ELSEVIER

Biophysical Chemistry 110 (2004) 59–72

Biophysical Chemistry

www.elsevier.com/locate/bpc

# The persistence exponent of DNA

Douglas Poland*

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

Accepted 5 January 2004
Available online 5 May 2004

**Abstract**

Using the complete genome of *Thermoplasma volcanium*, as an example, we have examined the distribution functions for the amount of $C$ or $G$ in consecutive, non-overlapping blocks of $m$ bases in this system. We find that these distributions are very much broader (by many factors) than those expected for a random distribution of bases. If we plot the widths of the $C–G$ distributions relative to the widths expected for random distributions, as a function of the block size used, we obtain a power law with a characteristic exponent. The broadening of the $C–G$ distributions follows from the empirical finding that blocks containing a given $C–G$ content tend to be followed by blocks of similar $C–G$ content thus indicating a statistical persistence of composition. The exponent associated with the power law thus measures the strength of persistence in a given DNA. This behavior can be understood using Mandelbrot's model of a fractional Brownian walk. In this model there is a hierarchy of persistence (correlation between blocks) between all parts of the system. The model gives us a way to scale the $C–G$ distributions such that all these functions are collapsed onto a master curve. For a fractional Brownian walk, the fractal dimension of the $C–G$ distribution is simply related to the persistence exponent for the power law. The persistence exponent for *T. volcanium* is found to be $\gamma = 0.29$ while for a 10 million base segment of the human genome we obtain $\gamma = 0.39$, similar to but not identical with the value found for the microbe.

*Keywords:* Persistence exponent; *Thermoplasma volcanium*; DNA; Fractal dimension; Power law

## 1. Introduction

The statistical mechanics of the unwinding of the double helix of DNA is a problem we have been engaged with for many years [1–5]. Recently, we studied the free energy distribution for blocks of double helix and found that there was a strong correlation between the free energy values of neighboring blocks of bases [4]. This correlation is also present when we examine a simpler property of consecutive blocks, namely, the content of $C$ or $G$ and it is the $C–G$ distributions we explore here.

The division of the genome into consecutive non-overlapping blocks containing $m$ bases is illustrated in Fig. 1. The upper row illustrates the assignment of bases to blocks for the case of $m = 10$. In the lower row, we indicate how we assign 0's to A or T bases and 1's to $C$ or $G$ bases and count the number of 1's ($C–G$ units) in each box. We then collect statistics on how many boxes in the molecule contain $n$ $C–G$ units, thus

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

0301-4622/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.01.003

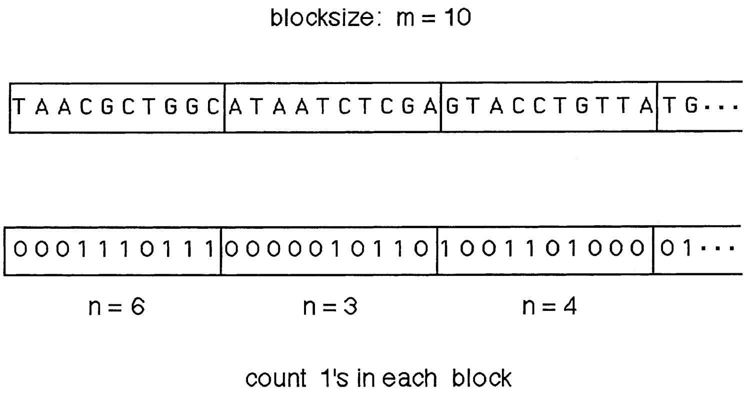

Fig. 1. Illustration of the use of blocks to obtain $C-G$ distributions. The upper row illustrates the division of the genome into consecutive non-overlapping blocks each containing $m$ bases. For the example shown $m = 10$. The lower row indicates the translation of the base composition into 0's ($A$ or $T$) and 1's ($C$ or $G$). One then simply counts how many 1's are in each block and catalogs the distribution.

giving the distribution function for the $C-G$ content in $m$-blocks.

The species we have chosen to treat as an example is that of the archaean *Thermoplasma volcanium* (Tv for short). The genome of this organism, which is found on a single circular chromosome and contains 1, 584, 804 base pairs, was determined by Kawashima et al. [6] and can be obtained from the institute for genomic research (Tigr) on the web [7]. This organism prefers to live in hot water and has an optimal growth temperature of 60 °C. We picked this organism as an example since the size of the genome is manageable and the evidence for a power law is particularly impressive in this species.

## 2. Random distribution for comparison

A key part of our approach is to show exactly how the empirical $C-G$ distributions in Tv differ from random distributions and so we begin by reviewing the basic relations governing random distributions. From the known Tv genome [6], we obtain the overall fractions of the two states ($A$ or $T$ and $C$ or $G$)

$$f_{\mathrm{at}} = \text{fraction } A \text{ or } T = 0.601$$

$$f_{\mathrm{cg}} = \text{fraction } C \text{ or } G = 0.399 \qquad (1)$$

with

$$f_{\mathrm{at}} + f_{\mathrm{cg}} = 1 \qquad (2)$$

If we assume independent units, we can obtain the distribution function for a block of $m$ bases as follows. We construct the generating function, $\Gamma_m$, for the distribution

$$\Gamma_m = \left(f_{\mathrm{at}} + f_{\mathrm{cg}} z\right)^m = \sum_{n=0}^{m} P(n) z^n \qquad (3)$$

where $P(n)$ is the probability that a block of $m$ bases contains $n$ bases that are $C$ or $G$ and is given by

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

The factor $z$ in Eq. (3) is a label parameter to keep track of the number of $C-G$ units; we ultimately set this parameter equal to one.

The nature of any distribution is most simply given in terms of the moments of the distribution. In terms of the generating function for an arbitrary distribution, the following relations give the first two moments of the distribution

$$\mu_1(m) = \sum_{nk=0}^{m} P(n) = \partial \Gamma_m / \partial(\ln z) = \langle n \rangle$$

$$\mu_2(m) = \sum_{n=0}^{m} n^2 P(n) = \partial^2 \Gamma_m \Big/ \partial(\ln z)^2 = \langle n^2 \rangle \qquad (5)$$

where we set $z$=1 after taking the appropriate derivatives. We take as a standard measure of the breadth of the distribution the root-mean-square width

$$\sigma_m = \sqrt{\mu_2(m) - \mu_1(m)^2} \tag{6}$$

We note that the use of this quantity does not necessarily imply that the distributions are Gaussian or random.

