. The straight line gives the best linear fit resulting in the parameters given in Eq. (15).

(expressed in terms of a characteristic exponent) with respect to the variable m.

Another way to demonstrate the nonrandom character of the block distribution functions is the construction of plus/minus maps which measure the tendency for blocks of a one size to follow blocks of another size. To construct this map, we simply collect statistics, f_{ij}, on how often blocks with i C–G units are followed by blocks with j C–G units for all i and j. We then compare these frequencies with the frequencies, $f_i f_j$, expected for a random distribution of block contents. The plus/minus map is then given by the difference in these two sets of frequencies

$$\Delta_{ij} = \mathrm{sign}\left(f_{ij} - f_i f_j\right) \tag{11}$$

where "sign" indicates that one takes the sign, plus or minus, of the quantity in brackets. If a given i–j combination occurs more often than random then this quantity will be positive and vice versa. Using the data given in the upper graph in Fig. 2, one obtains the plus/minus map shown in Fig. 4. The white and black squares indicate where Δ_{ij} is positive and negative, respectively. One sees that there is a strong tendency for positive correlations to occur along the axis $i=j$ which means that like tends to follow like with respect to block composition. Repeating this correlation from block to block leads to runs of blocks having similar composition (relative to a random distribution). We have previously shown [1] that this type of correlation tends to increase the "wings" of the distribution function, hence leading to the broadening shown in Fig. 3. This is the idea of persistence: like tends to follow like with respect to the net composition of blocks. And the unexpected result is that the tendency for this type of correlation persists to blocks of all sizes indicating that there is correlation operating over the entire sequence.

If one now repeats the same procedure for different values of m, say m=100, 200, and so on, one finds that the width of the block distribution function increases significantly, relative to the width expected for a random distribution, as a function of m. We define the width in terms of the moments of the distribution (which can be

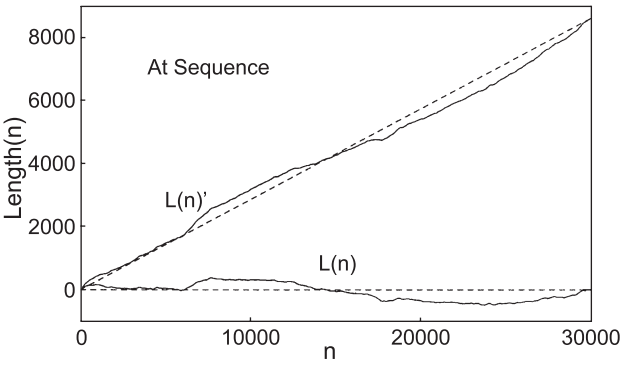

Fig. 7. The upper solid curve gives the random walk function $L(n)'$ defined in Eq. (17) for the At sequence used in Fig. 2. The upper dashed curve gives the locus of the average length of the walk as a function of the number of steps as given in Eq. (18). The lower solid curve gives the random walk function relative to the average length of the walk as defined in Eq. (20). The average of this function is zero as indicated by the lower dashed curve.

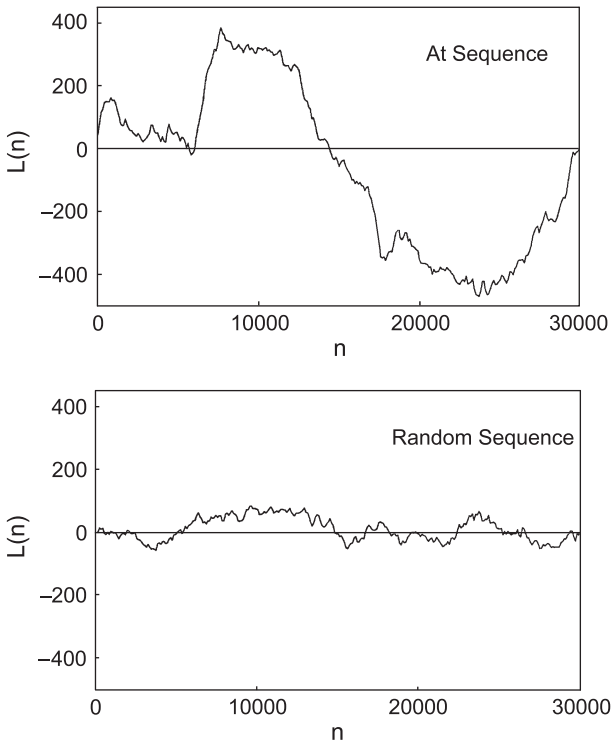

Fig. 8. The upper graph shows the function $L(n)$ for At as given in Fig. 7 shown on a closer scale. The lower graph shows $L(n)$ for a random sequence on the same scale.

