Biophysical Chemistry 106 (2003) 275–303

Biophysical Chemistry

www.elsevier.com/locate/bpc

Long-range correlations in the helix free energy distribution in DNA

Douglas Poland*

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

Received 27 June 2003; accepted 14 July 2003

Abstract

In this paper we explore the free energy distribution in the helical form of DNA using the genome of the virus *Rickettsia prowazekii* Madrid E as an example. The genome of this organism has been determined by Andersson et al. (Nature 396 (1998) 133) and is available on the World Wide Web (www.tigr.org). Using the helix statistical weights based on nearest-neighbor base pairs of SantaLucia (Proc. Natl. Acad. Sci. USA 95 (1998) 1460), we calculate the free energy in consecutive blocks of m base pairs in the DNA sequence and then construct the free energy distribution for these values. Using the maximum-entropy method we can fit the distribution curves with a function based on the moments of the distribution. For blocks containing 10–20 base pairs the distribution is slightly skewed and we require four moments to accurately fit the function. For blocks containing 100 base pairs or more, the distribution is well approximated by a Gaussian function based on the first two moments of the distribution. We find that the free energy distribution for $m=20$ can be reproduced using random sequences that have the local (singlet, doublet or triplet) statistics of *Rickettsia*. However, for much larger blocks, for example $m=500$, the width of the free energy distribution based on the actual *Rickettsia* genome is broader by almost a factor of 3 than the distributions based on random local statistics. We find that the distribution functions for the C or G content in blocks of m base pairs have almost the same behavior as a function of block size as do the free energy distributions. In order to duplicate the width of the distribution functions based on the actual *Rickettsia* sequence, we need to introduce tables (matrices) that correlate the states of consecutive blocks hundreds of base pairs long. This indicates that correlations on the order of the number of base pairs contained in the average gene are required to give the actual widths for either the C or G content or the helix free energy distributions. Above a certain m value, the distributions for larger m can be accurately expressed in terms of the distribution functions for smaller m. Thus, for example, the distribution for $m=5000$ can be expressed in terms of the generating function for $m=1000$.

Keywords: DNA statistical mechanics; Helix free energy distributions; Distribution moments; Maximum-entropy method; C or G content distributions; Correlation tables

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

0301-4622/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0301-4622(03)00213-8

1. Introduction

We recently published a paper on the statistical mechanics of the unwinding of the double helix of DNA using the genome of *Treponema pallidum*, the syphilis spirochete, as an example [4]. The statistical mechanical methodology required for that calculation was based on a previous paper [5] that outlined a method for calculating the statistical mechanics of a specific-sequence molecule with long-range correlations (in DNA this is the entropy effect of closed loops formed when the double helix is unwound in the interior of the molecule). Recently, the melting profile of the entire human genome has been determined [6] using the method mentioned above as a starting point.

One of the main features of the melting profiles found in these studies is that when the molecule is, say, half unfolded, there are distinct, clearly defined regions that are helix and regions that are clearly coil. The distribution function for helix probability is essentially a step-function, either zero or one, over a range of thousands of base pairs. This phenomenon is partly due to the long-range entropy effect of the loops and partly due to the intrinsic free energy of the helix arising from the interaction of base pairs. In the present paper, we will examine the distribution of the free energy in the helix form of DNA alone in order to see if we can determine some simple properties of this distribution that will shed light on the more complex problem of the complete equilibrium between helix and coil.

A key ingredient for this study is the helix statistical weight of a given base pair at site i in the chain followed by another given base pair at site $i+1$ relative to the free-coil state (no loops) as a reference. This is the analog of the s parameters used by Zimm and Bragg [7] and Zimm [8] in treatments of the helix–coil transition in polypeptides and DNA, respectively. SantaLucia [3] has given a consistent set of the required s parameters for the double helix of DNA. The s factors contain contributions from hydrogen bonds between bases and stacking interactions between neighboring base pairs and are expressed in terms of the following free energies as given in his Table 2:

$$\begin{aligned}
&\mathrm{G}_1=-7.9+0.0222T,\\
&\mathrm{G}_2=-7.2+0.0204T,\\
&\mathrm{G}_3=-7.2+0.0213T,\\
&\mathrm{G}_4=-8.5+0.0227T,\\
&\mathrm{G}_5=-8.4+0.0224T,\\
&\mathrm{G}_6=-7.8+0.0210T,\\
&\mathrm{G}_7=-8.2+0.0222T,\\
&\mathrm{G}_8=-10.6+0.0272T,\\
&\mathrm{G}_9=-9.8+0.0244T,\\
&\mathrm{G}_{10}=-8.0+0.0199T
\end{aligned}\tag{1}$$

where T is the absolute temperature in Kelvin. The free energies given in Eq. (1) have the units of kcal/mole. The s parameters are then given as

$$s_n=\exp(-\mathrm{G}_n/RT)\tag{2}$$

In units of kcal/mole, the quantity RT in Eq. (2) is given by

$$RT=(1.9872T)/1000\tag{3}$$

We can then construct an **S** matrix where the elements represent the base-pair interactions between nearest-neighbor base pairs (where the symbol T below stands for base thymine and is not to be confused with the temperature T used in Eqs. (1)–(3))

$$\mathbf{S}=\begin{array}{c|cccc} & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G}\\ \hline \mathrm{A} & s_1 & s_2 & s_5 & s_6\\ \mathrm{T} & s_3 & s_1 & s_7 & s_4\\ \mathrm{C} & s_4 & s_6 & s_{10} & s_8\\ \mathrm{G} & s_1 & s_5 & s_9 & s_{10}\end{array}\tag{4}$$

A general element in the matrix **S** gives the s value for the base at a general site i (row index) followed by the base at site $i+1$ (column index); the complementary bases in the antiparallel chain are understood. Thus, for example, the TC matrix element stands for the following pair interaction, i.e. TC in one chain and AG in the antiparallel chain

$$\mathrm{TC}=\mathrm{TC}/\mathrm{AG}=5'\text{-TC-}3'/3'\text{-AG-}5'\tag{5}$$

We note from Eq. (1) that there are only 10 independent s values as given by Eq. (2) due to the symmetry of the double helix.

Here we will use as an example the entire genome of the virus *Rickettsia prowazekii* Madrid E as given by Andersson et al. [1] and available on the web [2]. This genome is 1 111 523 base pairs long and has the following fractional composition:

$$f_{\mathrm{A}}=0.353743; \quad f_{\mathrm{T}}=0.356254; \qquad f_{\mathrm{C}}=0.143796; \quad f_{\mathrm{G}}=0.146207 \tag{6}$$

Given the known genome, one can tabulate the number of occurrences of each of the base-pair interactions as given in the matrix of Eq. (4). Then using the free energies given in Eq. (1), one can calculate the temperature at which the total helix free energy is zero (relative to the completely unwound chains). This temperature is found to be

$$T_{\mathrm{m}}=364.217\ \mathrm{K} \tag{7}$$

and represents the temperature at which the average s value in the molecule is 1. It is also approximately the melting temperature of the double helix. Thus, at this temperature $s>1$ favors helix and $s<1$ favors coil. Throughout this paper we will use the T_{m} given in Eq. (7) as the value of the temperature. At T_{m} the elements of the **S** matrix of Eq. (4) have the following values:

$$\mathbf{S}=\begin{array}{c|cccc} & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \\ \hline \mathrm{A} & 0.77 & 0.73 & 1.40 & 1.23 \\ \mathrm{T} & 0.46 & 0.77 & 0.17 & 1.38 \\ \mathrm{C} & 1.38 & 1.23 & 2.83 & 2.61 \\ \mathrm{G} & 1.17 & 1.40 & 3.53 & 2.83 \end{array} \tag{8}$$

One sees that, in general, interactions involving A and T favor coil, while interactions involving C and G favor helix.

2. Block-*s* profiles

Using the matrix of s values (evaluated at $T=T_{\mathrm{m}}$ of Eq. (7)) for all possible nearest-neighbor base pairs given in the matrix of Eq. (8), one can then use the known sequence of *Rickettsia* and calculate the average s values in non-overlapping blocks of m base pairs as a function of the position of the given block in the chain, thus producing a profile of average s as a function of chain position. We let the index i denote the position of a base pair in the chain and the index j denote the number of the block containing m base pairs (where both i and j are measured starting from the left-hand end of the chain). Thus, for blocks containing $m=10$ base pairs we have $i=1$–10 for the first block ($j=1$), $i=11$–20 for the second block ($j=2$) and so on. Formally, the block index j is given as a function of base-pair locus i for blocks containing m base pairs as follows

$$j=\left[1+\frac{(i-1)}{m}\right] \tag{9}$$

where the square bracket indicates that one takes the integer part of the quantity enclosed. Given a block of m base pairs that begins at base pair i in the chain, one first calculates the product of m consecutive base-pair interactions and then takes the $1/m$ root, as shown below, to give the geometric mean value of s for the block j, which we will refer to as S_j

$$S_j=\left(\prod_{i=\gamma+1}^{\gamma+m} s(i,\ i+1)\right)^{1/m} \tag{10}$$

where

$$\gamma=m(j-1) \tag{11}$$

We recall that j is the number of the block starting from the left end of the molecule with $j=1$. Note that when we get to the end of a block, the nearest-neighbor of the last unit is simply the next unit in the chain, i.e. the first base pair of the next block.

From Eq. (2) one sees that the factors s are related to the Gibbs free energy for base-pair interactions. Thus, $\ln S_j$ gives the negative of the arithmetic mean Gibbs free energy (divided by RT) per base pair in the block. The condition $\ln S_j>0$ favors helix, while the condition $\ln S_j<0$

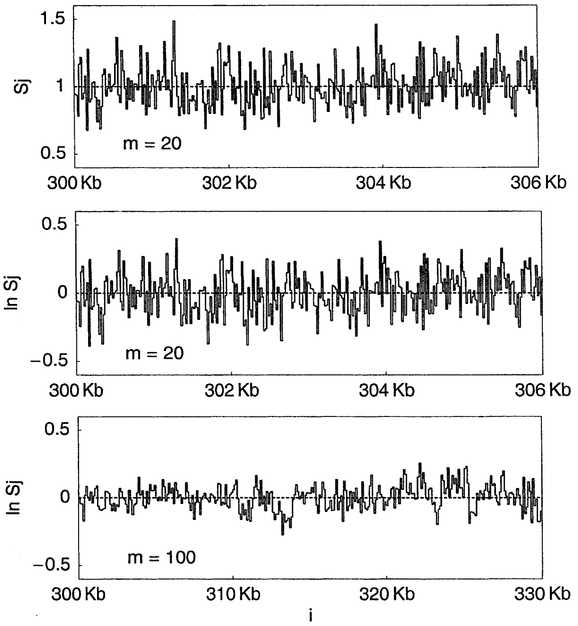


Fig. 1. The average S values for blocks of m base pairs, as defined by Eq. (10), as a function of location in the *Rickettsia* genome. Eq. (9) gives the relation between block number, j, and base-pair number, i, in the genome. The upper curve gives S_j for blocks of 20 base pairs and covers base pairs 300 000–306 000 in the sequence. The middle curve shows $\ln S_j$ for the same m value and sequence. The lower curve shows $\ln S_j$ for blocks of 100 base pairs and covers base pairs 300 000–330 000 in the sequence. The dashed curve in the upper graph represents the locus of $S=1$, which is the overall average for the entire genome at $T=T_m$ given by Eq. (7). The dashed curves for the other two graphs give the locus of $\ln S=0$.

favors coil (just as $S_j>1$ favors helix and $S_j<1$ favors coil). For $S_j\approx 1$ the behavior of profiles with respect to both S_j and $\ln S_j$ are very similar. This is illustrated in Fig. 1 where the top two graphs compare sequences of S_j (top graph) and $\ln S_j$ (middle graph) for blocks of $m=20$ base pairs at T_m. The step-graphs assign the average S_j or $\ln S_j$ to all of the base pairs in an m-unit block. The sequence illustrated starts at base pair number 300 000 in the *Rickettsia* sequence and shows 6 kilobases of sequence. One sees that the two profiles are very similar. The lower graph shows the profile of $\ln S_j$ for $m=100$ at T_m. For this case we illustrate the profile for 30 kilobases of sequence starting, as before, at base pair number 300 000 and treating the same number of blocks as was the case for $m=20$ (the block size for $m=100$ is five times the block size for $m=20$). Notice that the vertical scale for the middle and the lower graphs is the same thus illustrating the fact that as the block size increases, the fluctuations in average $\ln S$ value decrease. It is precisely how the width of the distribution of average $\ln S$ values decreases that will be one of our main interests in the present work. The dashed line in the upper graph gives the locus of $S=1$, while in the lower two graphs the dashed lines give the locus of $\ln S=0$, which is the average $\ln s$ value at temperature T_m of Eq. (7). Throughout the rest of this paper we will give profiles of $\ln S_j$ that show how the average free energy per base pair in a block of m base pairs varies as a function of the location of the block in the sequence.