For the special case of a random distribution, we can obtain an explicit result for $\sigma_m$ defined in Eq. (6) in terms of the fractions given in Eq. (1). Using the relation for $P(n)$ given in Eq. (4) and the definitions of the moments in Eq. (5) one obtains the following result for the width of the $C$–$G$ distribution in blocks of $m$ bases when the distribution is random

$$\sigma_m(\text{random distribution}) = \sqrt{m}\sqrt{f_{\text{cg}} - f_{\text{cg}}^2} = 0.49\sqrt{m} \tag{7}$$

The square-root of $m$ term in the above equation is a standard result for random walks.

## 3. Distributions

We are interested in how the actual $C$–$G$ distributions obtained from the Tv genome deviate from the dependence on $m$ given in Eq. (7). To this end, we define the following function, which gives the width of the empirical distribution, $\sigma_m$, relative to the $m$-dependence of the random distribution given in Eq. (7)

$$\zeta(m) = \frac{\sigma_m}{\sqrt{m}} \tag{8}$$

If the distribution of the $C$–$G$ content in $m$-blocks is random, then this function should equal a constant value or approach a constant value as $m$ is increased. The deviation of this function from a constant value indicates the existence of an $m$-dependence in addition to the standard square-root behavior for the random distribution indicated in Eq. (7).

In Fig. 2, we illustrate the empirical $C$–$G$ distributions based on the complete genome for Tv for two values of $m$. In the top graph, the solid dots give the empirical distribution function obtained for the block size $m$=20 while the solid line joins the points obtained from the random distribution given by Eq. (4). For this case, the empirical distribution and the random distribution are almost superimposable. Thus, while the specific base sequence contains the genetic information for this organism, the net occurrence of $C$–$G$ is given accurately by a random distribution. In the bottom graph of Fig. 2, we show the results for blocks of $m$=500 where again the solid dots represent the empirical distribution and the solid curve joins the points for the random distribution given by Eq. (4); we will discuss the broader dashed curve shortly. For this case, one sees that the empirical distribution is very much broader than the random distribution. We note that there appears to be some scatter in the

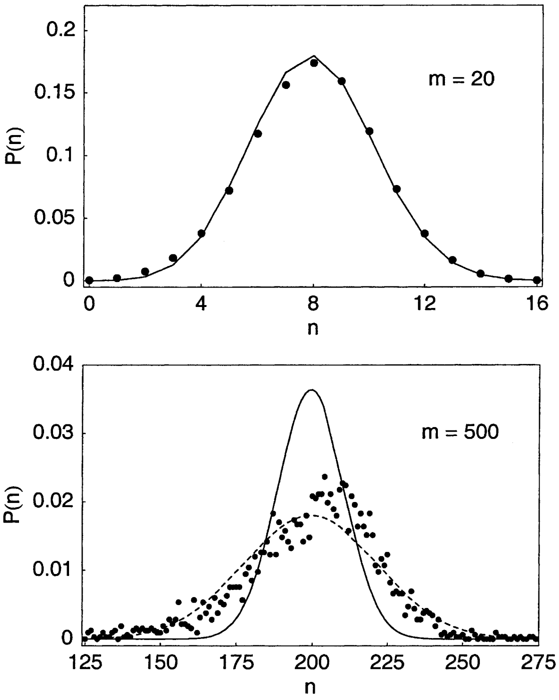

Fig. 2. Distribution functions for the $C$–$G$ content in the Tv genome. $P(n)$ gives the probability that $n$ $C$ or $G$ units will be found in a block containing a total of $m$ units. The solid dots give the empirical points. The upper graph gives the distribution function for the case $m$=20 where the solid curve gives the random distribution given by Eq. (4). The lower graph gives the distribution function for the case $m$=500 where, again, the solid curve gives the random distribution given by Eq. (4) while the lower dashed curve gives the distribution given by Eq. (14) with $H$=0.79.

empirical points. It is important to realize that this is not experimental error since each dot is obtained from the specific information of the complete Tv genome. It is this broadening of the $C-G$ distributions for large $m$ that is our focus in the present paper.

## 4. Plus/minus maps

We can gain some insight as to why the empirical distributions are very much broader than the random distributions for large $m$ by examining the correlation of successive states of $m$-blocks in the molecule. We can do this most simply by examining the statistics for the frequency with which the various states of a block follow one another. Specifically, we 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 expected for a random distribution of block contents, namely, $f_i f_j$. To measure the difference between these two frequencies we construct the function

$$\Delta_{ij} = \mathrm{sign}\left(f_{ij} - f_i f_j\right) \tag{9}$$

which gives the sign of the difference between the empirical doublet frequency and the random doublet frequency. If this quantity is positive, then the particular $i-j$ combination occurs more often than random and vice versa. A plot of the function $\Delta_{ij}$ gives a plus/minus map that simply indicates the nature of the correlation between the $C-G$ compositions of successive $m$-blocks in the molecule.

Fig. 3 shows the quantity $\Delta_{ij}$ defined above for blocks with $m$=100 and $m$=200 where the axes give the possible values of $i$ and $j$. The white squares indicate the $i-j$ combinations where $\Delta_{ij}$ is positive (more correlation than random) while the black squares indicate the $i-j$ combinations where $\Delta_{ij}$ is negative (less correlation than random); gray indicates that there is no difference. For the case of $m$=100 shown in Fig. 3, we see that the white squares tend to lie along the diagonal axis $i$=$j$ with a noticeable concentration of points at the extreme values of $i$ and $j$. This means that there is a distinct tendency for blocks containing $n$ $C-G$ units to be followed by blocks with similar $C-G$ content, particularly for small and large values of $n$. Thus there is a persistence of $C-G$ content as one goes along the chain from one $m$-block to the next. For the case of $m$=200, the pattern of positive and negative correlations is similar to that seen for the case of $m$=100 but with more scatter of the points. We note that the correlations we are discussing are