determined empirically using, for example, the data (solid dots) in Fig. 3) as given in Eq. (5)

$$\mu_1(m) = \sum_{nk=0}^{m} nP(n) = \langle n \rangle$$

$$\mu_2(m) = \sum_{n=0}^{m} n^2 P(n) = \langle n^2 \rangle \tag{12}$$

For a random distribution, we expect the width of the distribution to have the square-root dependence on m given in Eq. (6). We then define a function that measures the width of the empirical distribution relative to that for the m dependence for a corresponding random distribution as follows

$$\zeta(m) = \frac{\sigma_m}{\sqrt{m}} \tag{13}$$

If the empirical distributions are indeed random, then this function should have a constant value independent of m. We find that in fact this function has a strong variation with m that can be accurately fit by a power law of the form

$$\zeta(m) = A\ m^{\gamma} \tag{14}$$

Fig. 5 shows the function $\zeta(m)$ obtained from our large At sample. The solid dots give the numerical values of $\zeta(m)$ for m values that are multiples of 100 for m=100 up to m=20,000. The test for power law behavior as shown in Eq. (14) is that $\ln[\zeta(m)]$ should be linear in $\ln[m]$. This plot is shown in Fig. 6 and one sees that a linear relation fits the data very well. The solid line in Fig. 6 is given using the best-fit parameters

$$A = 0.424 \text{ and } \gamma = 0.196 \tag{15}$$

The solid curve in Fig. 5 is a plot of Eq. (14) using these parameters. If the distributions were random, then $\zeta(m)$ would be a constant independent of m and the value of this constant is indicated by the dashed curve in Fig. 5. One sees that the broadening due to block correlation (persistence) is a very large effect: the actual distributions are four to six times broader than the corresponding random distributions for the range of m treated in Fig. 5.

Finally, we can gain some insight into the correlations in DNA sequences by interpreting the sequence as a random walk. To do this, we make the following assignments to each base in the sequence:

$$\alpha_i = -1 \text{ if C or G}$$

$$\alpha_i = +1 \text{ if A or T} \tag{16}$$

One can think of this as a random walk where α_i=+1 indicates a step to the right and α_i=−1 indicates a step to the

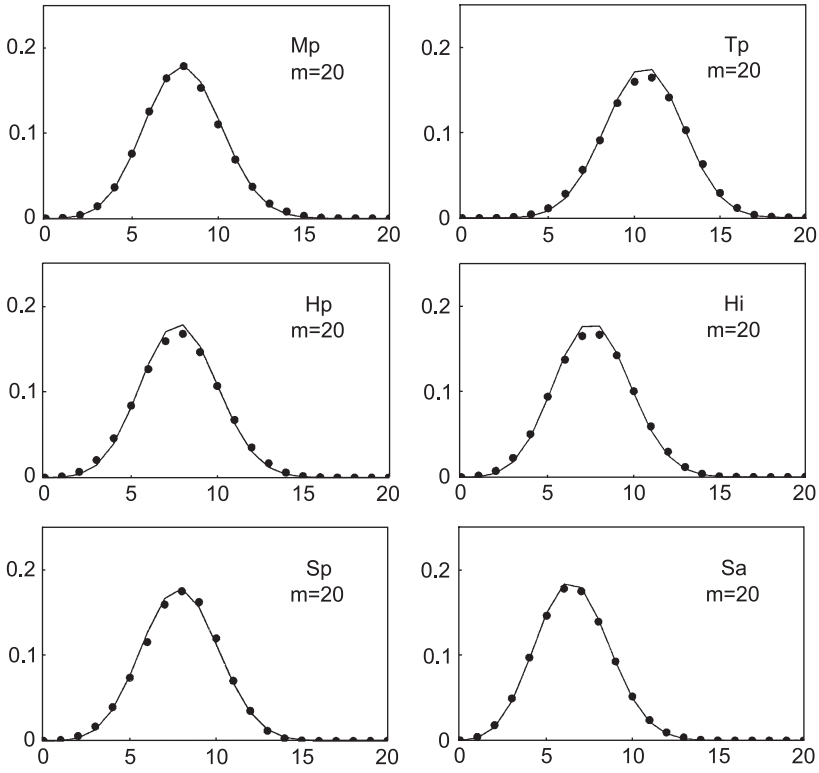

Fig. 9. The block distributions for m=20 for the set of six bacteria given in the Species list. The solid points give the fraction of m=20 blocks containing n C or G bases obtained empirically from the entire genomes for these species. The solid curve joins the points for a random sequence having the same overall composition as given by Eq. (4).