In Fig. 2 we show another profile of $\ln S_j$. In this case the value of m is much larger with $m=2000$ and we show the profile for the entire genome (rounding off the size of the genome at the right end to give an integer number of blocks). Notice that in this case the range of the vertical

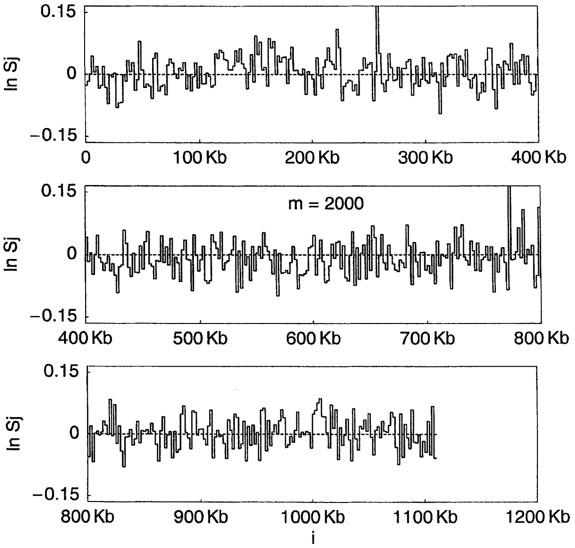


Fig. 2. The average $\ln S_j$ values for the same conditions used in Fig. 1 for blocks of $m=2000$ base pairs as a function of location in the *Rickettsia* genome for the entire genome. The dashed curves give the locus of $\ln S=0$ for reference.

scale is approximately one-third of that used in the lower two graphs in Fig. 1, reflecting the fact that as the block size gets bigger, the range of fluctuation in the average free energy decreases. There are two blocks where the average s value exceeds the scale used in Fig. 2; the average $\ln S_j$ values in these two cases are both approximately equal to 2.

The first impression one has of the free energy profiles shown in Figs. 1 and 2 is that they appear to represent random variations in the free-energy values. We will see that this is both true and false: the profiles for $m=20$ are well represented by a random distribution of base pairs, while the profiles for $m=2000$ have pronounced deviations from random behavior. Thus, there is a transition from random behavior for small m values to nonrandom behavior at large m values and it is this phenomenon that is the focus of the present paper.

3. Block distributions

We now examine data such as shown in Figs. 1 and 2 to see if there is any pattern to the distribution function for the different $\ln S_j$ values shown. To this end we collect the number of $\ln S$ values that occur in a given $\ln S$ range and then plot the distribution function. We will take the grain size for the $\ln S$ values as

$$\Delta \ln S = 0.01 \tag{12}$$

We note that $\ln S$ and $\Delta \ln S$ are dimensionless quantities. The values of $\ln S$ that lie between $(k-1)\Delta \ln S$ and $(k)\Delta \ln S$ will be put in bin-k and this bin will be assigned a mean $\ln S$ value half way between these two bounds

$$\ln S(k) = \left(k - \frac{1}{2}\right)\Delta \ln S \tag{13}$$

The distribution function $P(k)$ gives the probability of a $\ln S$ value being in a given k-bin and is simply proportional to the number of blocks having a value of $\ln S$ in the prescribed range. This probability distribution is normalized according to the following condition (discrete analog of the integral over the distribution function):

$$\sum_k P(k)\Delta \ln S = 1 \tag{14}$$

The quantity $P(k)\Delta \ln S$ is then the probability of $\ln S$ occurring in bin-k.

We then take the data from a profile such as either of those given in the lower two graphs in Fig. 1 and classify each plateau as to which particular bin-k it belongs in. The step-function curve shown in the upper graph in Fig. 3 shows the distribution function obtained in this manner for the case of blocks with $m=20$ such as shown in the middle graph of Fig. 1 (using similar data for the entire genome). One sees that in this case one obtains a very well behaved function for the distribution of helix free energies. We now show that one can fit this function using a small number of moments of the distribution.

We have recently addressed the question of obtaining distribution functions from moments for a variety of different problems concerning biological macromolecules [9–19]. The overall approach utilizes the maximum-entropy method that gives an approximate distribution function based on a small finite number of moments of the distribution. In the present case we can use Eq. (14) to calculate the moments of the distribution function. The following relation gives the nth moment of $\ln S(k)$ distribution:

$$\mu_n = \sum_k \ln S(k)^n P(k)\Delta \ln S \tag{15}$$

In the maximum-entropy method [9] the distribution is given by the following functional form (finite polynomial in x)

$$f(x) = \exp\left[-\sum_{j=0}^{n} \lambda_j x^j\right] \tag{16}$$

Given n moments of the distribution, one can trade these values for the values of the λ_j parameters used in Eq. (16) (this is accomplished by a straightforward nonlinear iteration procedure).

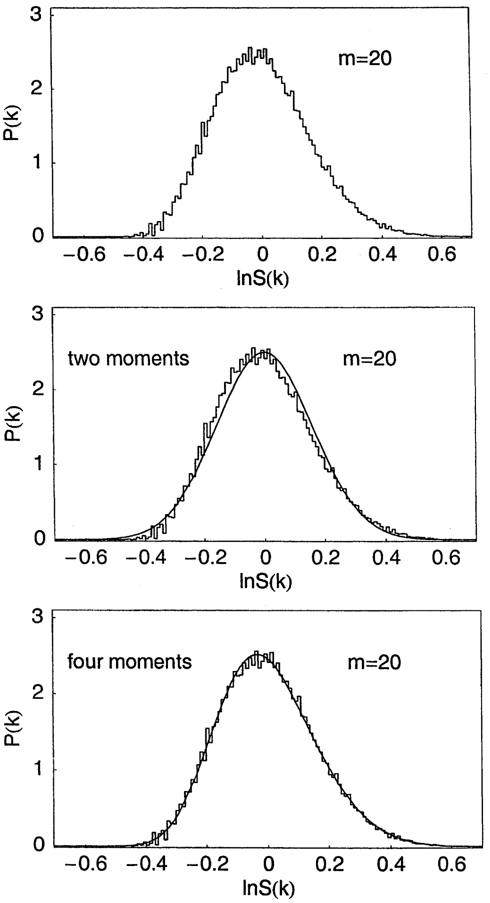


Fig. 3. The upper graph gives the distribution function for the occurrence of average $\ln S$ values for blocks of $m=20$ base pairs, based on data similar to that shown in the middle graph in Fig. 1, for the entire *Rickettsia* genome. The $\ln S$ values are sorted into bins with $\Delta\ln S=0.01$; the k-index gives the number of the bin as defined in Eq. (13). The middle graph repeats the upper graph and compares it with the maximum-entropy approximation (solid curve) to the distribution constructed using two moments. The lower graph is similar to the middle graph, but shows the maximum-entropy distribution constructed using four moments.

In the two graphs in Fig. 3 we show the results of this procedure. The step-function curve gives the actual distribution obtained from the *Rickettsia* genome, while the solid curve in the upper graph (labeled 'two moments') gives the maximum-entropy approximate distribution constructed using two moments of the actual distribution. The solid curve in the lower graph (labeled 'four moments') gives the maximum-entropy approximate distribution constructed using four moments of the actual distribution. For the case where one uses only two moments, the sum in Eq. (16) is a quadratic function and hence the distribution function in this case is essentially equivalent to a Gaussian distribution. The two-moment, or Gaussian, approximation is seen to work quite well in this case, while the use of four moments, as in the lower graph, gives an even better approximation.

As the block size, *m*, is increased, the two-moment (Gaussian) approximation becomes an increasingly good fit for the step-function distribution. We illustrate this point in Fig. 4 where we show the step-function distribution for $\ln S$ for the case of $m=100$ using the entire genome at T_{m}. The solid curve represents the maximum-entropy (or, in this case, Gaussian) distribution based on two moments. Thus, for $m>100$ we will use the maximum-entropy distribution function based on two moments as a good approximation to the actual distribution.

From the examples given above we see that the block-distribution functions of the free energy are well-behaved functions, which, for $m>100$, can be approximated very well by a Gaussian (two-moment) function. The question then arises as to whether these Gaussian functions derived from the actual specific sequence of the *Rickettsia* genome

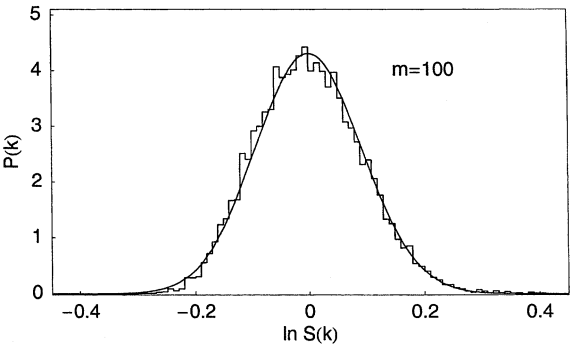


Fig. 4. The distribution function for the occurrence of average $\ln S$ values for blocks of $m=100$ base pairs. The step-graph gives the block-distribution function for the entire genome, while solid curve gives the maximum-entropy approximation based on two moments.

can also be derived from the statistics of the local occurrence of different bases. Of course, the primary determinant of the base sequence in DNA is the genetic information the sequence contains with the resultant thermal stability of the helix being a secondary property. Here we want to see if the thermal stability behaves as if one had a random sequence.

4. Models based on local occurrence statistics

In this section, we will see if we can reproduce the block distributions such as shown in Figs. 3 and 4 from simple models based on the statistics of local base-pair occurrence. First, we use the net base fractions given in Eq. (6) to construct the singlet distribution, which depends only on the type of base at a given location and not on the state of neighboring bases:

Rickettsia singlet frequencies

A	T	C	G
0.3537	0.3563	0.1438	0.1462

(17)

Next, we count the number of overlapping doublets (this means, for example, that the sequence abcd would yield the doublets ab, bc and cd). The *Rickettsia* genome has the following nearest-neighbor frequencies:

Rickettsia doublet frequencies

	A	T	C	G
A	0.1314	0.1234	0.0439	0.0551
T	0.1230	0.0138	0.0465	0.0530
C	0.0522	0.0541	0.0213	0.0162
G	0.0472	0.0450	0.0321	0.0220

(18)

We can check immediately if these are significantly different from the frequencies that one would obtain if the placement of bases were completely random, in which case one would have (independent units)

$$f_{ij}=f_i f_j \tag{19}$$

For example, the fraction of TC doublets would be $f_T f_C=(0.3563)(0.1438)=0.0512$ (taking the singlet frequencies from the table in Eq. (17)), which is to be compared with the actual frequency given in Eq. (18) of 0.0465. A table of nearest-neighbor frequencies based on random placement using Eqs. (17) and (19) above is given below:

Random doublet frequencies

	A	T	C	G
A	0.1251	0.1260	0.0509	0.0517
T	0.1260	0.1269	0.0512	0.0521
C	0.0509	0.0512	0.0207	0.0210
G	0.0517	0.0521	0.0210	0.0214

(20)

One notes that while there are some small differences between the numbers given in Eqs. (18) and (20), the *Rickettsia* doublet frequencies in general are given quite accurately as the product of the singlet frequencies. One notes that the table based on random pairs is symmetric, while that for the actual doublet frequencies is not.

The quantity of interest for constructing distribution functions is the matrix of conditional probabilities. The general form for the doublet conditional probability is given below:

$$P(i|j)=\text{Probability that given } i,\ j \text{ follows} \tag{21}$$

and is constructed from the table of doublet frequencies given in Eq. (16) as follows

$$P(i|j)=f_{ij}/f_i \tag{22}$$

If the doublet frequencies are random (given by Eq. (19)), then one has

$$P(i|j)=f_{ij}/f_i=f_i f_j/f_i=f_j \tag{23}$$

Using the data given in Eqs. (17) and (18) we obtain the following matrix of doublet conditional probabilities for *Rickettsia*

$$P_D=(P(i|j))=\begin{array}{c|cccc} \overset{j}{i} & A & T & C & G \\ \hline A & 0.3714 & 0.3489 & 0.1240 & 0.1557 \\ T & 0.3452 & 0.3755 & 0.1306 & 0.1487 \\ C & 0.3630 & 0.3760 & 0.1485 & 0.1125 \\ G & 0.3226 & 0.3078 & 0.2194 & 0.1502 \end{array} \tag{24}$$

We can then go on to consider overlapping triplets. Of course, the genetic code is in terms of non-overlapping triplets, but we are considering triplets here simply as the next step in describing local correlations. The triplet conditional probabilities are given in general as follows

$$P(i\,j|k) = \text{Probability that given } i\,j,\ k \text{ follows} \tag{25}$$

To determine this probability we count overlapping triplets (this means, for example, that the sequence abcde would yield the triplets abc, bcd and cde) and construct the quantities

$$P(i\,j|k) = f_{ijk}/f_{ij} \tag{26}$$

where f_{ijk} is the fraction of a particular triplet of bases. If the occurrence of bases is random, then one has

$$P(i\,j|k) = f_i f_j f_k / f_i f_j = f_k \tag{27}$$

The table of triplet conditional probabilities obtained from the *Rickettsia* sequence is given below:

$P_T = (P(ij|k)) =$ (28)

i	j \ k	A	T	C	G
A	A	0.3888	0.3450	0.1122	0.1540
A	T	0.3534	0.3684	0.1434	0.1348
A	C	0.3244	0.3733	0.1869	0.1154
A	G	0.3234	0.3055	0.2222	0.1489
T	A	0.3541	0.3536	0.1425	0.1498
T	T	0.3280	0.3937	0.1275	0.1508
T	C	0.3734	0.3757	0.1492	0.1075
T	G	0.3313	0.2739	0.2444	0.1505
C	A	0.3789	0.3158	0.1308	0.1745
C	T	0.3319	0.3727	0.1298	0.1656
C	C	0.3592	0.3756	0.1164	0.1488
C	G	0.2879	0.3165	0.1990	0.1966
G	A	0.3597	0.3844	0.1010	0.1549
G	T	0.3897	0.3437	0.1059	0.1607
G	C	0.4034	0.3804	0.1161	0.1001
G	G	0.3254	0.3891	0.1668	0.1187

If the base occurrence was random, then all of the entries under the column index A would be $f_A = 0.3537$ as given by the singlet distribution of Eq. (17), and so on. One sees that from the point of view of the occurrence of triplets, the distribution is not far from random.