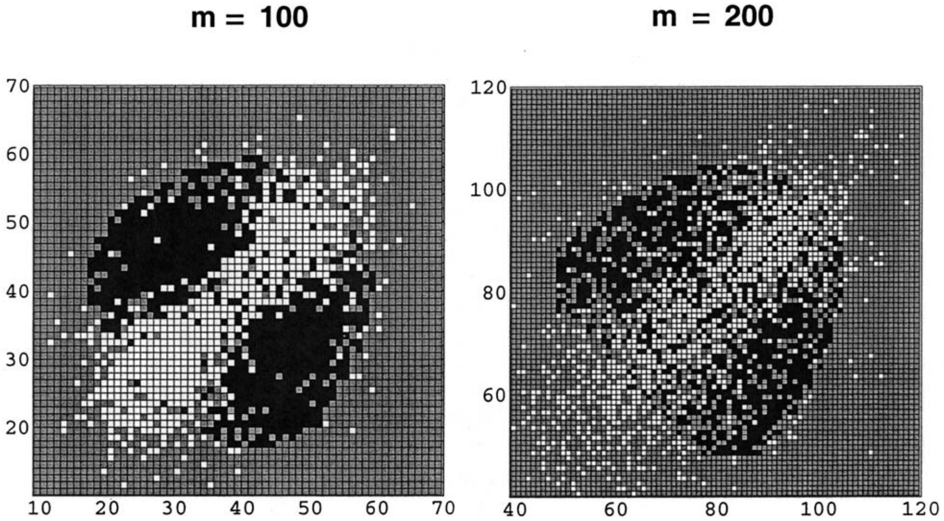

Fig. 3. Plus/Minus maps for the $C-G$ content in the Tv genome. The axes index the number of $C-G$ units in consecutive blocks for the cases of $m$=100 and $m$=200 units. The white squares indicate that state-$i$ tends to be followed by state-$j$ more often than random occurrence, while the black squares indicate just the opposite.

on a large scale, specifically, in Fig. 3, between blocks with $m$=100 and $m$=200. The trend for persistence means that blocks with few or many $C$–$G$ units will be more probable than random (they tend to follow one another) and hence the wings of the $C$–$G$ distribution are made more probable relative to the random distribution [4] thus giving the broadening illustrated in Fig. 2.

## 5. Power law

We now show how the empirical width function defined in Eq. (8) varies as a function of $m$. This is shown in Fig. 4 where we plot $\zeta(m)$, as a function of $m$, where the solid dots give $\zeta(m)$ for $m$ values in steps of 100 from $m=100$ to $m=12\,000$. The dashed curve shows the value of $\zeta(m)$ one would obtain if the distributions were random, namely, the constant 0.49 given in Eq. (7). We will explain the solid curve shortly. For small values of $m$ (such as $m=20$ as illustrated in Fig. 2) there is no difference between the actual width of the distribution and that for a random distribution. However, as the value of $m$ is increased the actual width of the distribution becomes much larger than the corresponding width for a random distribution. One sees in Fig. 4 that $\zeta(m)$ for the empirical distributions is larger than the same quantity for random distributions by a factor of two-to-five in the range shown and thus the distribution widening we find is a major effect.

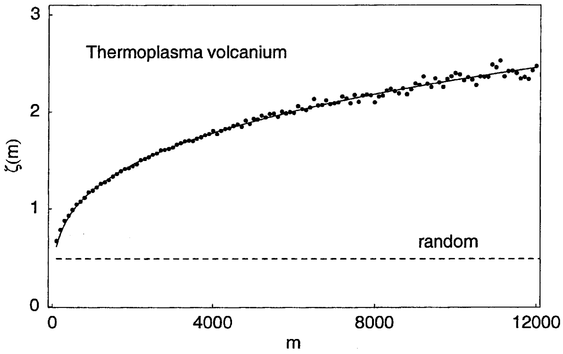

Fig. 4. The relative widths of the $C$–$G$ distributions for the Tv genome as a function of block size. The width function $\zeta(m)$ defined in Eq. (8) is plotted as a function of $m$ (the block size). The solid dots give the empirical points while the solid curve is a plot of the power law given in Eq. (10) with the persistence exponent $\gamma=0.29$. The dashed curve gives the result for random distributions.

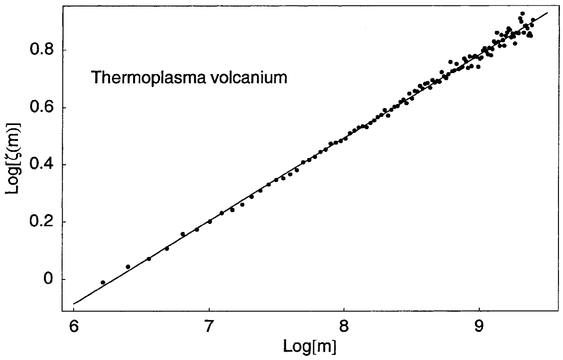

Fig. 5. A log–log plot of the data given in Fig. 4. The solid line is the linear fit to the data.

The behavior of $\zeta(m)$ in Fig. 4 strongly suggests a power law of the form

$$\zeta(m) = Am^{\gamma} \tag{10}$$

The standard test for power law behavior is a log–log plot and this is shown in Fig. 5 where log[$\zeta(m)$] is plotted as a function of log[$m$], using natural logarithms for both. The solid line gives a linear fit to the data, which is seen to be excellent. The results of this fit are

$$A = 0.161 \text{ and } \gamma = 0.290 \tag{11}$$

As previously noted, the small irregularities in $\zeta(m)$ at large values of $m$ do not represent experimental error since these values are obtained from the exact genome for Tv. For simplicity, we will round off the value of the exponent to give $\gamma=0.29$. The power law defined in Eq. (10) with the parameters given in Eq. (11) is plotted (the solid line) in Fig. 4 and is seen to give an excellent fit to the empirical points. This power law holds very well up to $m=12\,000$. For $m$-values larger than this, variations in $\zeta(m)$ become more pronounced and there is a slow tendency for the values of $\zeta(m)$ to begin to fall below those given by the power law curve. Of course, ultimately the power law must fail since the DNA of Tv is finite (1, 584, 804 bp).

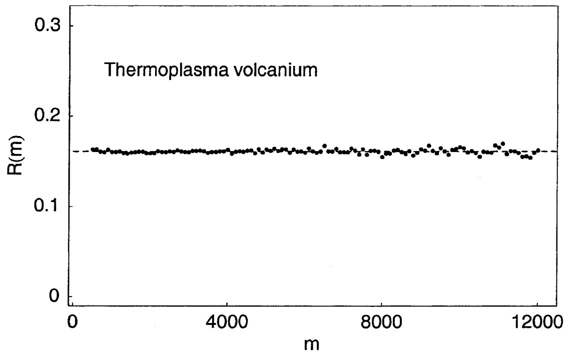

Fig. 6. Goodness of fit of the power law. A plot of the function $R(m)$ defined by Eq. (12) using the persistence exponent $\gamma = 0.29$.

To see how good a model the power law given in Eq. (10) is at fitting the data given in Fig. 4 we form the function

$$R(m) = \zeta(m)/m^{\gamma} \tag{12}$$

This function is plotted in Fig. 6 over the range $m = 100 - 12\,000$. The root-mean-square variance of $R(m)$ from the value of $A = 0.161$ is $\pm 0.0025$ and one sees that there is no tendency for this function to either increase or decrease from a constant value over the range shown. Thus, the function $\zeta(m)$ shown in Fig. 4 is very accurately represented by a power law over the range shown.