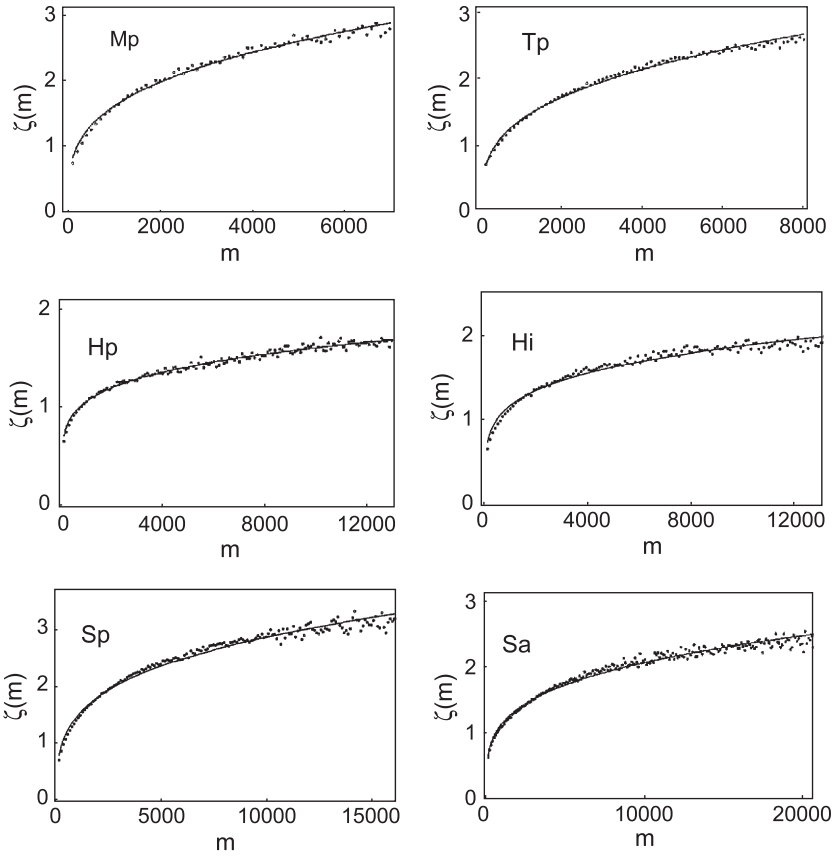

Fig. 10. A plot of the relative width function $\zeta(m)$ defined in Eq. (13) for the set of six bacteria treated in Fig. 9. The solid dots give the empirical points while the solid line is a plot of power law given in Eq. (14) using the parameters obtained from plots similar to Fig. 6 and listed in Table 1.

left. The distance of the walk from the origin after n steps (n bases) is then given by

$$L(n)' = \sum_{i=1}^{n} \alpha_i \tag{17}$$

On the average, this function will have the value

$$\langle L(n)' \rangle = n\Delta f \tag{18}$$

where

$$\Delta f = f_{at} - f_{cg} \tag{19}$$

It is convenient to define the length of the walk relative to the average walk given above. This is given by

$$L(n) = \sum_{i=1}^{n} \alpha_i - n\Delta f \tag{20}$$

The average value of this quantity is, by definition, zero.

Both of the quantities $L(n)'$ and $L(n)$ are plotted for our large At sequence in Fig. 7. The dashed curves give the respective average values as given by Eq. (18) for $L(n)'$ and zero for $L(n)$. One sees that the function $L(n)$ is more convenient to treat than $L(n)'$. In the upper graph in Fig. 8, we show the details of $L(n)$ for our At sequence on a finer scale. In the lower graph, we show $L(n)$ for a random sequence with the same overall base composition as our sample. Again the phenomenon of persistence is obvious: the walk for the At sequence has basically two very large swings, first persistence in one direction and then persistence in the other direction. Thus the graph of the walk for the actual At sequence shows structure and hence correlations on the scale of the entire length of the sequence.

Table 1
The persistence exponent γ and the fractal dimension D for assorted species

Species	f_{cg}	N (Mb)	γ	D
(1) Mp	0.399	0.82	0.297	1.20
(2) Tp	0.527	1.14	0.318	1.18
(3) Hp	0.390	1.64	0.178	1.32
(4) Hi	0.380	1.83	0.205	1.30
(5) Sp	0.396	2.16	0.282	1.22
(6) Sa	0.327	2.82	0.255	1.25
(7) Pf	0.193	2.27	0.234	1.27
(8) At	0.363	9.50	0.196	1.30
(9) Sc	0.379	1.53	0.057	1.44
(10) Ce	0.357	13.78	0.191	1.31
(11) Dm	0.421	21.00	0.262	1.24
(12) Kv	0.426	0.24	0.183	1.32
(13) Rp	0.289	1.11	0.075	1.43
(14) Tv	0.399	1.58	0.290	1.21
(15) Hs	0.406	10.00	0.386	1.11