We will now generate random, but specific, sequences that have, successively, the singlet, doublet and triplet distributions characteristic of *Rickettsia*. For simplicity we replace the base designators A, T, C and G with the numbers 1, 2, 3 and 4. We start with the singlet distribution where we designate the net frequency of occurrence of each type as f_1, f_2, f_3 and f_4 (with $f_1 + f_2 + f_3 + f_4 = 1$). We then define the four numbers

$$L_1 = f_1,\ L_2 = L_1 + f_2,\ L_3 = L_2 + f_3,$$
$$L_4 = L_3 + f_4 = 1 \tag{29}$$

We let k indicate the type of unit ($k = 1$–4). To obtain the value of k we pick a random number, R_n, between 0 and 1. The type of unit is then determined as follows:

$$\begin{array}{ll} \text{if } R_n \leqslant L_1 & \text{then } k = 1 \\ \text{if } R_n \geqslant L_1 \text{ and } \leqslant L_2 & \text{then } k = 2 \\ \text{if } R_n \geqslant L_2 \text{ and } \leqslant L_3 & \text{then } k = 3 \\ \text{if } R_n \geqslant L_3 \text{ and } \leqslant L_4 & \text{then } k = 4 \end{array} \tag{30}$$

When this process is repeated, say a million times, one will generate a specific sequence of k values that has, statistically, the overall composition given by the appropriate singlet fs.

In order to generate a random but specific sequence that has a given set of doublet conditional probabilities, one proceeds in the same fashion. One begins at the left end of the sequence and picks the first unit according to the given singlet probabilities for the molecule as in Eq. (30). Then one defines a set of frequencies for the next unit as follows

$$f(k)=P(j|k) \tag{31}$$

Notice that in this case the $f(k)$ values depend on the preceding unit-j and are not constant as was the case for singlets. One then picks the value of k as given in Eqs. (29) and (30). For the case of triplet conditional probabilities, it is done the same way. The first two units in the chain on the left are picked, respectively, according to the given singlet and doublet probabilities. One then defines the following set of frequencies for the next unit as follows

$$f(k)=P(i\,j|k) \tag{32}$$

The values of $f(k)$ in this case now depend on the states of the two preceding units in the chain. One then uses Eqs. (29) and (30) to pick the type of unit for k. In this manner one can generate specific-sequence chains that are generated randomly but have a specified singlet, doublet or triplet distribution.

We now use our tables for singlet, doublet and triplet probabilities obtained from the *Rickettsia* genome to construct $\ln S$ distributions and observe how the block distributions so obtained compare with the actual distributions. In Fig. 5, we show the distribution function for blocks of 20 base pairs obtained from specific sequences generated with the specific singlet, doublet and triplet distributions given above. In each case the specific sequence was 1 111 520 base pairs long (giving 55 576 blocks of $m=20$). To obtain the distribution functions for these sequences, one uses the same procedure employed for the actual sequence with $\ln S$ boxes of $\Delta\ln S=0.01$. In Fig. 5 the smooth solid curve in each graph is the maximum-entropy distribution obtained using four moments for $m=20$ blocks for the actual *Rickettsia* sequence as shown in the graphs in Fig. 3. One sees that all three local distributions, singlet, doublet and triplet, give essentially identical results for the distribution of $\ln S$ in blocks of $m=20$, and all agree very well with the actual distribution of $\ln S$ found in *Rickettsia*.

The results shown in Fig. 5 suggest that the distribution of double helix free energy in m-blocks is described well by the local statistics of occurrence of singlets, doublets or triplets with the simplest, the singlet distribution, giving as good a representation of the *Rickettsia* 20-block distribution as the others. Since big blocks represent an average over the behavior of constituent smaller blocks, one would expect that the trend one sees in Fig. 5 (good representation of the $\ln S$ distributions by simple local distributions) would, if anything, only improve as the block size increases. It turns out that this is not the case and that as m

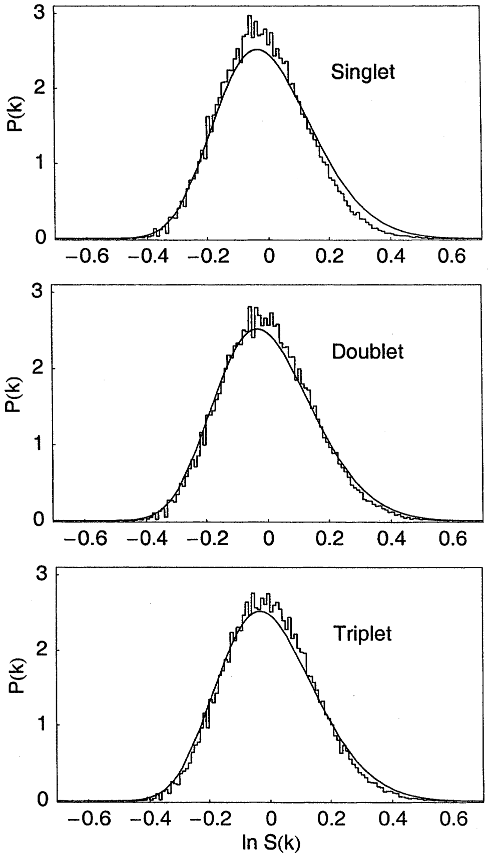


Fig. 5. Block-distribution functions of $\ln S$ for $m=20$ base pairs for specific sequences of 1 111 520 base pairs generated to satisfy the specific singlet, doublet or triplet statistics for *Rickettsia* as indicated in Eqs. (17), (24) and (28). The generation of the specific sequences is outlined in Eqs. (29) and (30). The continuous curve in each graph gives the four-moment maximum-entropy approximation to the actual block-distribution function for *Rickettsia* as shown in the lower graph in Fig. 3.

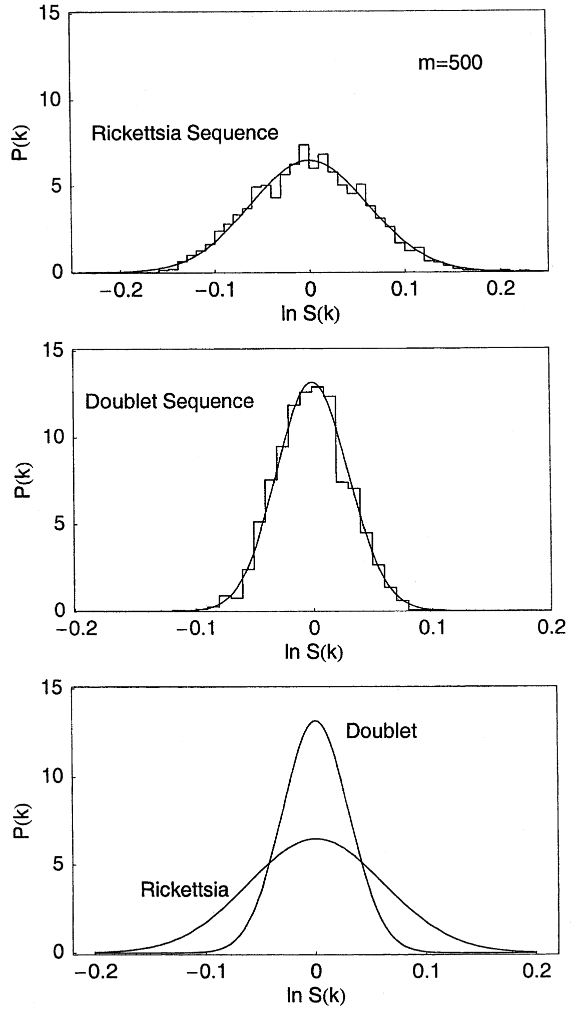


Fig. 6. Distribution functions of ln S for $m=500$. The upper curve gives the actual step-graph for the *Rickettsia* genome; the solid curve is the two-moment maximum-entropy approximation. The middle curve gives the step-graph for a specific sequence of 1 111 500 generated to give the *Rickettsia* doublet frequencies; the solid curve is the two-moment maximum-entropy approximation. The lower graph compares the two-moment maximum-entropy distributions for the actual *Rickettsia* distribution and the specific-sequence doublet distribution.

increases the difference between the ln S distributions obtained from local statistics and those actually found in *Rickettsia* increase in a dramatic fashion.

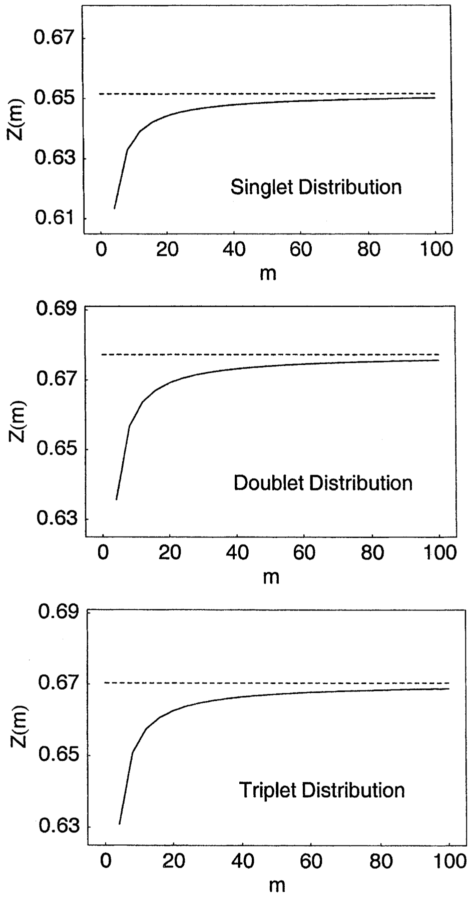


Fig. 7. A plot of the quantity $Z(m)=\sigma_m\sqrt{m}$, where σ_m is the standard deviation defined in Eq. (52) as a function of m for the singlet, doublet and triplet probability distributions constructed as outlined in the text. The dashed curves represent the limiting value for large m obtained from the largest eigenvalue of the appropriate matrix as given in Eq. (57).

We illustrate this phenomenon in Fig. 6. The upper graph in Fig. 6 shows the $m=500$ block bin-distribution function for ln S as obtained from the actual *Rickettsia* genome. The smooth solid curve is the two-moment (Gaussian) fit to the distribution. The middle graph shows the ln S distribution for a specific sequence of 1 111 500 (to give an integer number of $m=500$ blocks) base pairs generated according to Eqs. (30) and (31) for the doublet distribution of *Rickettsia*. Again, the smooth solid curve is the two-moment (Gaussian) fit to the distribution. Note that all of the graphs in Fig. 6 are drawn on the same scale. Finally, in the lower graph in Fig. 6, we compare the Gaussian curves for the distribution function actually found in *Rickettsia* and the distribution function based on the doublet distribution in *Rickettsia*. Unlike the case for $m=20$ where the actual *Rickettsia* distribution and the singlet, doublet and triplet distributions all were in good agreement, here, for $m=500$, there is an enormous difference between the actual distribution and the doublet distribution. The result that is clear in Fig. 6 is as follows: the actual ln S distribution (or free energy distribution) for blocks of $m=500$ is very much broader than the distribution based on local statistics (doublets). Thus, there must be a tendency for weak helix formers and strong helix formers to cluster (like with like) on the scale of $m=500$, which is way beyond the range dictated by the local statistics of base-pair occurrence. In other words, there must be a correlation in helix strength on the scale of $m=500$.

Clearly, the feature of interest in Fig. 6 is the great difference in the widths of the two distributions shown. In order to understand this phenomenon better we need to have a general method to generate the widths of distribution functions and we turn to this task in the next section. The results shown in Fig. 6 are for one specific sequence with the indicated local (doublet) statistics. If one generates another such sequence, one would obtain slightly different results. One can generate the average over all possible specific sequences with given local statistics using a matrix product and we outline this procedure in the next section.