The exponent $\gamma$ for the power law given in Eq. (10) measures the extent of the broadening of the $C$–$G$ distributions as given in Fig. 4. In turn the broadening is a result of the persistence of $C$–$G$ content from block to block as illustrated by Fig. 3. Thus, the exponent $\gamma$ is a measure of the persistence of $C$–$G$ content characteristic of a given DNA.

## 6. Model of fractional Brownian walk

The net variation of $\sigma_m$ with $m$ is obtained by combining Eqs. (8) and (10) to give

$$\sigma_m = Am^{1/2+\gamma} \tag{13}$$

A similar relation for the width of a distribution occurs in Mandelbrot's model of a fractional Brownian walk [8,9]. This is a model for a random walk in terms of a continuous variable $x$ (the analog or our discrete variable $n$), where the walks have a standard Gaussian distribution

$$B_H(x) = (\sigma_m)^{-1}\exp[-(x-\bar{x})^2/2\sigma_m^2] \tag{14}$$

where $\bar{x} = f_{cg}m$ is the mean position. What is non-standard in this model is the width parameter $\sigma_m$, which is taken to follow a power law with respect to $m$. In Mandelbrot's notation $\sigma_m$ is given by

$$\sigma_m = Am^H \tag{15}$$

For an ordinary Brownian walk, one has

$$H = 1/2 \tag{16}$$

while if $H$ is not equal to 1/2, then one has a fractional Brownian walk. In particular, if one has the case

$$H > 1/2 \tag{17}$$

then one has a fractional Brownian walk with persistence, which means that steps in the same direction tend to follow one another more often than random, which is exactly the case we find for the $C$–$G$ distributions in the Tv genome as graphically illustrated by the plus/minus maps in Fig. 3. Comparing Eqs. (13) and (15), we obtain the value of $H$ for the Tv genome (where $\gamma = 0.29$ from Eq. (11))

$$H = 1/2 + \gamma = 0.79 \tag{18}$$

Eq. (14) can be viewed as the continuous analog of the discrete $C$–$G$ distribution. For an ordinary Brownian walk, one has $H = 1/2$ and the value of $\sigma_m$ is then given by Eq. (7). Under these conditions Eq. (14) is the continuous analog of the discrete random distribution given by Eq. (4); it gives, for example, an extremely close fit to the discrete random distribution for $m = 20$ shown by the solid line in Fig. 2. For a fractional Brownian walk with persistence, where $H > 1/2$, Eq. (14) also is the continuous analog of the discrete block distribution, but the width of the distribution, given by Eq. (15), is now much broader than that for a random distribution. An important property of the fractional Brownian walk is that the persistence of the walk not only involves neighboring blocks, but requires correlations that span the entire genome.

We illustrate Mandelbrot's model of a fractional Brownian walk by using it to describe the $C-G$ distribution in Fig. 2 for blocks of $m=500$. The dashed curve shown in Fig. 2 is a plot of the Gaussian distribution given in Eq. (14) using $\sigma_m$ given by Eq. (15) with $H=0.79$, and is seen to accurately duplicate the width of the empirical Tv distribution. We note that the use of $\sigma_m$ from Eq. (15) with $H=1/2+\gamma$ and the relation for the mean, $\bar{x}=f_{\mathrm{cg}}m$, guarantees that Mandelbrot's model will have the same first and second moments as the empirical $C-G$ distributions in DNA.

## 7. Generating functions as matrix products

We mentioned above that the model of a fractional Brownian walk requires correlations that span the entire genome. We illustrate that feature here by showing that correlations that are limited to neighboring blocks are not sufficient to explain the power law behavior for $\zeta(m)$ shown in Fig. 4. The idea is illustrated in Fig. 7. In Fig. 7a, we illustrate a chain of nearest-neighbor correlations between blocks of a given size. We will show that for this case $\zeta(m)$ is always asymptotic to a constant value as $m$ increases. Fig. 7b illustrates a hierarchy of correlations where each block has direct correlations with blocks down the chain. It is this latter type of correlation that is required to obtain the power law behavior we find for $\zeta(m)$.

We begin our construction of generating functions for $C-G$ distributions with a correlation table similar to that used for the plus/minus maps illustrated in Fig. 3. We first choose a reference block containing $k$ bases. The correlation table considers the possible states $i$ of a given $k$-block with the possible states $j$ of the following $k$-block where the values of $i$ and $j$ can run from zero to $k$. The main ingredient in the table is the conditional probability that given state $i$, it is followed by state $j$. This is constructed from the

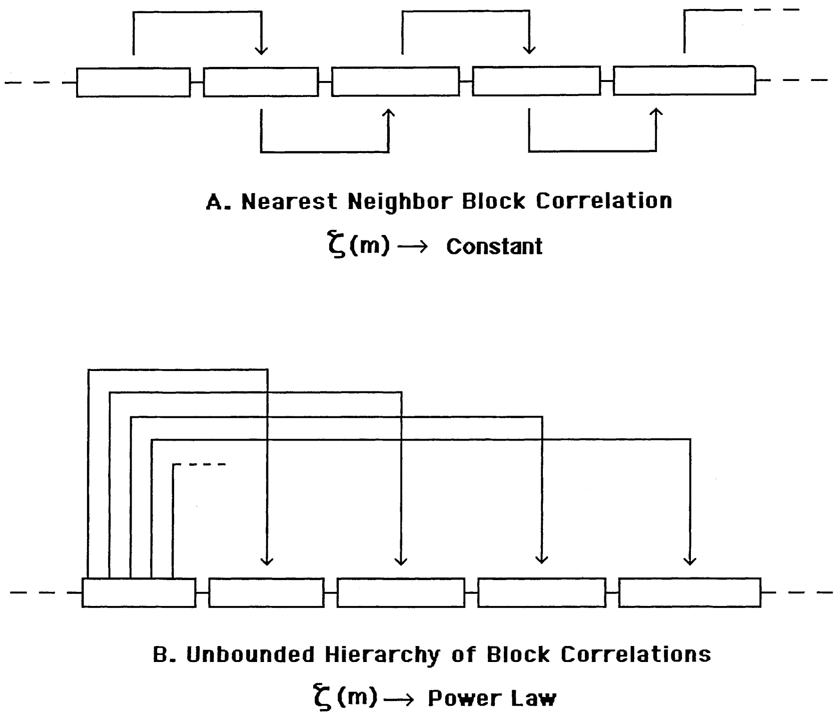
A. Nearest Neighbor Block Correlation
ζ(m) ⟶ Constant
B. Unbounded Hierarchy of Block Correlations
ζ(m) ⟶ Power Law