3. Distribution functions for six bacteria

We treat here the distribution functions for the bacteria listed as the first six organisms in the Species list in the Introduction. For the case of these bacteria, the complete genome is a single circular chromosome and thus the distributions we treat here reflect the entire genomes for these organisms. We first show the C–G distribution functions for these organisms for a small value of m, namely m=20. From the behavior of the width function $\zeta(m)$ for the plant At illustrated in Fig. 5, we see that the empirical points (solid dots) approach the constant value of the width function for a random system (dashed line) at small values of m (say, m less than 50). Thus we expect that the empirical C–G distribution functions for m=20 for these six bacteria will be accurately fit by the distribution function for a random sequence as given by Eq. (4). The C–G distributions for m=20 for these organisms are shown in Fig. 9 where the solid dots give the empirical points and the solid lines are plots of the distribution for random sequences given by Eq. (4). One sees that the empirical distributions and the distributions for random sequences are essentially identical thus confirming our expectation from the behavior shown in Fig. 5 that the empirical and random distributions are very similar for small values of m.

What is also evident in Fig. 5 is that as the value of m increases the width of the empirical distributions becomes very much greater than that for the corresponding random distribution. In Fig. 3, the empirical (solid dots) and random (solid line) distributions for At are shown for m=100, illustrating the fact that for this value of m the empirical distribution is significantly wider than the random distribution.

In Fig. 10, we show the relative width Function $\zeta(m)$ for our set of six bacteria for a very wide range of m values. The solid dots give the empirical points where we give all of the points for m values that are multiples of 100. The solid lines give the power law fit of the data using the form of Eq. (14). The values of γ that give the best fit are shown in Table 1. If the power law behavior of Eq. (14) holds, then the following function should be a constant independent of m

$$R(m) = \zeta(m)/m^{\gamma} \tag{21}$$

The function $R(m)$ constructed using the values of γ given in Table 1 is plotted for our set of six bacteria in Fig. 11 and one sees that in all cases $R(m)$ is essentially constant over the range of m values shown giving outstanding evidence for the power law behavior of the width function.

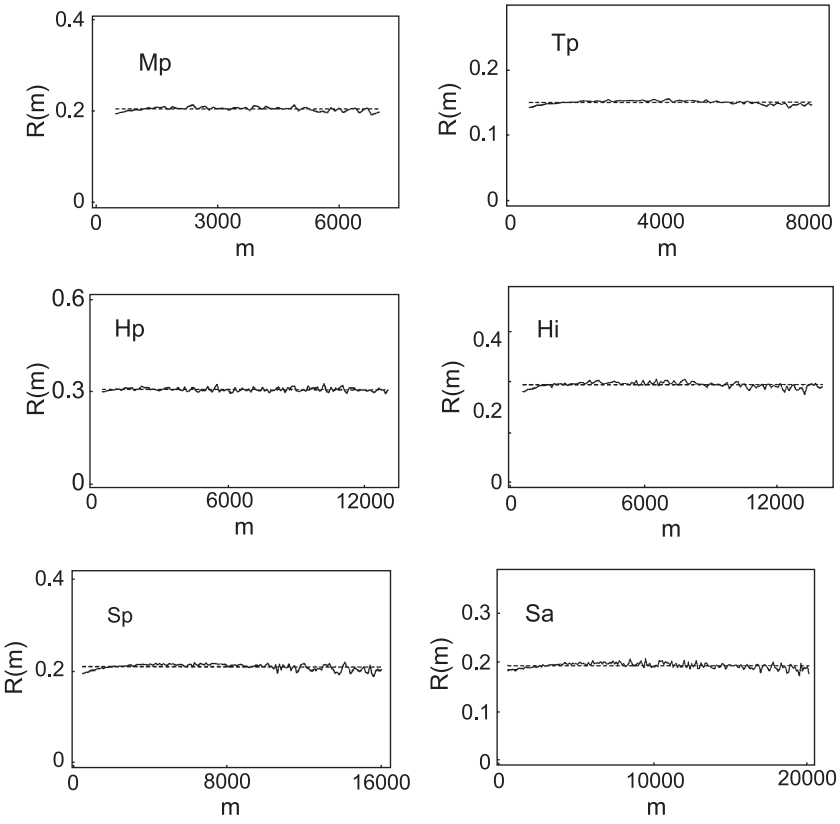

Fig. 11. The function $R(m)$ defined in Eq. (21) based on the data given in Fig. 10 for the set of six bacteria.