5. Matrix generation of specific-sequence averages

First, we give the overall form of the correlation matrices required. As before we use the following notation: A = 1, T = 2, C = 3 and G = 4. Then, the general matrix, $\mathbf{M}_D$, for treating doublet correlations (states $i=1$–4 in the chain followed by states $j=1$–4) is given below:

$$\mathbf{M}_{\mathrm{D}} = \begin{array}{c|cccc} j & 1 & 2 & 3 & 4 \\ i & & & & \\ \hline 1 & 11 & 12 & 13 & 14 \\ 2 & 21 & 22 & 23 & 24 \\ 3 & 31 & 32 & 33 & 34 \\ 4 & 41 & 42 & 43 & 44 \end{array} \tag{33}$$

For the case of triplet correlations we require the following 16×16 matrix, $\mathbf{M}_T$, giving all possible

$$\mathbf{M}_{\mathrm{T}} = \begin{array}{cc|cccccccccccccccc} & k & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \\ i & j & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \\ \hline 1 & 1 & 111 & 112 & 113 & 114 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 121 & 122 & 123 & 124 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 131 & 132 & 133 & 134 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 141 & 142 & 143 & 144 \\ 2 & 1 & 211 & 212 & 213 & 214 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 & 0 & 0 & 221 & 222 & 223 & 224 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 231 & 232 & 233 & 234 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 241 & 242 & 243 & 244 \\ 3 & 1 & 311 & 312 & 313 & 314 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 2 & 0 & 0 & 0 & 0 & 321 & 322 & 323 & 324 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 331 & 332 & 333 & 334 & 0 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 341 & 342 & 343 & 344 \\ 4 & 1 & 411 & 412 & 413 & 414 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 2 & 0 & 0 & 0 & 0 & 421 & 422 & 423 & 424 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 431 & 432 & 434 & 434 & 0 & 0 & 0 & 0 \\ 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 441 & 442 & 443 & 444 \end{array} \tag{34}$$

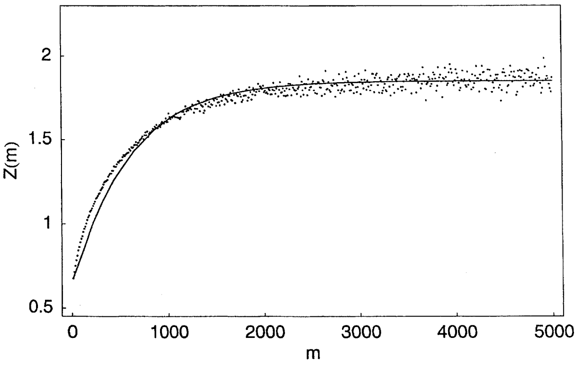


Fig. 8. A plot of the quantity $Z(m) = \sigma_m \sqrt{m}$, where σ_m is the standard deviation defined in Eq. (52) as a function of m for the *Rickettsia* genome. Each point represents an m value; 500 values of m are shown in steps of 10. The solid curve represents the empirical fit of the data given in Eq. (58).

triplets of consecutive states i, j and k in the chain:

To calculate the moments of the ln S distribution for m-blocks that satisfy a set of doublet frequencies, one requires a vector and a matrix. The elements of the vector are simply the singlet probabilities of the four bases

$$\mathbf{p} = (p_i) \tag{35}$$

The matrix has the general structure of Eq. (33) with the specific matrix elements indicated below:

$$\mathbf{P}_\mathrm{D} = \left(P(i|j)\exp[\alpha q_{ij}]\right) \quad \text{where } q_{ij} = \ln\, s_{ij} \tag{36}$$

and α is dummy parameter that will enable us to generate moments. We choose the definition of q_{ij} given in Eq. (36), since on taking the derivative of the quantity $\mathbf{P}_\mathrm{D}$ with respect to α a factor of ln s_{ij} will be brought down and we have chosen to deal with the distribution of ln s. The following matrix product generates all possible sequences of base pairs for a block m units long with the proper singlet and doublet probabilities assigned

$$\Gamma_m(\alpha) = \mathbf{p}\mathbf{P}_\mathrm{D}^m\mathbf{v} \tag{37}$$

where $\mathbf{v}$ is a column 4-vector with all of the elements equal to 1. We note that the matrix $\mathbf{P}_\mathrm{D}$ is raised to the mth power in Eq. (37) since from Eq. (10) the average S for a block includes the interaction between the last base pair in the block and the first base pair of the next block.

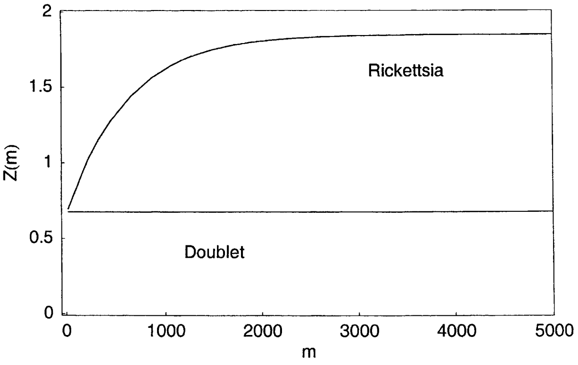


Fig. 9. A plot of the quantity $Z(m) = \sigma_m \sqrt{m}$, where σ_m is the standard deviation defined in Eq. (52) as a function of m. The curve labeled '*Rickettsia*' is the empirical function of Eq. (58) and represents the data from the *Rickettsia* genome. The curve labeled 'Doublet' is the limiting value for the doublet distribution as shown in the middle graph in Fig. 7.

The matrix product in Eq. (37) has the following general form:

$$\Gamma_m(\alpha) = \sum_\nu \prod P(\nu)\exp\left[\alpha\sum f(\nu)\right] \tag{38}$$

where ν indicates a general index for a specific sequence in the m-block; Γ is then a sum over all such specific sequences. If one takes the following derivatives of Γ with respect to α, one generates the moments of the distribution given below (hence Γ acts as a moment generating function)

$$\left(\frac{\mathrm{d}\Gamma(\alpha)}{\mathrm{d}\alpha}\right)_{\alpha=0} = \sum_\nu \left(\sum f(\nu)\right)\prod(P(\nu))$$

$$= \left\langle\sum q\right\rangle = \mu'_1 = m\langle\ln S\rangle$$

$$\left(\frac{\mathrm{d}^2\Gamma(\alpha)}{\mathrm{d}\alpha^2}\right)_{\alpha=0} = \sum_\nu \left(\left(\sum f(\nu)\right)^2\prod P(\nu)\right)$$

$$= \left\langle\left(\sum q\right)^2\right\rangle = \mu''_2 = m^2\left\langle(\ln S)^2\right\rangle \tag{39}$$

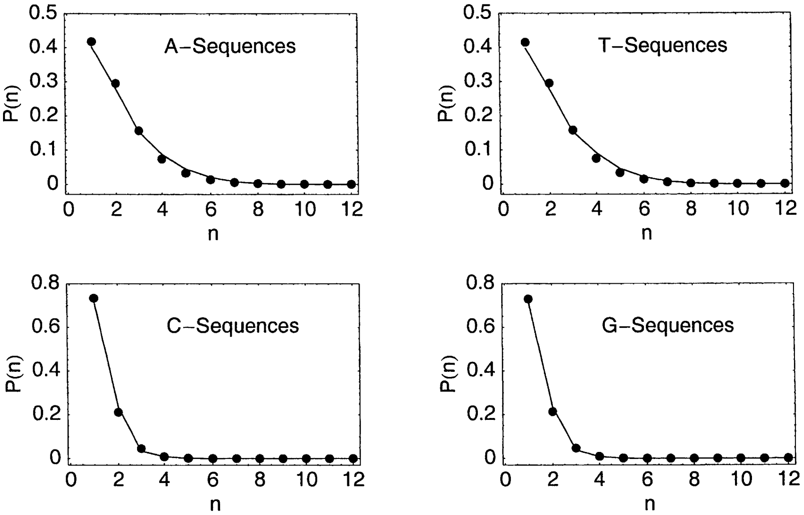


Fig. 10. The probability, $P(n)$, that a base is in a sequence of like bases n units long. The solid dots are plots of $P(n)$ given in Eq. (67) based on random occurrence of bases using the net fractions of each base from *Rickettsia* as given in Eq. (6). The solid line gives the actual $P(n)$ distribution found in *Rickettsia*.

We designate these moments with primes since they are related to the moments of the $\ln S$ distribution but are not the same; we will give the relation between the above moments and the moments of the $\ln S$ distribution shortly. The matrix product in Eq. (37) then generates all possible sequences of an m-block with the prescribed doublet frequencies. The moments given by Eq. (39) are the analog of the moments given in Eq. (15) based on the empirical block distribution for given m. We note that the quantities S and $\ln S$ are defined as the mean values per base pair for the block and hence one has the m factors in the expressions in Eq. (39) involving the averages of $\ln S$.

For the case of the singlet probability distribution the quantity s still depends on the nearest-neighbor composition and we require the structure of the doublet matrices just outlined for that case. To treat the singlet probability distribution one uses the following form (Eq. (23)) for the conditional probabilities in Eq. (35):

$$P(i|j)=p(j) \tag{40}$$

To satisfy the distribution of triplet frequencies, we require the following quantities, all of which are based on the form of the matrix given in Eq. (34). The analog of the vector $\mathbf{p}$ given in Eq. (35) is a vector with 16 elements composed with the general structure of the i-column of $\mathbf{M}_{\mathrm{T}}$ in Eq. (34). Thus, the first four elements are the probability p_1, the next four are p_2 and so on. The analog of $\mathbf{P}_{\mathrm{D}}$ given in Eq. (36) has the structure of $\mathbf{M}_{\mathrm{T}}$, but utilizes only the indices i and j with the general matrix element given by Eq. (35). Finally, we require a matrix that uses all of the information contained in $\mathbf{M}_{\mathrm{T}}$, the general matrix element of which is given below

$$\mathbf{P}_{\mathrm{T}}=\big(P(i\,j|k)\exp[\alpha q_{jk}]\big) \tag{41}$$

The moments of the $\ln S$ distribution for an m-block are then given by the analog of Eq. (37), which for triplets is

$$\Gamma_m(\alpha)=\mathbf{p}\mathbf{P}_{\mathrm{D}}\mathbf{P}_{\mathrm{T}}^{m-1}\mathbf{v} \tag{42}$$

where, in this case, $\mathbf{v}$ is a column 16-vector with all of the elements equal to 1.

Given the first two moments of the distributions, one can then construct the two-moment or Gaussian approximation such as shown in the lower graph of Fig. 6. Thus, one can obtain this distribution function from the matrix products given above without having to generate specific sequences and one has the moments as a general function of the block size *m*.

In order to obtain the first two moments of a distribution according to Eq. (39), one must take the derivatives with respect to α of a matrix product (for singlets or doublets this is the matrix product given in Eq. (37), while for triplets this is the matrix product given in Eq. (42)). We note that the task of taking the derivative of a matrix product is made simpler by utilizing the properties of Toeplitz matrices [20,21]. In this approach one does not first take the matrix product (generating very complex expressions for the matrix elements) and then take the derivative, but rather the other way around. For example, if one wants to take the derivative of a matrix product $\mathbf{W}(x)^m$ with respect to a general variable *x*, one first takes the derivative of **W** with respect to *x* (which we designate as $\mathbf{W}'$) and then evaluates this matrix for a given numerical value of *x*. Thus, all of the matrix elements of **W** and $\mathbf{W}'$ are numbers and not functions. We next construct the following Toeplitz hypermatrix (matrix of matrices) where 0 is the null matrix having the same size as **W**

$$\mathbf{A}=\begin{pmatrix}\mathbf{W} & \mathbf{W}'\\ \\ \mathit{0} & \mathbf{W}\end{pmatrix} \tag{43}$$

We then set $\mathbf{B}=\mathbf{A}^m$ and finally, letting $\Gamma=\mathbf{W}^m$, we have

$$\mathrm{d}\Gamma/\mathrm{d}x=\mathbf{B}[1,2] \tag{44}$$

that is, the required derivative of the matrix product Γ is simply the (1,2) element of the hypermatrix **B** (this element itself is a matrix) and is obtained by raising the matrix **A** to the *m*th power. Since the elements in the matrices **W** and $\mathbf{W}'$ are simply numbers (one has set *x* equal to a number), the matrix multiplication does not require the multiplication of functions and the storage of the complex results. This approach represents an easy way to take the derivatives of complex matrix products. Higher-order derivatives of Γ are obtained in a similar manner.

One can also calculate the asymptotic forms for the moments in the limit of large *m*. In that case Γ is given in terms of the largest eigenvalue, λ_1, of the appropriate **W** matrix as follows

$$\Gamma(\alpha)=\lambda_1(\alpha)^m \tag{45}$$

For $\alpha=0$, Γ is simply the sum of the probabilities of all sequences and is equal to 1. Thus, one has

$$\lambda_1(0)=1 \tag{46}$$

The derivatives required in Eq. (39) are then given as follows

$$\begin{aligned}\mathrm{d}\Gamma/\mathrm{d}\alpha&=m\lambda_1^{m-1}\lambda_1'\\ \mathrm{d}^2\Gamma/\mathrm{d}\alpha^2&=m(m-1)\lambda_1^{m-2}(\lambda_1')^2+m\lambda_1^{m-1}\lambda''_1\end{aligned} \tag{47}$$

where

$$\lambda_1'=\mathrm{d}\lambda_1/\mathrm{d}\alpha,\ \lambda_1''=\mathrm{d}^2\lambda_1/\mathrm{d}\alpha^2 \tag{48}$$

Evaluated at $\alpha=0$, the moments are given by the following relations (using Eq. (46))

$$\begin{aligned}\mu_1'&=m\lambda_1'\\ \mu_2''&=m(m-1)(\lambda_1')^2+m\lambda_1''\end{aligned} \tag{49}$$

Thus, to obtain the first two moments of the distribution (and hence the Gaussian distribution function) one need only obtain the first two derivatives of the largest eigenvalue, λ_l, evaluated at $\alpha=0$.