Fig. 7. Different types of inter-block correlation. (a) The schematic illustration of the correlation of nearest-neighbor blocks. This type of correlation leads to a function $\zeta(m)$ that is asymptotic to a constant value. (b). The schematic illustration of the case where there is an unbounded range of correlation. This type of correlation leads to a function $\zeta(m)$ that is asymptotic to a power law.

empirical fractions of states we introduced in Eq. (9) as follows:

$$P_k(i|j) = f_{ij}/f_i \tag{19}$$

We then define a $(k+1)$ by $(k+1)$ matrix, the general element of which is given by the conditional probability given in Eq. (19) multiplied times a label parameter $z$ that is required to count $C-G$ states

$$\mathbf{P}_k = \left[P_k(i|j)z^i\right] \tag{20}$$

This matrix is then used to construct a generating function analogous to that given in Eq. (3) for the case of independent units

$$\Gamma_m = \mathbf{p}_k \mathbf{P}_k^{L-1} \mathbf{v}_k \tag{21}$$

where $m=kL$ and $L$ is any positive integer. Thus, while the generating function given in Eq. (3) generates the $C-G$ distributions for the case of independent (random) units, the generating function given in Eq. (21) generates the $C-G$ distributions when empirical correlations between the $C-G$ content of successive $k$-block are taken into account.

In Eq. (21) the vectors $\mathbf{p}_k$ and $\mathbf{v}_k$ are required, respectively, to initiate and terminate the chain. The general element of the vector $\mathbf{p}_k$ is simply the a priori probability of state $i$ multiplied times a factor of $z^i$ to count $C-G$ states while the general element of $\mathbf{v}_k$ is simply one. Thus the two vectors are given by

$$\mathbf{p}_k = \left[f_i z^i\right], \quad \mathbf{v}_k = [1] \tag{22}$$

Given the generating function of Eq. (21) constructed using the matrix of Eq. (20) and the vectors of Eq. (22), one then uses Eqs. (5) and (6) to calculate the width function $\zeta(m)$ of Eq. (8). For large values of $L$, the generating function $\Gamma_m$ can be expressed in terms of the largest eigenvalue, $\lambda_1(k,z)$, of the matrix $\mathbf{P}_k$ as follows

$$\text{(limit of large } L) \quad \Gamma_m \sim \lambda_1(k,z)^L \tag{23}$$

Setting the label parameter $z=1$ one has $\lambda_1(k)=1$, which follows since $\Gamma_m$ is the sum of the probabilities of all possible states (see Eq. (4)) which must sum to one. Using the asymptotic form for $\Gamma_m$ given in Eq. (23), it follows that for large $m$ the function $\zeta(m)$ is asymptotic to a constant value independent of $m$. Thus a calculation of the width function $\zeta(m)$ will always give a curve that becomes flat for large $m$ and can never give a power law dependence of the type shown in Fig. 4.

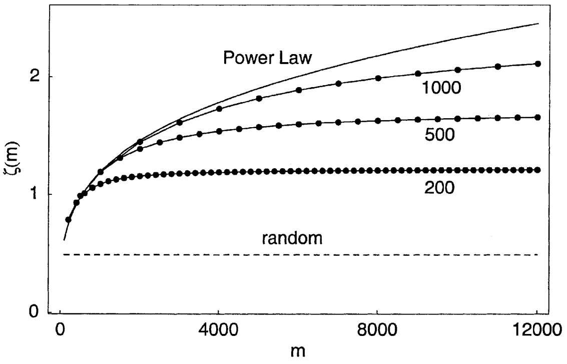

Fig. 8. Calculation of the width function from correlation matrices. The curves show the construction of the function $\zeta(m)$ obtained from the matrix-product generating function given in Eq. (21) based on blocks of $k=200$, 500 and 1000 bases. All of these curves are asymptotic to constant values as m is increased. The power law given by Eq. (10) is shown for comparison.

We illustrate the behavior just described in Fig. 8 where we show $\zeta(m)$ functions calculated from the generating function given in Eq. (21) based on the reference blocks with $k=200$, 500, and 1000. The dashed curve gives the result for random distributions while the upper solid curve gives the power law of Eq. (10) using the parameters of Eq. (11). One sees that the larger the value of $k$ (the size of the reference block) used to construct the generating function, the closer the $\zeta(m)$ points come to the empirical power law. However, one also sees that the $\zeta(m)$ curves for all of the $k$-values shown eventually become independent of $m$. Thus the nearest-neighbor correlation of $k$-blocks, as indicated in Part A of Fig. 7, will never produce the empirical power law behavior found in Fig. 4. In order to obtain true power law behavior the hierarchy of block correlations, as shown in part B of Fig. 7, is required.

## 8. Scaling

In addition to implying a hierarchy of interaction over the whole molecule, the fractional Brownian model has two additional features. The first is the property of scaling whereby a proper choice of scaled variables can result in the collapse of all of the distribution functions for different values of the block size $m$ onto a single distribution. From the continuous Gaussian function of Eq. (14) we obtain the following scaling relations

$$y = (x - x_m)/\sigma \quad \text{and} \quad \mathrm{d}y = \mathrm{d}x/\sigma \tag{24}$$

which gives the scaled function

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left[-y^2/2\right] \tag{25}$$

The analog of the continuous distribution function given in Eq. (14) is the discrete sum

$$\sum_{n=0}^{m} P(n)\Delta = 1 \tag{26}$$

where Δ=1. The scaling for this sum that is analogous to that given in Eq. (24) is

$$n' = n/\sigma \quad \text{and} \quad \Delta' = \Delta/\sigma \tag{27}$$

and

$$P(n') = \sigma P(n) \tag{28}$$

The scaling indicated above preserves the normalization of the sums

$$\sum_{n} P(n)\Delta = \sum_{n'} P(n')\Delta' = 1 \tag{29}$$

In Fig. 9, we illustrate the scaling outlined above for the empirical discrete distributions and for the continuous Mandelbrot model. The solid dots in the upper graph give the empirical distributions for the probability of finding $n$ $C$–$G$ units in blocks of $m$=100 and $m$=200 units. The solid curves in the same graph give the corresponding Mandelbrot distributions given by Eq. (14) with $H$=0.79. The lower graph shows the scaled distributions for both the discrete and continuous cases using the scaling relations given above. The scaled continuous distributions are exactly the same (one curve) while the discrete points are merged onto a single curve to a very good approximation.