4. Distribution functions for five eukaryotes and a virus

We next examine the behavior of the relative width function $\zeta(m)$ for species 7 through 12 in our Species list. Fig. 12 shows the behavior of $\zeta(m)$ for Sc (yeast), At (plant), and Kv (virus). As before, the solid dots give the empirical points for m values that are multiples of 100 while the constant dashed lines give the behavior for the appropriate random distributions. In each case, the solid curve gives the form of $\zeta(m)$ given by the power law form of Eq. (14) using the values of the persistence exponents given in Table 1. Fig. 13 gives the same functions for Pf (protozoan), Ce (nematode) and Dm (fruit fly). We conclude that the width function $\zeta(m)$ is fit very well by the power-law form of Eq. (14) for the C–G distributions of the six organisms treated in Figs. 12 and 13.

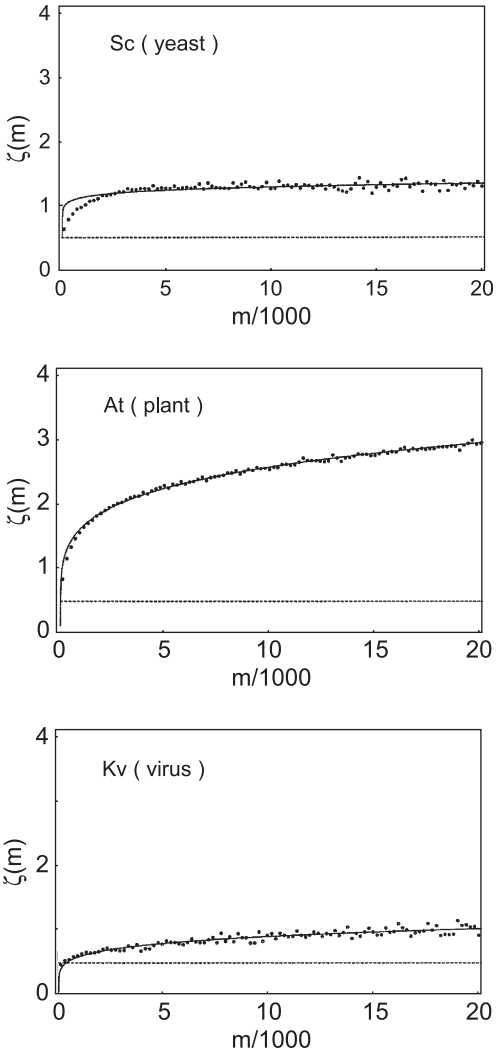

Fig.12. A plot of the relative width function $\zeta(m)$ defined in Eq. (13) for Sc (yeast), At (plant) and Kv (virus). In each case, the solid dots give the empirical points obtained from the sequence used while the solid curve is the power law given in Eq. (14) using the parameters listed in Table 1.

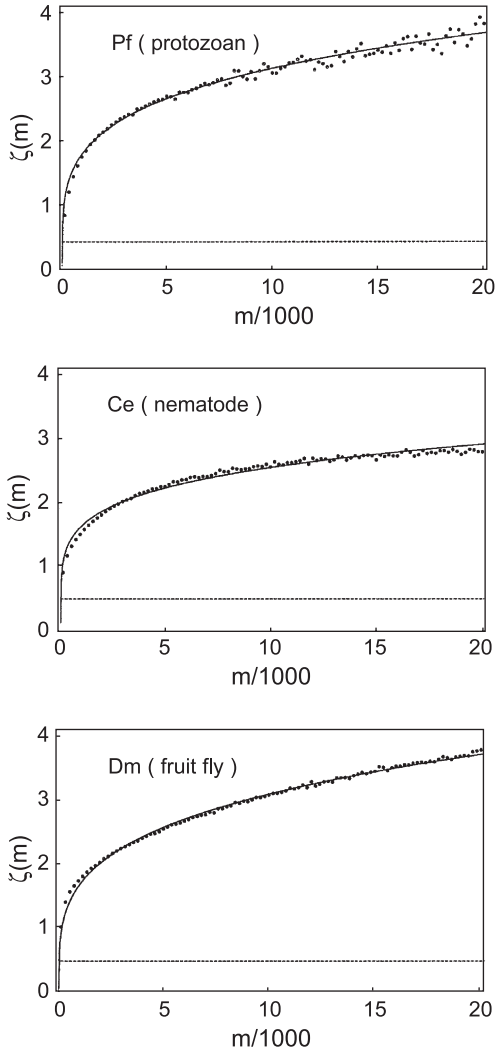

Fig. 13. A plot of the relative width function $\zeta(m)$ defined in Eq. (13) for Pf (protozoan), Ce (nematode) and Dm (fruit fly). In each case, the solid dots give the empirical points obtained from the sequence used while the solid curve is the power law given in Eq. (14) using the parameters listed in Table 1.