In order to obtain the derivatives of λ_1 one need not have an explicit equation for the eigenvalues. Rather one can evaluate the derivatives simply using the secular equation of the appropriate matrix directly. Thus, we form the following determinant (where **I** is the identity matrix the same size as **W**):

$$|\mathbf{W}-\lambda\mathbf{I}|=0 \tag{50}$$

and on expanding this equation (which can be done by computer) one obtains a polynomial in λ, which we illustrate for the case where $\mathbf{W}$ is a 4×4 matrix

$$a_0(\alpha)+a_1(\alpha)\lambda(\alpha)+a_2(\alpha)\lambda(\alpha)^2 +a_3(\alpha)\lambda(\alpha)^3+a_4(\alpha)\lambda(\alpha)^4=0 \tag{51}$$

From the expansion in Eq. (50) one obtains explicit expressions for the coefficients $a_n(\alpha)$ given in Eq. (51). One then takes the first two derivatives of Eq. (51), explicitly with respect to the $a_n(\alpha)$ coefficients and implicitly with respect to $\lambda(\alpha)$. One next sets $\alpha=0$, and, from Eq. (46), $\lambda_1(0)=1$. Finally, one solves the two equations (the first and second derivatives of Eq. (51)) for λ_1' and λ''_1. All of this can be done simply on the computer (which can take explicit derivatives and assign symbols for implicit derivatives).

Thus, there is no problem in calculating the derivatives of the largest eigenvalue required in Eq. (49) to give the first two moments. Note that one need not solve Eq. (50) (an example of which is Eq. (51)) explicitly for λ in order to obtain the derivatives exactly.

The moments given in Eq. (49) refer to the net averages per block. We want to obtain the average per base pair in a block used in Eq. (10). In that case we want the moments

$$\mu_1=\mu_1'/m \quad \text{and} \quad \mu_2=\mu_2''/m^2 \tag{52}$$

We have seen that the distribution functions for $\ln S$ are accurately given as two-moment or Gaussian distributions. A convenient measure of the width of a Gaussian distribution is the standard deviation as given below in terms of the moments given in Eq. (49)

$$\sigma_m'=\sqrt{\mu_2''-(\mu_1')^2}=\sqrt{m\left[\lambda_l''-(\lambda_l')^2\right]} \tag{53}$$

and we obtain the familiar result that the width of the distribution varies as the square root of the size of the system (which is m, the size of the block).

If we use the moments in the 'per unit' system given in Eq. (52), then we obtain the result

$$\sigma_m=\sqrt{\lambda_l''-(\lambda_l')^2}/\sqrt{m} \tag{54}$$

indicating that in this system the width of the distribution as measured by the standard deviation gets narrower as the inverse of the square root of the block size. It is useful to define a quantity that is constant as a function of m. From Eq. (54) this is obtained by the following function

$$Z(m)=\sigma_m\sqrt{m} \tag{55}$$

The above argument was based on using the largest eigenvalue of the appropriate matrix. Thus, $Z(m)$ as given in Eq. (55) will be a constant, C, in the asymptotic sense as m becomes large

$$Z(m)=\sigma_m\sqrt{m}\sim C \tag{56}$$

We can now use the methods outlined above to calculate $Z(m)$ as a function of m for the singlet, doublet and triplet distributions characteristic of the *Rickettsia* genome. The results are shown in Fig. 7 where $Z(m)$ is given for each distribution for $m=4$–100. The dashed line in each graph gives the asymptotic value obtained from the largest eigenvalue as illustrated in Eq. (56). This constant has the following values, respectively, for the singlet, doublet and triplet distributions

$$C=0.652,\ 0.677 \text{ and } 0.670 \tag{57}$$

which are very close in value to one another. Notice that in Fig. 7 the value of $Z(m)$ has already settled down to the asymptotic values given in Eq. (57) at approximately the point $m=100$. In addition, note that the total range of variation in $Z(m)$ for the range of m shown is very small. Thus, for the doublet and triplet distributions $Z(4)$ is approximately 0.63, while $Z(100)$ is approximately 0.67, indicating little variation.

We now return to the actual distributions for the *Rickettsia* sequence and show the actual variation of the quantity $Z(m)$ defined in Eq. (56). This quantity is plotted as a function of m in Fig. 8 where we give all values of m in multiples of 10 up to $m=5000$ (in all, 500 points are plotted). One sees that there is some scatter, but a clear pattern emerges. As we expect from our discussion above, the quantity $Z(m)$ does level off to a constant value, but the range of variation of the results shown in Fig. 8 for the actual *Rickettsia* sequence is very much greater than that for the singlet, doublet and triplet distribution as shown in Fig. 7.

The solid curve shown in Fig. 8 is an empirical fit to the data using the very simple form

$$Z(m)=c\Big(1-b\exp\big[-m/m^*\big]\Big) \tag{58}$$

The constants b and c are as follows where c is the limiting value of $Z(m)$ at large m

$$b=0.645 \text{ and } c=1.85 \tag{59}$$

The other constant, m^*, is the most interesting of the three parameters in Eq. (58) and has the value

$$m^*=606 \text{ base pairs} \tag{60}$$

Note that in Eq. (58) the constant m^* has the significance of being the characteristic block size for the relaxation of $Z(m)$ to its asymptotic value. One sees in Fig. 8 that at this value of $m=m^*$ the curve $Z(m)$ is approximately half way to its limiting value.

We can now combine the empirical function of Eq. (58), which represents the actual data of the *Rickettsia* sequence shown in Fig. 8, with that of the local distributions shown in Fig. 7. Since the results for the local distributions are quite similar, we pick the case of the doublet distribution for comparison. On the scale of Fig. 8, the doublet curve is simply a straight line with $Z(m)=0.677$ (of Eq. (57)). This equation and Eq. (58) are shown in Fig. 9. In terms of this figure we can understand our previous results. In Fig. 5 we found that the *Rickettsia* distribution for $m=20$ agreed about equally well with the singlet, doublet and triplet local distributions. This is confirmed in Fig. 9 since for $m=20$ the width function $Z(m)$ agrees for the *Rickettsia* and doublet distributions. On the other hand, for $m=500$, as illustrated in Fig. 6, the *Rickettsia* and doublet distributions are very different. Ultimately, for large m, the $Z(m)$ functions in Fig. 9 level off for both the doublet and the *Rickettsia* distributions. The ratio of the limiting values is then given by (using the data in Eqs. (59) and (57))

$$\frac{Z(\textit{Rickettsia})}{Z(\text{doublet})}=\frac{\sigma(\textit{Rickettsia})}{\sigma(\text{doublet})}=\frac{1.85}{0.677}=2.73 \tag{61}$$

Thus, the actual *Rickettsia* distributions are approximately three times broader than the doublet distribution for m greater than approximately $m=2000$. This difference in the widths of the distribution functions is thus no minor effect involving a few percentage points but rather a major feature of the free energy distributions in this genome. Note that in Eq. (61) the ratio of the Zs equals the ratio of the σs since the square root of m in Eq. (56) cancels on taking the ratios.

What the graph in Fig. 9 shows is that the free energy distribution in m-blocks is very much broader than expected on the basis of local statistics, which means that there are long-range correlations in the distributions for large m. One usually associates the effects of correlation, or cooperation, in biomacromolecules with the sharpening of transitions such as the melting transition. In the present case the correlations make the extremes of the distribution more probable giving rise to a broader distribution.

A reason for this broadening that immediately comes to mind is the gene structure of the DNA. In *Rickettsia* there are 1 111 523 base pairs and 834 protein-coding genes [1,2]. Thus, the average number of base pairs per protein is

$$\frac{1\,111\,523 \text{ base pairs}}{834 \text{ proteins}}=1333 \text{ base pairs/protein} \tag{62}$$

and the average number of amino acids per protein is 444. Now the number given in Eq. (62) has the following relation to the characteristic number m^* given in Eq. (60)

$$1333\approx 2m^* \tag{63}$$

This number is thus seen to be approximately the

m value where $Z(m)$ in the curve given in Fig. 9 levels off to the asymptotic value. Thus, there are significant correlation effects on the scale of the number of base pairs per gene. One can imagine these correlation effects arising due to the different amino acid compositions of different proteins, for example, the difference between the compositions of hydrophobic and hydrophilic proteins, with this difference reflected in the base composition for the DNA sequence over the gene length.

6. Distributions of same-base sequences

Another way to examine the apparent randomness of the distribution of the bases in DNA is to examine the probability of sequences of different lengths containing the same base. Thus, one counts the number of A singlets, doublets, triplets and so on. In general, using the actual *Rickettsia* sequence, we count the number of sequences of a given base type that contains n bases of the specified type. If we let $N(n)$ be the number of n-sequences for a given base type, then the probability that a sequence picked at random from the total set will have n-bases is simply given by

$$p(n)=N(n)/\sum_n N(n) \tag{64}$$

Another distribution of interest is the probability that a base picked at random from the total set of bases of a given type is in a sequence n-units long. This probability is given by

$$P(n)=nN(n)/\sum_n nN(n) \tag{65}$$

If the occurrence of base types in the molecule is random with fraction f for a given base type, then the two probabilities given above are proportional to the following quantities

$$p(n)\sim f^n \text{ and } P(n)\sim nf^n \tag{66}$$

Using the forms given above, the two normalized probability distributions are then

$$p(n)=\left(\frac{1-f}{f}\right)f^n \text{ and } P(n)=\left(\frac{(1-f)^2}{f}\right)nf^n \tag{67}$$

In Fig. 10 we show the distribution $P(n)$, the probability that a base is in a sequence of n units for the actual *Rickettsia* sequence and for a random distribution having the fs characteristic of *Rickettsia* (given in Eq. (6)). The solid curves show the data for the actual *Rickettsia* distribution, while the solid dots give the results of using $P(n)$ from Eq. (67). One notes that with respect to this test, the random singlet model gives the same-base sequence distributions very accurately. Thus, just as we found that the statistics for the free energy distribution for block sizes of the order of $m=20$ (Fig. 5) are accurately given by the random singlet distribution, we find here that the sequence distributions are also very accurately given by Eq. (67), which assume random placement. But again, this is for structure on the order of 10–20 base pairs. From Figs. 6 and 9 we see that on the scale of m greater than approximately 100, the assumption of random distributions fails dramatically. To understand better the nature of this long-range correlation in the DNA helix free energy distribution, we turn to a simpler distribution, namely, the distribution for C or G content that we find parallels the behavior of the free energy distribution.

7. Distributions of C or G content

We have seen that there is a dramatic difference in the distributions of the helix free energy given by random occurrence statistics and the actual *Rickettsia* genome for large block sizes. We now examine a simpler quantity, namely, the distribution of C or G content. The s parameters of Eq. (8) already show that there is a correlation between base type and free energy. Although the free energies given in Eq. (8) do show marked dependence on neighboring base type, we will examine distribution functions for a single variable, the net amount of C or G in a block.

We will use the following notation to indicate the base composition:

$$a=\text{A or T and } c=\text{C or G} \tag{68}$$

As with the free energy distribution, we will examine the *Rickettsia* genome in non-overlapping blocks of m base pairs and record the number of cs in each block. From Eq. (6) we have the following net fractions for *Rickettsia*

$$f_a = 0.7100 \text{ and } f_c = 0.2900 \tag{69}$$

with

$$f_a + f_c = 1 \tag{70}$$

If the states 'a' and 'c' are placed at random, then all of the possible compositions of a block of m units are produced by the following generating function

$$\Gamma_m(z) = (f_a + zf_c)^m = \sum_{k=0}^{m} a_k z^k \tag{71}$$

The parameter z is simply a dummy variable that is inserted in Eq. (71), in this case, to count the number of c states; in the final result the numerical value of z is set equal to 1 and from Eq. (70) one has the result that $\Gamma_m(z=1)=1$ for all m. The probability of having k c-states is simply given by

$$P(k) = a_k \tag{72}$$

where a_k is the coefficient of the z^k term in Eq. (71). This relation gives the singlet, or random unit, distribution for the probability of finding k c-states in an m-block.

We next turn to the doublet distribution for the two states given in Eq. (68). The doublet probability distribution gives the nearest-neighbor pair counts in the actual *Rickettsia* distribution with respect to the states a and c. From the data given in Eq. (18) one has the following result

$$\begin{aligned} f_{aa} &= 0.5116, \; f_{ac} = 0.1984 \\ f_{ca} &= 0.1984, \; f_{cc} = 0.0916 \end{aligned} \tag{73}$$

The doublet conditional probabilities are then given by Eq. (22) using Eqs. (69) and (73). The matrix of conditional probabilities for *Rickettsia* is then

$$\mathbf{P}_{\mathrm{D}} = \begin{pmatrix} P(a|a) & zP(a|c) \\ P(c|a) & zP(c|c) \end{pmatrix} = \begin{pmatrix} 0.7205 & 0.2795z \\ 0.6843 & 0.3157z \end{pmatrix} \tag{74}$$

while the generating function for the doublet distribution is given by

$$\Gamma_m(z) = \mathbf{p}\mathbf{P}_{\mathrm{D}}^{m-1}\mathbf{v} = \sum_{k=0}^{m} a_k z^k \tag{75}$$

where $\mathbf{p}$ and $\mathbf{v}$ are the following vectors

$$\mathbf{p} = (f_a, \; zf_c), \quad \mathbf{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{76}$$

We note that the matrix $\mathbf{P}_{\mathrm{D}}$ is raised to the $(m-1)$ power in Eq. (75), since unlike the case treated in Eq. (37), there is no interaction included between successive blocks. The generating function given in Eq. (75) is again a finite polynomial in powers of the dummy variable z and the distribution giving the probability of finding k c-units in an m-block subject to the doublet frequencies given in Eq. (73) is then given by Eq. (72).