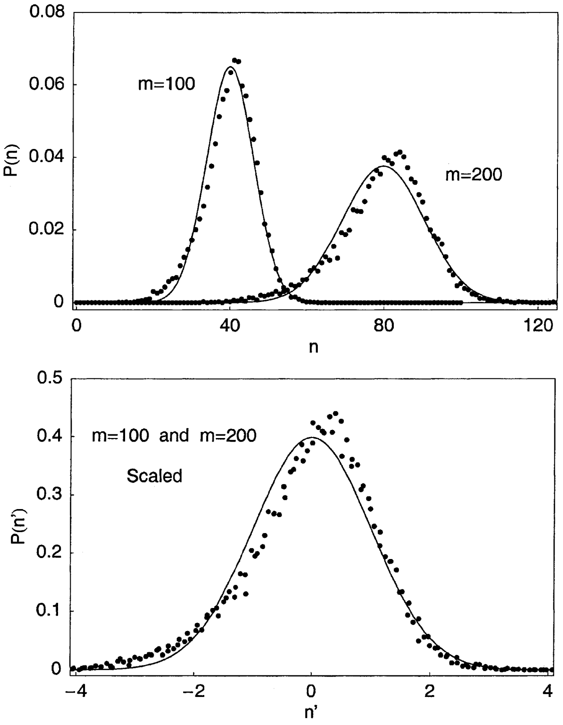

Fig. 9. Scaling of the $C$–$G$ distributions. The curves illustrate scaling of $C$–$G$ distributions from the Tv genome. The solid dots in all cases are the empirical points while the solid curves are the distribution functions given by Eq. (14) with $H$=0.79. The upper curves give the unscaled $C$–$G$ distributions for $m$=100 and $m$=200 while the lower curves give the scaled versions of these distributions using the scaling relations given in the text.

## 9. Fractal dimension

Finally, the model of the fractional Brownian walk allows us to calculate the fractal dimension of DNA with respect to the $C$–$G$ distributions we have been examining. Mandelbrot [8,9] gives the following relation for the fractal dimension of a fractional Brownian walk:

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

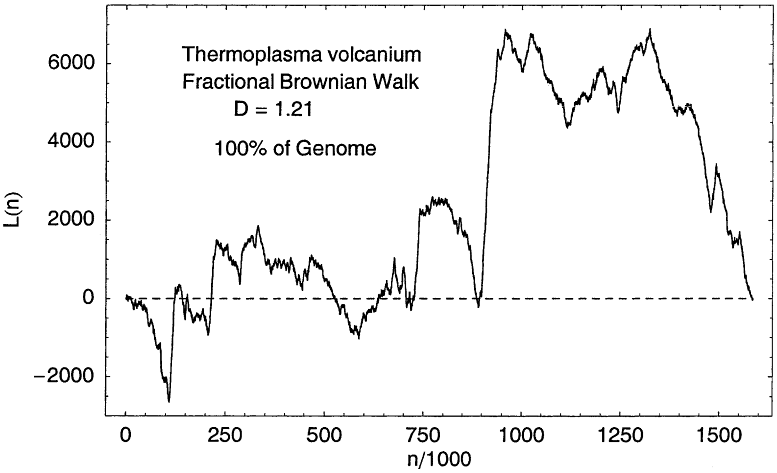

Fig. 10. A fractal walk based on the Tv genome. The curve is the fractional Brownian walk given by Eq. (34) for the entire Tv genome. The fractal dimension for this curve is $D=1.21$.

For ordinary Brownian motion $H=1/2$ and we have

$$D = 1.50 \tag{31}$$

For the value of $H=0.79$ characteristic of the Tv genome, we find

$$D = 2-0.79 = 1.21 \tag{32}$$

We can construct a fractal curve for DNA based on the $C-G$ content as follows. We assign each base pair in the chain the number $-1$ or $+1$ as follows:

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

$$\alpha_i = +1 \quad \text{if } A \text{ or } T \tag{33}$$

This is analogous to the toss of a coin where one assigns a factor of +1 if heads and −1 if tails. One then simply sums over these factors from the first unit in the chain out to a general site n defining the function

$$L(n) = \sum_{i=1}^{n} \alpha_i - n\Delta f \tag{34}$$

where $\Delta f = f_{at} - f_{cg}$. The $\Delta f$ term is included to make the average value of $L(n)$ equal to zero.

In Fig. 10, we show the curve defined in Eq. (33) for the complete Tv genome, that is, for $n=1$ to 1, 584, 804. The jagged appearance of this curve is characteristic of a fractal curve. It is interesting to note that the fractal dimension of a fractional Brownian walk as given in Eq. (30) decreases as $H$ increases above the value of $H=1/2$ (for a random Brownian walk). This is due to the persistence of the walk when $H>1/2$, which

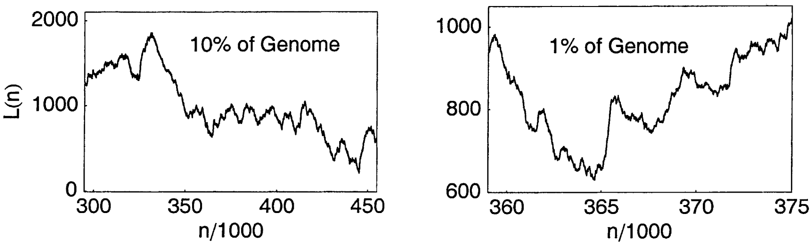

Fig. 11. A fractal walk based on the Tv genome. The curves shown zoom in on portions of the fractal curve shown in Fig. 10 to give, respectively, 10% and 1% of the Tv genome.

produces dramatic features on a large scale, but is less jagged (hence less fractal) on a local scale.

In Fig. 11, we show enlarged sections of the fractal curve shown in Fig. 10. In the left-hand graph, we zoom in by a factor of 10 and show the fractal curve for 10% of the genome while in the right-hand graph we zoom in by a factor of 100 and show 1% of the total curve. In each case the resultant curve looks just as jagged as the original graph for the overall genome shown in Fig. 10. The fact that the curves look qualitatively the same on all scales is a hallmark of fractal curves.

Fig. 12 shows three fractal curves constructed from random distributions based on the overall fractions of $A-T$ and $C-G$ for the Tv genome given in Eq. (1). These graphs give the fractal curves for ordinary Brownian walks illustrating the case of $D=1.50$. On comparing the scales used in Figs. 10 and 12, one sees that the fractional Brownian walk shown in Fig. 10, based on the actual Tv genome, undergoes much larger fluctuations in magnitude than do the walks based on a random distribution of steps as shown in Fig. 12. This is due to the presence of persistence in the case of the Tv genome, that is, the walker has a tendency to keep going in the same direction. The ordinary Brownian walk is more jagged on a local scale (hence $D=1.50$) than the fractional Brownian walk with $D=1.21$.