Fig. 14 gives the random-walk function $L(n)$ for the species treated in Fig. 12 (yeast, plant and virus). The data in Fig. 14 is plotted in terms of the relative function $L(n)/N$ as a function of n/N where N is the number of bases in the particular sample treated. Fig. 15 shows the same function for the species treated in Fig. 13 (protozoan, nematode and fruit fly). In all of the graphs in Figs. 14 and 15, the curve showing the large dramatic variations and swings is based on the actual base sequence for the species treated; the other curve with less variations is a random walk function based on a random sequence having the same overall AT/CG composition as the sequence treated. The striking feature about all of the graphs shown in Figs. 14 and 15 is that there is structure in all of these walk curves on the scale of the overall length of the piece of DNA treated.

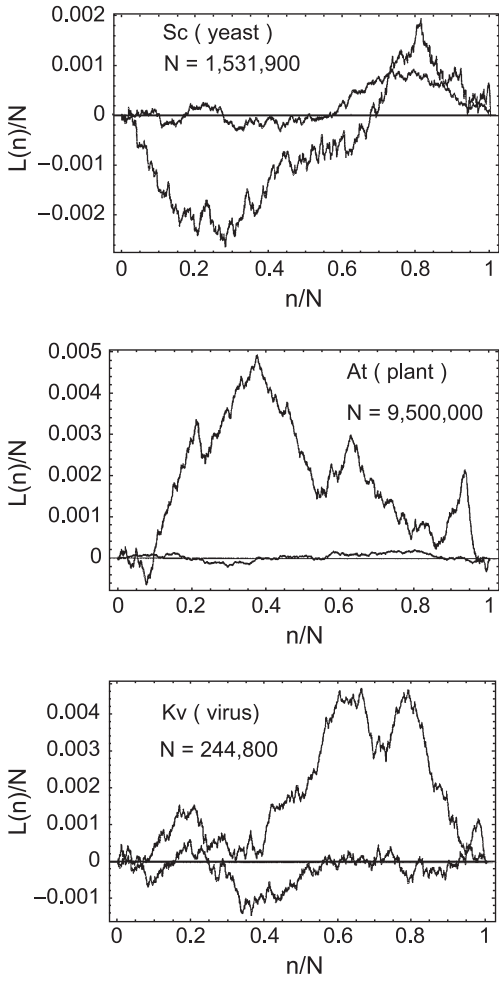

Fig. 14. A plot of the relative random walk length function $L(n)$ defined in Eq. (20) for Sc (yeast), At (plant) and Kv (virus). In each case, the curve with a smaller amplitude gives $L(n)$ for a random sequence for comparison.

To obtain a single number that characterizes the length of the random walk, we can calculate the average $L(n)$ as follows

$$\langle L\rangle = \left[\frac{1}{N}\sum_{n=1}^{N} L(n)^2\right]^{1/2} \tag{22}$$

We can construct this quantity for all of the walks shown in Figs. 14 and 15, both for the walks based on the actual sequence and the walks for the appropriate random sequences. In Fig. 16, we plot the averages obtained using Eq. (22) as the natural logarithm of $\langle L\rangle$ as a function of the natural logarithm of N (the length of the DNA sequence treated). This plot gives a very rough idea of the variation of $\langle L\rangle$ with N. If we fit the data given in Fig. 16 to straight lines, as shown by the dashed curves, then we obtain the following power law for the variation of $\langle L\rangle$ with N

$$\langle L\rangle = c\, N^{\alpha} \tag{23}$$

where c is a constant. From the data given in Fig. 16, we obtain the following rough values of

$$\alpha = 0.50 \text{ (random sequences)}$$

$$\alpha = 1.08 \text{ (actual sequences)} \tag{24}$$

Since the quantity $\langle L\rangle$ is a measure of the overall length of the random walk, the exponents given in Eq. (24) indicate that $\langle L\rangle$ is much larger for a given actual sequence as compared to a corresponding random sequence and that this difference increases with N. This behavior is clearly evident when one compares the random walks plotted in Figs. 14 and 15.

5. Discussion

We have shown [2] how one can construct the $\zeta(m)$ function from a matrix product using a matrix based on