In Fig. 11 we compare the c-unit distribution for blocks of $m=20$, 100 and 500 units. In each case the solid smooth lines represent the distributions for random distributions as described above. In each case the curve that is slightly higher at the maximum value is for the singlet distribution (Eq. (71)), while the other is for the doublet distribution (Eq. (75)). One sees that there is not much difference between the distributions obtained using the singlet and doublet frequencies characteristic of *Rickettsia* as given in Eqs. (69) and (73). The solid dots (joined by solid line segments) are the probability distributions based on the actual

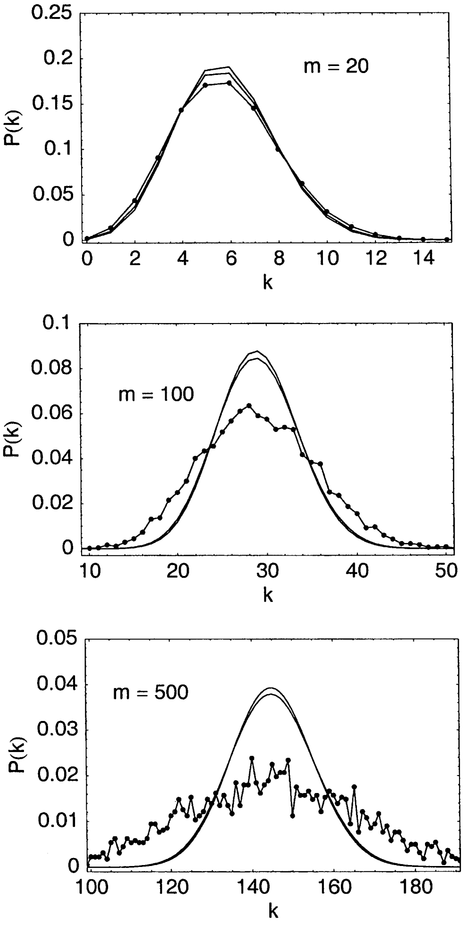


Fig. 11. The probability distribution, $P(k)$, that non-overlapping blocks of m bases in *Rickettsia* have k bases that are C or G. The smooth solid lines are the distributions based on singlet and doublet distributions calculated using the generating functions of Eq. (71) and Eq. (75), respectively, for $m=$ 20, 100 and 500. In each graph the upper smooth curve at the maximum is for the singlet distribution. The solid dots joined by solid lines give the distributions actually found in the *Rickettsia* sequence.

sequence of *Rickettsia* obtained from non-overlapping m-blocks. One sees that for $m=20$ there is little difference between the actual and random distributions. For $m=100$ the actual distribution is now noticeably broader than either of the random distributions and, finally, for $m=500$ the actual distribution is very much broader than the random distributions. This is exactly the pattern of behavior that we saw for the free energy distributions, where, as in Fig. 5, there is little difference between the distribution for $m=20$ and the random distributions and then, for the case of $m=500$ illustrated in Fig. 6, the actual free energy distribution is very much broader than the random distributions.

In Fig. 12, we illustrate the fact that for large m the actual distribution is broader than the random distributions in another way. We treat the case of

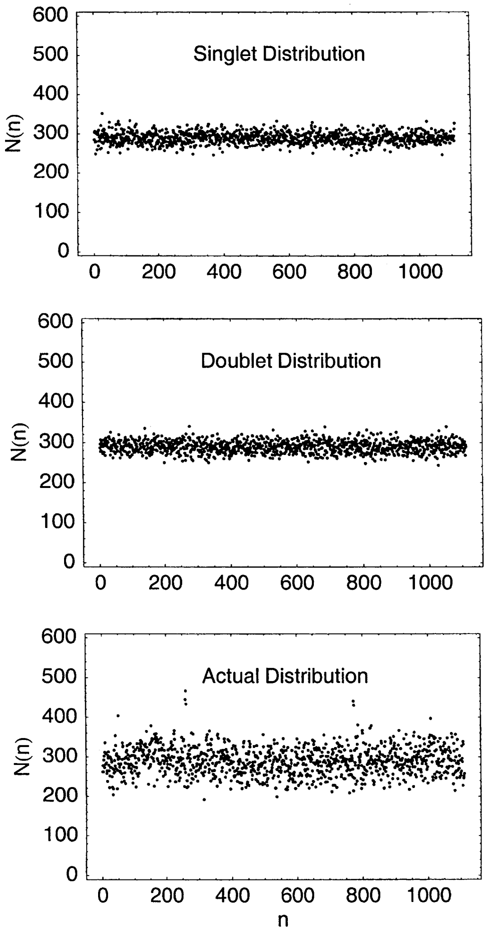


Fig. 12. The number, $N(n)$, of bases that are C or G as a function of the number of the block (starting the counting from the left end of the molecule) in non-overlapping blocks of $m=$ 1000 bases. The upper two graphs are for specific sequences generated to have the singlet or doublet distributions of *Rickettsia*, while the lower graph is based on the actual *Rickettsia* sequence.

$m=1000$ and label the successive blocks, starting from the left end with the index n. So $n=1$ is the first block, $n=2$ is the second and so on. As a function of n we then plot on the vertical axis the number of c-units, $N(n)$, in that block number-n; thus there is one number per block plotted. The upper curve is for a specific sequence generated to have the singlet distribution of *Rickettsia* (given in Eq. (60)), while the middle graph is for the case of a specific sequence generated to have the doublet distribution of *Rickettsia* (given in Eq. (73)). One sees that the singlet and doublet distributions so obtained are very similar. The lower graph is for the actual distribution in *Rickettsia* and here one finds that the spread of points is very much greater than that found in either the singlet or doublet distributions.

We can generalize this information about the width of the actual distributions following the approach used with respect to Fig. 8. First we note that in this case the moments and hence the standard deviations of the distribution are given as derivatives of the generating function Γ_m with respect to z

$$\mu_1=(\mathrm{d}\Gamma/\mathrm{d}z)_{z=1} \quad \text{and} \quad \mu_2=(\mathrm{d}/\mathrm{d}z\ \ (\mathrm{d}\Gamma/\mathrm{d}\ln z))_{z=1} \tag{77}$$

For the case of the singlet generating function of Eq. (71) we have

$$\mu_1=mf_c \quad \text{and} \quad \mu_2=m(m-1)f_c^2+mf_c \tag{78}$$

giving for the standard deviation

$$\sigma(m)=\sqrt{\mu_2-\mu_1^2}=\sqrt{m}\left(\sqrt{f_c-f_c^2}\right) \tag{79}$$

We note that we obtain the standard result for a random distribution that the width of the distribution varies as the square root of the sample size. As with the free energy distribution it is convenient to have a function that is asymptotic to a constant. Thus, we define the following quantity (using f_c of Eq. (69))

$$\zeta(m)=\sigma(m)/\sqrt{m} \tag{80}$$

For the singlet distribution result given in Eq. (79), we have the constant value

$$\zeta(m)=\sqrt{f_c-f_c^2}=0.4538 \tag{81}$$

We note the difference between this quantity and the quantity $Z(m)$ defined in Eq. (55) for the free energy distributions. The difference in these two quantities with respect to the placing of the square root of m arises because the free energy distribution deals with the average free energy per unit in the block.

We plot the quantity $\zeta(m)$ defined above in Fig. 13. The irregular solid line gives the results for m-blocks in the actual *Rickettsia* genome, while the dashed line is the constant value characteristic of the singlet distribution given in Eq. (81). One sees that the behavior shown in Fig. 13 is very similar to that shown in Fig. 8 for the width of the free energy distributions. For small m (say, $m=20$) the widths of the actual and random distributions are very similar. For large m (say, greater than $m=100$) the actual distributions have a much greater width with the quantity $\zeta(m)$ approaching an asymptotic value. As was the case for the curve for the free energy distributions, we can fit a simple empirical function to the data. Our result is

$$\zeta(m)=1.30(1-0.643\exp[-m/557]) \tag{82}$$

which is very similar to the function of $Z(m)$ given in Eq. (58). In particular, the characteristic relaxation parameters, m^*, have similar values with $m^*=606$ for the free energy distributions and $m^*=557$ for the C or G distributions given here. Recall that m^* is the value of m where the function is well on the way to its asymptotic value. The ratio of the asymptotic values of the curves shown in Fig. 13 is given by

$$\frac{\zeta(Rickettsia)}{\zeta(\text{singlet})}=\frac{\sigma(Rickettsia)}{\sigma(\text{singlet})}=\frac{1.30}{0.454}=2.86 \tag{83}$$

where we note again that on taking the ratio of the $\zeta(m)$s, the square root of m in Eq. (80) cancels.

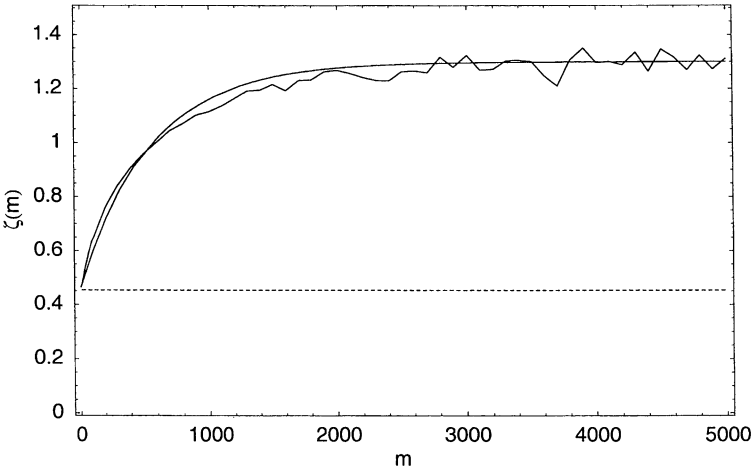


Fig. 13. The width, as described by the function $\zeta(m)$ given in Eq. (80), of the distribution for the number of bases that are C or G in blocks of m non-overlapping blocks in *Rickettsia*. The solid jagged line gives the actual results found for the *Rickettsia* sequence, while the dashed line gives the expected result based on the singlet distribution as given by Eq. (81). The smooth solid line is the empirical fit of the *Rickettsia* data as given by Eq. (82).

The number above is almost the same as the corresponding ratio for the free energy distributions, 2.73, as given in Eq. (61).

8. Long-range correlation tables

We have seen in Fig. 5 that the free energy distribution for blocks of $m=20$ is given quite accurately using local occurrence statistics (singlet, doublet and triplet distributions). On the contrary, when m is made large, say $m=500$, as illustrated in Fig. 6, the actual distribution is much wider than the distribution based on local occurrence statistics. Fig. 11 shows that the same behavior holds for the distribution of C or G content. This is further illustrated in Fig. 12, which shows that for $m=1000$ the actual distribution is again much broader that that given by using singlet and doublet distributions based on *Rickettsia* statistics. Finally, in Fig. 13 we compare the width function $\zeta(m)$ of Eq. (80) for the *Rickettsia* genome with the value for the singlet distribution and again obtain a very large difference between the results obtained using local statistics and actual sequence for m greater than approximately $m=100$.

Since the free energy distribution, illustrated in Fig. 8, and the C or G content distribution, illustrated in Fig. 13, show essentially the same behavior, we will use the behavior of the $c=$C or G content to try and understand the origin of the distribution broadness that we find. It is clear that in order to understand this behavior we must take into account correlations between consecutive m-blocks. This is easy to do for the c content since we can characterize a block simply by the number of c units it contains and this number can vary from $i=0$ to $i=m$, that is, $m+1$ integer values. Note that we are not going to introduce any variables describing the order or arrangement of the Cs and Gs in the block, but just the net number of Cs or Gs.