## 10. Other genomes

The question of course arises as to whether the occurrence of a power law with a characteristic exponent $\gamma$ describing the width of the $C-G$ distributions is a general feature of the DNA for all organisms. In addition to the behavior of the genome for *T. volcanium* described here, we have studied the genomes of seven bacteria: *Mycoplasma pneumoniae*, *Treponema pallidum*, *Chlamydia pneumoniae*, *Haemophilus influenzae*, *Helicobacter pylori*, *Streptococcus pneumoniae* and *Staphylococcus aureus*. All of these species exhibit power-law behavior for the width of the $C-G$ distributions similar to that shown in Fig. 4. In a future publication, we will consider these results in detail. We

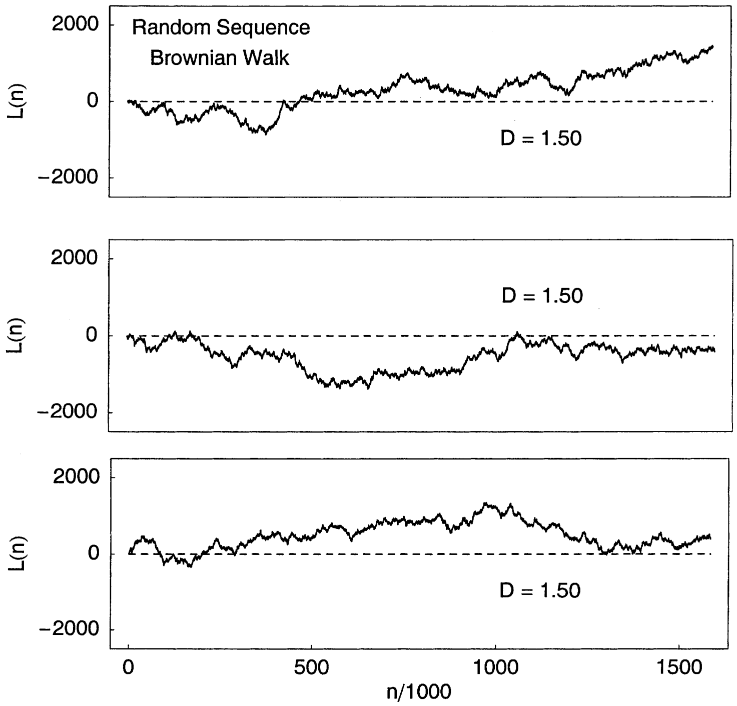

Fig. 12. Three examples of random Brownian walks. The walks shown are derived from random sequences of steps with the overall occurrence statistics of the Tv genome.

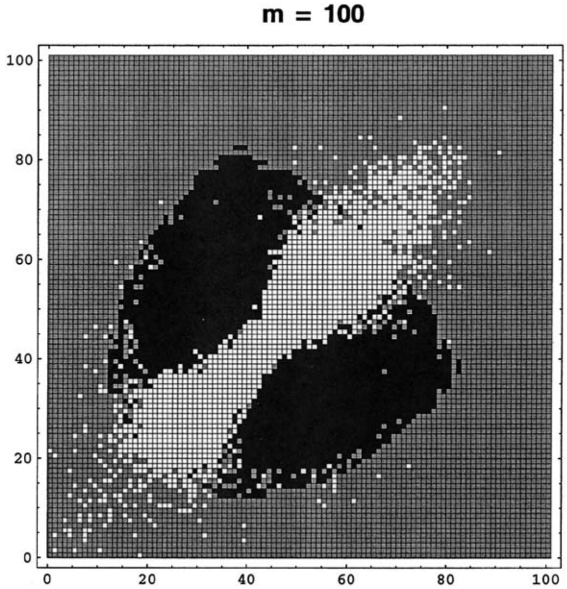

Fig. 13. Plus/Minus map for the $C-G$ content in a section of the human genome. The graph shows the statistics obtained from blocks with $m=100$. The axes index the number of $C-G$ units in consecutive blocks. The white squares indicate that state-$i$ tends to be followed by state-$j$ more often than random occurrence, while the black squares indicate just the opposite.

have found [4] one organism, namely the species *Rickettsia prowazekii*, that does not show this behavior. The function $\zeta(m)$ for this organism is qualitatively like that shown in Fig. 8 for the case of $k=200$. Andersson et al. [10] note that this species contains the highest proportion of non-coding DNA of any known bacteria, namely 24%. One can speculate that the occurrence of the junk DNA breaks up the persistence that is required for the power law behavior and that this is the reason why the power law is not found in this species.

We conclude by examining the properties of a section of the human genome. There are still a number of base pairs in the human genome whose identity is not yet resolved. For study, we chose a section of 10 million base pairs in chromosome #3 that contains only a small number of undetermined base pairs [11]. Specifically, for the sequence $n=1\,000\,001-11\,000\,000$ in chromosome #3, there are five isolated and well-separated unknown base pairs. We arbitrarily took these all as adenine, which has a negligible effect on the overall distributions. For this section of the human genome, we have $f_{\rm at}=0.5936$ and $f_{\rm cg}=0.4064$. Since 97% of the human genome is non-coding (junk DNA) one might anticipate that the persistence required to give power law behavior would be severely disrupted as we argued for the case of *Rickettsia* mentioned above. However, that is not the case.

In Fig. 13, we give the plus/minus map defined in Eq. (9) for our section of the human genome for the case of $m=100$. This map looks very much like the corresponding maps for Tv given in Fig. 3. Recall that the white squares indicate more correlation than random and the black squares the opposite. Thus, in this section of the human genome, there is a strong tendency for blocks to follow like blocks with respect to overall $C-G$ content thus leading to broadening of the $C-G$ distribution. This broadening is illustrated in Fig. 14 where the solid dots give the empirical $C-G$ distribution for the case of $m=100$ and the solid curve is the random distribution given by Eq. (4). As indicated in Fig. 13, one finds that the empirical $C-G$ distribution is much broader that the random distribution, indicating a persistence of correlation of $C-G$ content from block to block.