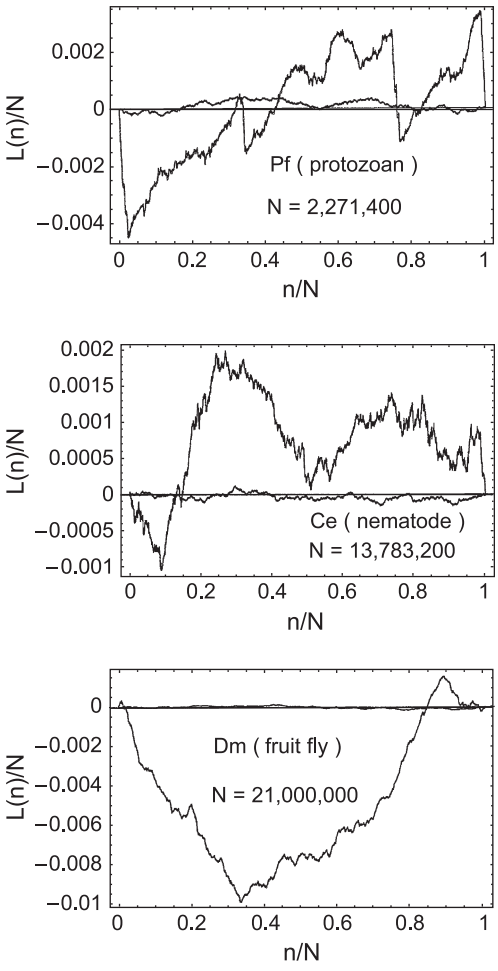

Fig. 15. A plot of the relative random walk length function $L(n)$ defined in Eq. (20) for Pf (protozoan), Ce (nematode) and Dm (fruit fly). In each case, the curve with a smaller amplitude gives $L(n)$ for a random sequence for comparison.

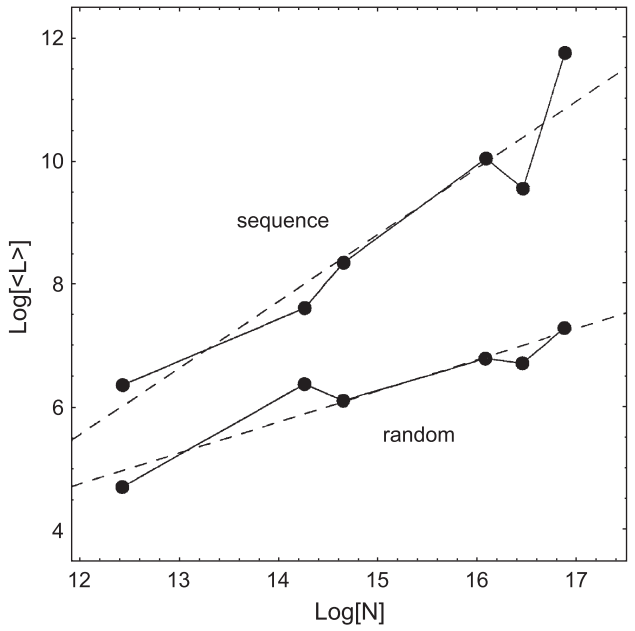

Fig. 16. The natural logarithm of the root-mean-square random walk length defined in Eq. (22) as a function of the natural logarithm of N, the number of bp treated, for the walk functions illustrated in Figs. 14 and 15. The upper set of solid dots is based on the specific sequence for the species treated while the lower set of solid dots gives the results based on the random sequences illustrated in Figs. 14 and 15. The dashed lines give the natural logarithm of the function defined in Eq. (23) using the parameters given in Eq. (24).

statistics for a reference block containing m_o bases. The basic ingredients required to construct the matrix are the conditional probabilities that a block containing j C–G units follows a block containing i C–G units. One can than repeat this construction for reference blocks of different sizes and then compare the $\zeta(m)$ functions obtained for the different m_o values. The result of such a calculation is shown in Fig. 17 for the bacteria Sa using the values m_o=50, 100 and 200. The upper, irregular curve is the actual $\zeta(m)$ function found empirically. One sees that as m_o, the size of the reference block, is increased the calculated values of $\zeta(m)$ approach the empirical curve for $\zeta(m)$ more closely; but the important feature for all of these curves is that the variation of $\zeta(m)$ calculated as a matrix product for a given value of m_o always flattens out as m increases becoming asymptotic to a constant value. In order to fit the empirical $\zeta(m)$ curve for all values of m, it is necessary to consider reference block sizes up to the length of the DNA treated. Thus this persistence is a whole-molecule phenomena that requires the correlation of the entire sequence treated.

Finally, we comment on the relation between the power law exponent γ for the relative width function $\zeta(m)$ as given in Eq. (14) and Mandelbrot's model of a fractional Brownian (random) walk [2,21,22]. Mandelbrot's model describes a random walk in terms of a continuous variable where the walks have a standard Gaussian form. The nonstandard feature of the model is the m dependence of the width parameter σ_m as defined in Eq. (5) which is given by

$$\sigma_m = A\, m^H \tag{25}$$

where

$$H = 1/2 \tag{26}$$

for a regular Brownian walk. The case

$$H > 1/2 \tag{27}$$

represents a fractional Brownian walk where steps in the same direction tend to follow one another more often than random, i.e. there is persistence in the walk process. This is exactly the kind of behavior we find for the C–G distribution in all of the species we have examined.