We thus introduce a correlation table that gives the number of non-overlapping m-blocks with i c-units that are followed by blocks containing j c-units. As an example we consider blocks with $m=20$. The number of such non-overlapping blocks (starting from the left end of the molecule) in the *Rickettsia* genome is given below

$$\text{Number of 20-blocks} = \text{Integer part}(1\,111\,523/20) = 55\,576 \tag{84}$$

The number of consecutive non-overlapping pairs

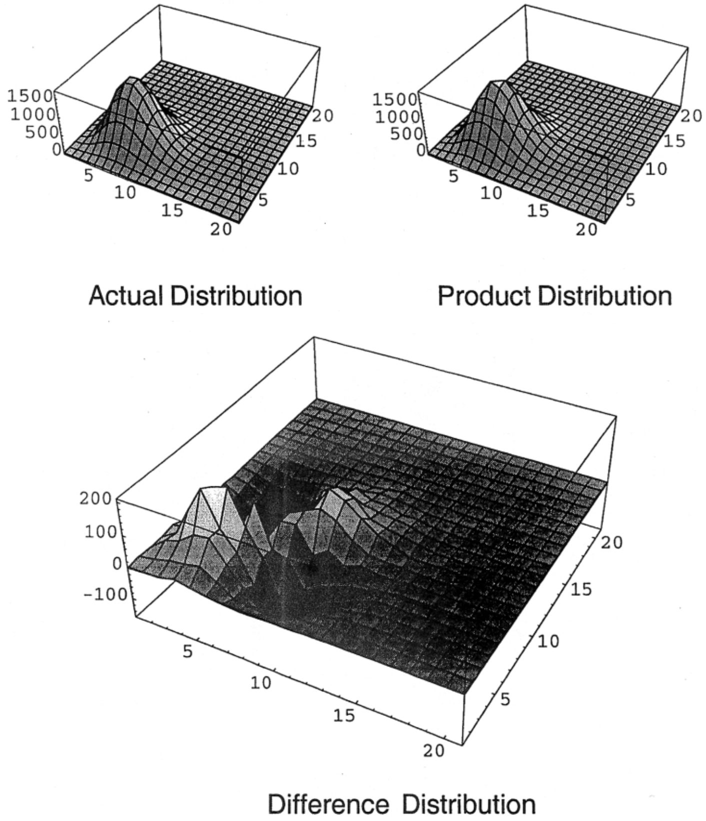


Fig. 14. The distribution, $N_b(i, j)$ of consecutive non-overlapping pairs of sequences for $m=20$ in terms of the number of C or G in each sequence. The upper left-hand graph gives the actual distribution for the *Rickettsia* genome, while the upper right-hand graph gives the distribution assuming random placement, $P(i)P(j)$, as given in Eq. (87). The bottom graph shows the difference in the two distributions, $\Delta N_b(i, j)$ given by Eq. (89), showing that there are distinctly two mountains and two valleys in this difference distribution.

of m-blocks is one less than this or

$$\text{Number of pairs} = 55\,575 \qquad (85)$$

We then set up an $(m+1)(m+1)$ table (m/m for short) that gives the number of (i, j) pairs of blocks, which we will refer to as $N_b(i, j)$. The resulting table is approximately, but not exactly, symmetric, that is, it is not so that $N_b(i, j)=N_b(j, i)$ exactly. For $m=20$ the largest entries in this table for the *Rickettsia* genome are

$$\begin{aligned} N_b(5,\ 5)&=1758 \quad N_b(5,\ 6)=1644 \\ N_b(6,\ 5)&=1634 \quad N_b(6,\ 6)=1727 \end{aligned} \qquad (86)$$

The complete correlation table is shown graphically in the upper-left graph in Fig. 14.

We can compare this with the numbers expected if the m-blocks occurred at random. Then we would have

$$N_b{}^*(i, j)=(55\,575)P(i)P(j) \qquad (87)$$

where $P(i)$ is the random probability that an m-block will contain i c-units as given by Eq. (72). The number in brackets is the number of $m=20$ pairs given in Eq. (85). Note that unlike the actual distribution function, $N_b(i, j)$, the distribution function given in Eq. (87) is symmetric, that is, $N_b{}^*(i, j)=N_b{}^*(j, i)$. Using the $P(i)$ obtained from the

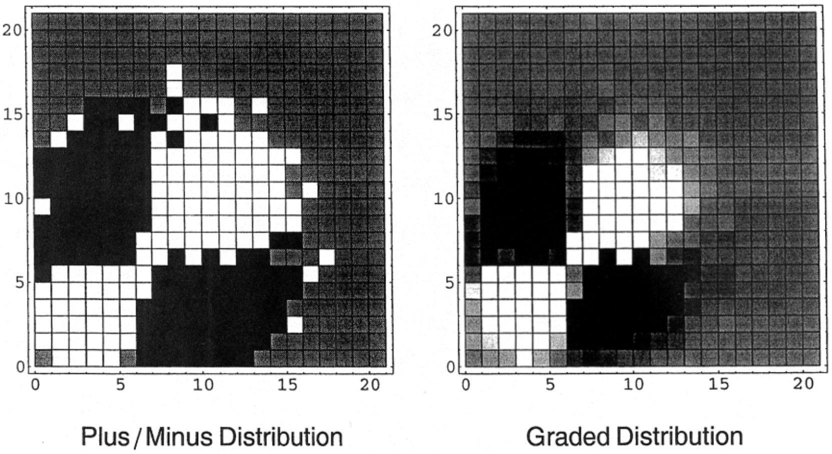


Fig. 15. Illustration of the difference distribution, $\Delta N_b(i, j)$ of Eq. (89), shown in Fig. 14 for $m=20$. In the left-hand graph the white and black areas show where the difference distribution is positive and negative, respectively; in the gray area the difference distribution is 0. The right-hand graph shows a graded distribution from positive to negative.

Rickettsia genome (using Eqs. (69)–(72)), the maximum terms in the N_b* distribution are found to be

$$\begin{aligned} N_b^*(5,\ 5)&=1620 \quad N_b^*(5,\ 6)=1642 \\ N_b^*(6,\ 5)&=1642 \quad N_b^*(6,\ 6)=1665 \end{aligned} \tag{88}$$

These values are not very different from the numbers given in Eq. (86) for the actual distribution in *Rickettsia*. The graph of $N_b*(i, j)$ is given by the upper-right graph in Fig. 14. One notes that the two distributions, $N_b(i, j)$, the actual distribution, and $N_b*(i, j)$, the distribution for independent blocks, are very similar and that the eye cannot make out much difference between the two.

There is, however, a significant difference between these two distributions and this difference will explain the occurrence of the very broad distributions we have been finding in *Rickettsia*. In order to see this, we construct the difference distribution as follows:

$$\Delta N_b(i, j)=N_b(i, j)-N_b*(i, j) \tag{89}$$

Using the table elements shown in Eqs. (86) and (88) as examples, we have

$$\begin{aligned} \Delta N_b(5,\ 5)&=138 \quad \Delta N_b(6,\ 5)=2 \\ \Delta N_b(6,\ 5)&=-8 \quad \Delta N_b(6,\ 6)=62 \end{aligned} \tag{90}$$

We note that the elements in ΔN_b can be positive or negative. The elements having the extreme positive and negative values are

$$\Delta N_b(3,\ 3)=201 \quad \Delta N_b(4,\ 9)\ =-152 \tag{91}$$

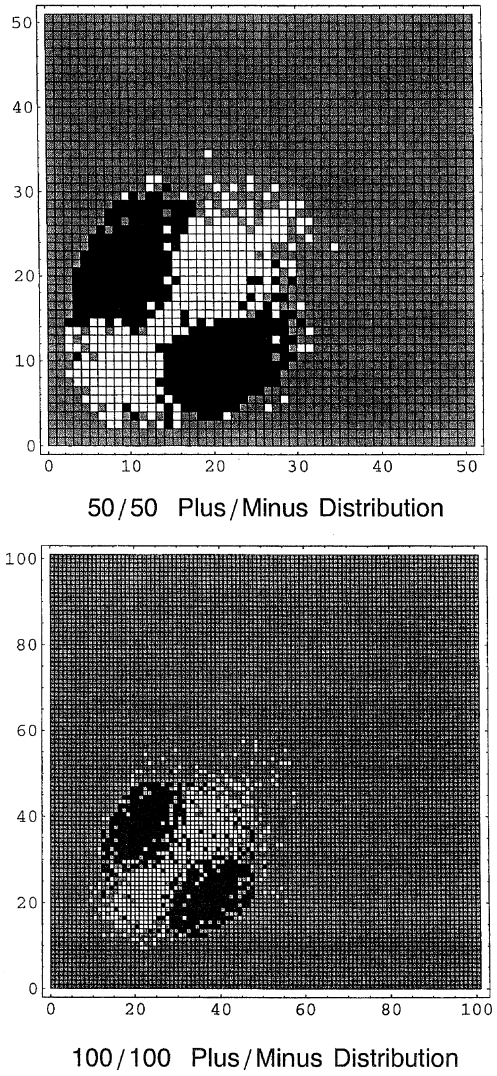
One notes that the extreme values given in Eq. (91) are approximately 10% of the maximum values given in Eq. (86) or Eq. (88). Thus, some pair correlations are larger than the values given by the random placement of blocks, while some are smaller and the magnitude of this effect is approximately a 10% change in correlation, a moderate, but important, amount.

The striking feature of the function $\Delta N_b(i, j)$ is not so much the magnitude of the effect, but the distribution of positive and negative deviations from random behavior. This is seen in the lower graph in Fig. 14 where one sees two distinct summits in the difference distribution and two distinct valleys. This feature is seen clearly in the left-hand graph in Fig. 15 where we plot simply whether ΔN_b is positive (white squares) or ΔN_b is negative (black squares); the gray squares represent no difference. The right-hand graph in Fig. 15 gives the gradual distribution, positive (white) to negative (black). Both of these graphs show the same feature: there is a definite correlation between where the white squares (+) and black squares (−) occur. The positive correlations (white) occur along the axis $i=j$, while the negative correlations (black) occur along an axis perpendicular to that. This means that there is a tendency for an m-block to be followed by an m-block with a similar C or G content. Thus, if one looks, for example, at $m=40$ blocks as made up of pairs of consecutive $m=20$ blocks, there will be a tendency for $m=20$ blocks with a given C or G content to be followed by like blocks. This correlation will make the extreme values of the distribution with respect to c-content more probable and explains the width of the distributions we have been finding. Of course, the reason why there is this correlation is a function of the information content of the genome (e.g. similar amino acids occur in certain proteins). In Appendix A we give an explicit example of distribution broadening caused by block correlation. In Fig. 16 we give the sign of ΔN_b for $m=50$ (top graph) and $m=100$ (bottom graph). One sees that the tendency for blocks with a given C or G content to be followed by like blocks persists to large m values.

9. Block distributions from correlation tables

In Eq. (75) we give the generating function, $\Gamma_m(z)$, for the probability of having k c-units in a block of m units as a matrix product utilizing the matrix of nearest-neighbor conditional probabilities $\mathbf{P}_D$, of Eq. (74). In that case the nearest-neighbors were nearest-neighbor base pairs. We now want to give the analogous result where the states that we are correlating are the states of an entire m-block with m of the order of 20 or larger.

The construction of the generating function in this case is similar to that used to obtain Eq. (75). We have a vector $\mathbf{p}$ whose elements are the fraction of m-blocks having a given c content that can vary from $i=0$ to $i=m+1$. The matrix $\mathbf{P}_D$ now gives the conditional probability that given an m-block containing i c-units it is followed by an m-block containing j c-units. Since both i and j can vary from 0 to $m+1$, $\mathbf{P}_D$ is an $(m+1)(m+1)$ matrix. The analog of $\mathbf{v}$ in Eq. (76) is an $(m+1)$ unit vector, each element of which is 1. If f_{ij} is the fraction of consecutive non-overlapping pairs of blocks each having m units, then the conditional



50/50 Plus/Minus Distribution

100/100 Plus/Minus Distribution

Fig. 16. Illustration of the difference distribution, $\Delta N_b(i, j)$ of Eq. (89), for $m=50$ and 100. As in the left-hand graph in Fig. 15, the white and black areas show, respectively, where the difference distribution is positive and negative, while the gray area shows where the difference distribution is 0.

probability $\mathbf{P}_{\mathrm{D}}(i|j)$ is given in general by Eq. (22), that is $\mathbf{P}_{\mathrm{D}}(i|j)=f_{ij}/f_{i\cdot}$.

We can now use these quantities to construct the generating function for a block that is composed of a general integer number of smaller m-blocks. Thus, the analog of Eq. (75) is

$$\Gamma_{m*n}(z)=\mathbf{p}(m,\ z)\mathbf{P}_{\mathrm{D}}(m,\ z)^{n-1}\mathbf{v}(m)=\sum_{k=0}^{m*n}a_k z^k \tag{92}$$

where $m*n$ is to be read as 'm times n' where n is a positive integer. Thus, the size of the resultant block, $m*n$, is m times n. As with Eq. (75), one can insert z factors to keep track of c-units in $\mathbf{p}$ and $\mathbf{P}_{\mathrm{D}}$. As shown in Eq. (92), the matrix product yields a sum over powers of z where the powers can range from 0 to $m*n$. The coefficient of z^k, the quantity a_k, in Eq. (92), is the probability that an $m*n$ block contains k c-units (in any arrangement).

In Fig. 17 we illustrate this procedure by constructing the block probability distribution for the case of $m*n=500$ blocks based on $\mathbf{P}_{\mathrm{D}}$ successively using $m=50$ ($n=10$) and $m=100$ ($n=5$) correlation tables. In both graphs the jagged solid line represents the actual distribution of $m*n=500$ blocks for the *Rickettsia* genome. The jagged appearance illustrates that for large values of m the actual distribution function need not be smooth. In the upper graph in Fig. 17 the smooth solid curve shows the c-unit distribution function constructed from the $m=50$ correlation table using the generating function given below

$$\Gamma_{500}=\mathbf{p}(50)\mathbf{P}_{\mathrm{D}}(50)^{9}\mathbf{v}(50) \tag{93}$$

The lower graph shows a similar calculation where Γ_{500} is constructed from the $m=100$ correlation table

$$\Gamma_{500}=\mathbf{p}(100)\mathbf{P}_{\mathrm{D}}(100)^{4}\mathbf{v}(100) \tag{94}$$

One sees that in both cases the distribution functions constructed from large-m correlation tables reproduce the actual *Rickettsia* distribution quite well, with the results from the $m=100$ correlation table being especially good.