The function $\zeta(m)$, giving the relative width function of Eq. (8) for our section of the human genome, is shown in Fig. 15 where we give the points for $m=100$ to $m=20\,000$ in steps of 200. The dashed curve shows the constant value for the case of random distributions. As with the similar curve for the Tv genome given in Fig. 4, the behavior of the function strongly suggests a power law of the form given in Eq. (10). A Log–Log plot

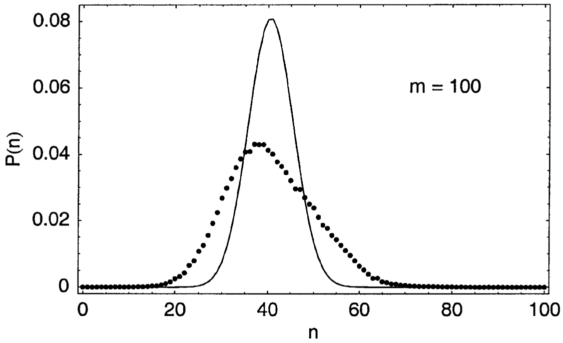

Fig. 14. Distribution functions for the $C-G$ content in a section of the human genome. $P(n)$ gives the probability that $n$ $C$ or $G$ units will be found in a block containing a total of $m=100$ units. The solid dots give the empirical points while the solid line gives the random distribution given by Eq. (4).

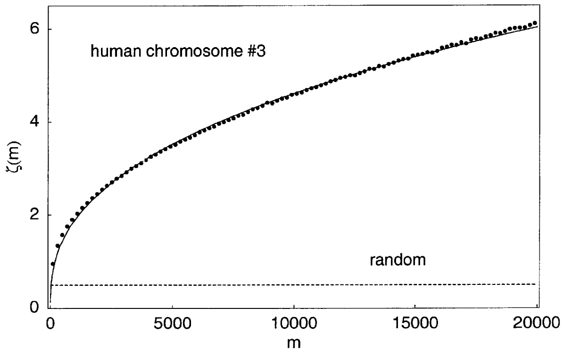

Fig. 15. The relative widths of the $C-G$ distributions for a section of the human genome as a function of block size. The width function $\zeta(m)$ defined in Eq. (8) is plotted as a function of $m$ (the block size). The solid dots give the empirical points while the solid curve is a plot of the power law given in Eq. (10) with the persistence exponent $\gamma = 0.39$. The dashed curve gives the result expected for a random distribution.

gives a very linear relation similar to that shown in Fig. 5 for the Tv genome. A fit of that linear plot gives the following parameters

$$A = 0.131 \text{ and } \gamma = 0.387. \tag{35}$$

Forming the function $R(m)$ defined in Eq. (12), one finds that this function has the constant value $A = 0.131$ over the range $m = 100$ to $m = 20\,000$ with a variance of ± 0.0034. We round off the value of the persistence exponent $\gamma$ to give $\gamma = 0.39$, which we compare with the value of $\gamma$=0.29 found for the Tv genome.

The $H$ parameter for the fractional Brownian walk in this case is $H = 1/2 + \gamma = 0.89$, which gives the fractal dimension for the 10 million base pair sequence in human chromosome #3 as

$$D = 2 - H = 1.11 \tag{36}$$

In Fig. 16, we show the behavior of the function $L(n)$, defined by Eq. (34), for this piece of the human genome. We note that this curve with $D = 1.11$ is less fractal than that for Tv shown in Fig. 10 with $D = 1.21$. The reason for this is that the larger the exponent $\gamma$ the more persistence there is in the walk, which enhances large overall features and reduces local features. Notice the difference in the vertical scales used in Figs. 10 and 16.

The most important consequence of the power law behavior that we find in DNA is that it requires correlation between all units in the genome. We conclude by noting that a property of macromolecules that also involves a power law and similarly requires the consideration of interactions between all of the units in the chain is that of the end-to-end distance of the macromolecule [12] when one includes the effect of excluded volume. For a chain containing $m$ units, the mean end-to-end distance varies as the square root of $m$ if excluded volume is not taken into account. If one does take excluded volume into account then one finds that the end-to-end distance now varies as a

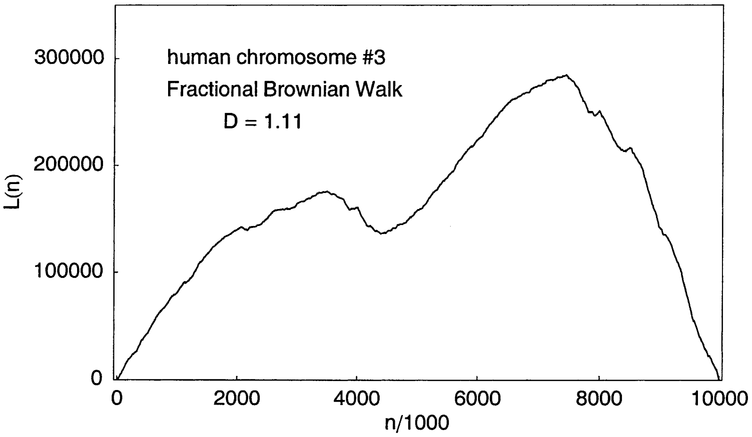

Fig. 16. A fractal walk based on a section of the human genome. Plotted is the fractional Brownian walk given by Eq. (34) for a 10 million base pair sequence of the human genome. The fractal dimension for this curve is $D = 1.11$.

power law, $m^v$ with $v$=3/5 for polymers in three dimensions. The question remains as to whether the power law that we find for the width of the $C-G$ distributions in DNA, reflecting the persistence of composition, is due to some strictly physical mechanism such as excluded volume or is an active product of evolution.

## References

[1] D. Poland, Biopolymers 13 (1859) 1974.

[2] D. Poland, H.A. Scheraga, Theory of Helix–Coil Transitions in Biopolymers, Academic Press, New York, 1970, p. 1970.

[3] D. Poland, Biophys. Chem. 104 (2003) 279.

[4] D. Poland, Biophys. Chem. 106 (2003) 275.

[5] D. Poland, Biopolymers (2003) (in press).

[6] T. Kawashima, N. Armano, H. Koike, S. Makino, Y. Kawashima-Ohya, K. Watanabe, et al., Proc Natl Acad Sci USA 97 (26) (2000) 14 257.

[7] On the world-wide web, Tigr (The institute for genomic research) can be reached at the address: http://www.tigr.org.

[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] S.G. Andersson, A. Zomorodipour, J.O. Andersson, T. Sicheritz-Ponten, U.C. Alsmark, R.M. Podowski, et al., Nature 396 (6707) (1998) 133.

[11] One can obtain the human genome on the world-wide web from UCSC Genome Bioinformatics at the address: http://genome.ucsc.edu.

[12] P.G. de Gennes, Scaling Concepts in Polymer Physics, Cornell University press, Ithaca, New York, 1979, Chapter I.