If we combine Eqs. (13) and (14), we have

$$\sigma_m = A\, m^{1/2+\gamma} \tag{28}$$

Comparing Eqs. (25) and (28), we obtain the following relation between Mandelbrot's exponent H in Eq. (25) and our exponent γ in Eq. (28)

$$H = 1/2 + \gamma \tag{29}$$

For the fractional Brownian walk, the fractal dimension of the walk is given by Mandelbrot as

$$D = 2 - H \tag{30}$$

or using Eq. (29)

$$D = 3/2 - \gamma \tag{31}$$

For a random walk with no persistence, we have γ=0 giving

$$D = 3/2 \tag{32}$$

For the species we have treated here, the fractal dimensions are listed in Table 1. One sees that the effect

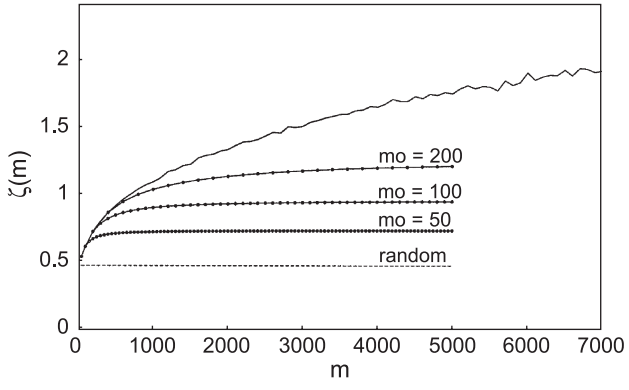

Fig. 17. The function $\zeta(m)$ defined in Eq. (13) for the bacteria Sa. The upper, more irregular, curve gives the empirical function based on the Sa genome while the lower dashed curve is the result for random placement of base pairs. The intermediate curves are obtained from the representation of $\zeta(m)$ in terms of a matrix of empirical conditional probabilities for a given reference box sizes m_o. The solid dots give $\zeta(m)$ based successively on larger reference boxes.

of persistence in the random walk always lowers the fractal dimension. From the result in Eq. (32), one sees that a true random walk has a fractal dimension (tendency to fill space) that is half way between one dimension and two dimensions. The effect of persistence is to move the fractal dimension more toward one dimension, that is, persistence gives one a function that is less space filling.

References

[1] D. Poland, Long-range correlations in the helix free energy distribution in DNA, Biophys. Chemist. 106 (2003) 275–303.

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

[3] C. Tudge, The Variety of Life, Oxford University Press, Oxford, 2000.

[4] R. Himmerlreich, et al., Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*, Nucleic Acids Res. 24 (22) (1996 Nov 15) 4420–4449.

[5] C.M. Fraser, et al., Complete genome sequence of *Treponema pallidum*, the syphilis spirochete, Science 281 (5375) (1998 Jul 17) 375–388.

[6] R.A. Alm, et al., Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*, Nature 397 (6715) (1999 Jan 14) 176–180.

[7] R.D. Fleishmann, et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae*, Rd. Science 269 (5223) (1995 Jul 28) 496–512.

[8] H. Tettelin, et al., Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*, Science 293 (5529) (2001 Jul 20) 498–506.

[9] T. Baba, et al., Genome and virulence determinants of high virulence community-acquired MRSA, Lancet 359 (9320) (2002 May 25) 1819–1827.

[10] R.W. Hyman, et al., Sequence of *Plasmodium falciparum* chromosome 12, Nature 419 (2002) 534–537.

[11] S. Sato, et al., Complete structure of the chloroplast genome of *Arabidopsis thaliana*, DNA Res. 6 (1999) 283–290.

[12] A. Goffeau, et al., Life with 6000 genes, Science 274 (5287) (1996 Oct 25) 563–567.

[13] J. Hodgkin, et al., *C. elegans*: sequence to biology, Science 282 (5396) (1998 Dec 11) 2011.

[14] D.L. Lewis, et al., *Drosophila melanogaster* mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons, Insect Mol. Biol. 4 (4) (1995) 263–278.

[15] E.S. Miller, et al., Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage, J. Bacteriol. 185 (17) (2003) 5220–5233.

[16] S.G. Andersson, et al., The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, Nature 396 (6707) (1998 Nov 12) 133–140.

[17] T. Kawashima, et al., Archaeal adaption to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*, Proc. Natl. Acad Sci. U. S. A. 97 (26) (2000 Dec 19) 14257–14262.

[18] One can obtain the human genome on the worldwide web from UCSC Genome Bioinformatics at the address: http://genome.ucsc.edu.

[19] The worldwide web address of The Institute for Genomic Research is: http://www.tigr.org.

[20] The worldwide web address of the National Center for Biotechnology Informatics, National Library of Medicine, National Institutes of Health is: http://www.ncbi.nlm.nih.gov.

[21] B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, New York, 1982.

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