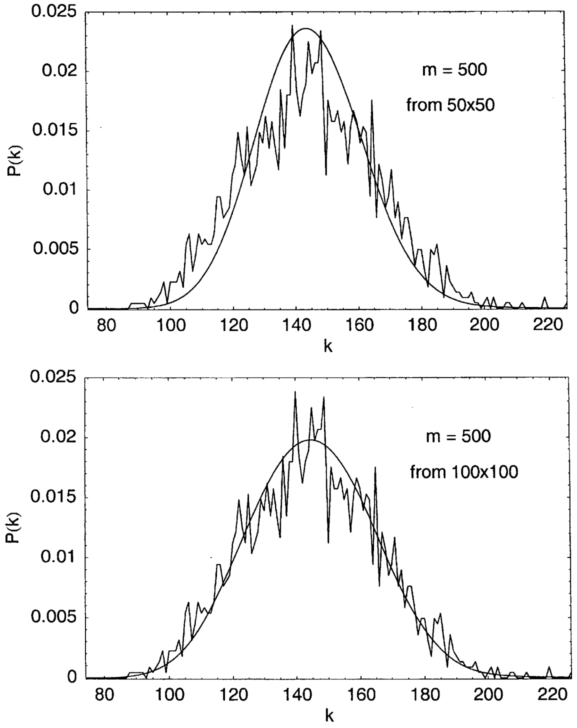


Fig. 17. The distribution of C or G content in sequences with $m=500$. The jagged curve in both graphs gives the actual distribution for the *Rickettsia* genome, while the smooth curve in each graph gives the distribution given by the Markov chain of Eq. (92), using the 50×50 correlation table in the top graph and the 100×100 correlation table in the lower graph.

We next want to see if we can use Eq. (92) to reproduce the behavior shown in Fig. 13. This figure gives the width of the m-block distribution as defined by the function $\zeta(m)$ of Eq. (80). The main feature of this function is that it levels off to an asymptotic value at approximately $m=2000$. We note that one can use Toeplitz matrices as outlined in Eqs. (43) and (44) to calculate the required derivatives of the matrix product contained in Eq. (92). In Fig. 18 we show the quantity $\zeta(m)$ calculated from the $m=50$ correlation table used in Eq. (92). The solid dots give the values of $m*n=50$, 100, etc., that is, integer multiples of 50. In Fig. 18 we also show $\zeta(m)$ calculated using the $m=200$ correlation table in Eq. (92). Here the solid dots give the values of $m*n$ that are integer multiples of 200. The actual $\zeta(m)$ curve

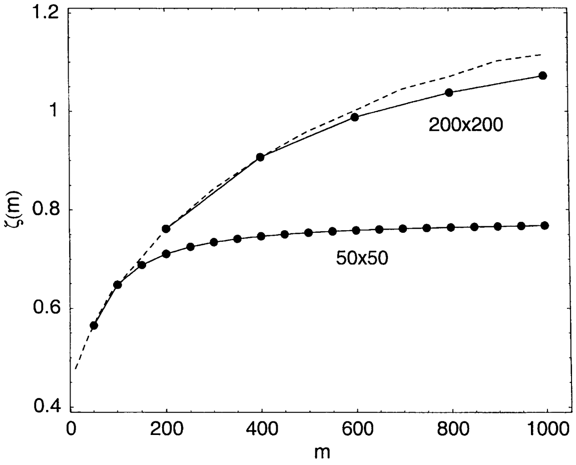


Fig. 18. The calculation of the $\zeta(m)$ function given by Eq. (80) and illustrated in Fig. 13 on the basis of sequence correlation tables (50×50 or 200×200, as labeled). The dashed curve shows part of the actual curve for the *Rickettsia* genome as shown in Fig. 13.

from the *Rickettsia* genome is given by the dashed curve. One sees that the $\zeta(m)$ function calculated from the $m=50$ correlation table levels off at a value well below the actual asymptotic limit given by the dashed curve. On the other hand, $\zeta(m)$ calculated using the $m=200$ correlation function is very close to the actual behavior found for *Rickettsia* (dashed curve) out to $m=1000$. Thus, we see that correlations between blocks on the order of $m=200$ are sufficient to explain the main features (especially the width) of the c-content distribution functions.

Once the system has reached the asymptotic limit, $m\sim1000$, the matrix product in Eq. (92) can be approximated well by using only the maximum eigenvalue, λ_1, of the matrix $\mathbf{P}_\mathrm{D}$. In this case one has

$$\Gamma_{m*n}=\lambda_1(m)^n \quad (m>1000) \tag{95}$$

Once the range of the limit $m>1000$ has been reached, then, if Eq. (95) is a good approximation, we have the following simple product relationship between generating functions

$$\Gamma_{m*n}=\Gamma_m^n \quad (m>1000) \tag{96}$$

where, again, n is a positive integer.

We first illustrate the fact that Eq. (96) does not hold for $m<1000$. In Fig. 19 we approximate the c-content distribution function for $m=1000$ using products of generating functions for smaller values of m. The lower, irregular curve is the actual distribution from the *Rickettsia* genome, somewhat smoothed by local averaging. The other

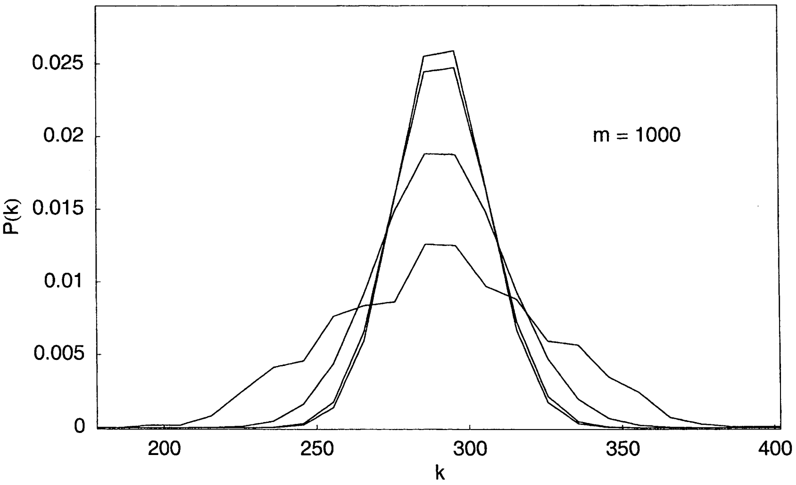


Fig. 19. The calculation of sequence generating functions as products for the case of $m=1000$. The lowest curve gives the actual sequence distribution function for $m=1000$. The upper three curves give, respectively, starting from the top, Γ_1^{1000}, Γ_{10}^{100} and Γ_{100}^{10}.

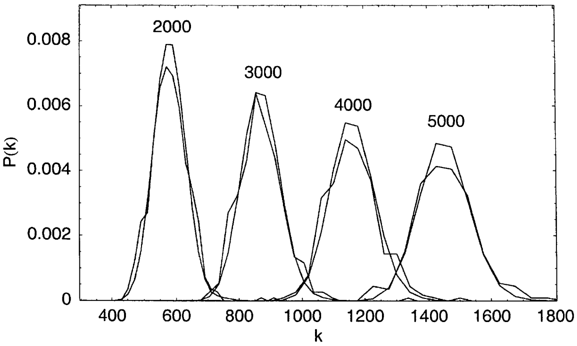


Fig. 20. The calculation of sequence generating functions as products for the cases of $m=2000$, 3000, 4000 and 5000. In each case the generating function was produced from the generating function for $m=1000$ as given in Eq. (98).

three curves represent the following product approximations using Eq. (96)

$$\Gamma_{1000}\approx\Gamma_{1}^{1000},\ \Gamma_{1000}\approx\Gamma_{10}^{100},\ \Gamma_{1000}\approx\Gamma_{100}^{10} \qquad (97)$$

The top three curves plotted in Fig. 19 give the functions listed in Eq. (97), the curves from top to bottom representing, respectively, the functions left to right. One sees that the product approximation gets better as the value of m used in the product generating function increases, but that none of the functions given in Eq. (97) gives a good approximation to the actual c-content distribution.

Finally, we construct the following distribution functions

$$\begin{aligned}&\Gamma_{2000}=\Gamma_{1000}^{2},\ \Gamma_{3000}=\Gamma_{1000}^{3}\\&\Gamma_{4000}=\Gamma_{1000}^{4},\ \Gamma_{5000}=\Gamma_{1000}^{5}\end{aligned} \qquad (98)$$

as illustrated in Fig. 20. In each case the lower curve is the actual c-content distribution based on the *Rickettsia* genome (again smoothed by local averaging). One now sees that for $m>1000$ the product approximation of Eq. (96) gives an excellent representation of the actual distributions, particularly with respect to the width of the distributions.

10. Summary

In this paper we have shown that if one looks at the free energy of the DNA helix of *Rickettsia* in non-overlapping consecutive blocks containing m base pairs, as shown in Figs. 1 and 2, one obtains a relatively smooth distribution. Using $m=20$ as an example, the distribution function obtained can be fit well using the maximum-entropy method using two to four moments as shown in Fig. 13. In addition, still for $m=20$, the distribution can be reproduced using local base occurrence statistics obtained from the *Rickettsia* genome (singlet, doublet or triplet statistics) as illustrated in Fig. 5. On the basis of the evidence for $m=20$, it would seem that there is nothing extraordinary about the free energy distribution: the base sequence contains the genetic information, but with respect to the helix free energy, the base occurrence can be considered to depend only on local statistics.

The picture changes dramatically when one looks at blocks containing hundreds of units as shown in Fig. 6 for the case of $m=500$. In this case the distribution function obtained from the actual *Rickettsia* genome is very much broader (by more than a factor of 2) than obtained using local statistics. This effect is dramatically shown in Fig. 9 where the function $Z(m)=\sigma_m\sqrt{m}$ of Eq. (55) is plotted for distributions obtained from the actual *Rickettsia* sequence and for distributions based on local (doublet) statistics where σ_m is the standard deviation of the appropriate distribution. We have suggested in the discussion surrounding Eq. (62) that the scale of the behavior of $Z(m)$ with respect to m shown in Fig. 9 indicates that the free energy content varies significantly from gene to gene, the average gene size being about the value of m ($\approx$1000 base pairs) at which the function begins to bend over toward its asymptotic value.

To explore this effect further we turned to a simper distribution function, namely, that of the net C or G content in a block of m base pairs. It turns out that the distributions for this variable closely resemble those for the free energy at comparable m values as shown in Fig. 11. The width function for these distributions, $\zeta(m)=\sigma_m/\sqrt{m}$, as shown in Fig. 13, also parallels the corresponding form for the free energy function shown in Fig. 9.

For the case of C or G content there is a single number characterizing the state of an m-block, namely, the number of C or G units (which can vary from 0 to m). Fig. 15 shows the possible states of a 20-block in terms of the states of a following 20-block, giving the difference between the doublet correlations for the actual *Rickettsia* distributions and those based on random occurrence. The white and black squares indicate, respectively, positive and negative differences between the two distributions. The fact that the white and black squares occur in distinctly different regions of the correlation table has enormous significance for the width of the distribution: blocks with a given C or G content tend to be followed by like blocks and vice versa. As illustrated in Appendix A, this effect naturally leads to a broadening of the distribution. Thus, in order to explain the width of the distribution for C or G content (or free energy content), we must consider correlation tables between consecutive m-blocks for m of the order of several hundred base pairs. When we do this, using $m=200$ as an example, then we can reproduce the behavior of the width function of Fig. 13 as shown in Fig. 18.

Appendix A: Correlation broadening

As an example of a case where correlation between successive blocks (sequences with similar c (C or G) content tend to follow one another) causes the distribution to get broader, we consider the following. Take the case where $m=2$, so there are three states per block, $i=0$, 1 and 2 where i is the number of C or G units in an $m=2$ block. We take the quantities f_0, f_1 and f_2 as the probabilities, respectively, of these three states. Next we construct the pair distribution for the case where the occurrence of successive $m=2$ blocks is random. This gives the following distribution for two consecutive $m=2$ blocks, or a single $m=4$ block

$$g=(f_0+f_1+f_2)^2 \tag{A1}$$

On multiplying this expression out, we obtain the random distribution for the five states of an $m=4$ block:

Random distribution

$$\begin{aligned} &P(0)=f_0^2, \qquad P(1)=f_0f_1+f_1f_0 \\ &P(2)=f_0f_2+f_1f_1+f_2f_0 \\ &P(3)=f_1f_2+f_2f_1, \quad P(4)=f_2f_2 \end{aligned} \tag{A2}$$

Now consider the case where the probability of the (0, 0) and (2, 2) states is increased, while the probability of the (0, 2) and (2, 0) states is decreased, where (i, j) indicates state-i of an $m=2$ block followed by j-state of the neighboring $m=2$ block. This case is the analog of the behavior we found in the correlation tables where the probabilities of the diagonal terms (like followed by like) are increased and the probabilities of the off-diagonal terms (unlike states) are decreased. As a specific example of this effect we modify the probabilities given in Eq. (A2) as follows:

Correlation distribution

$$\begin{aligned} &P(0)=\alpha f_0^2, \qquad P(1)=f_0f_1+f_1f_0 \\ &P(2)=\beta f_0f_2+f_1f_1+\beta f_2f_0 \\ &P(3)=f_1f_2+f_2f_1, \quad P(4)=\alpha f_2f_2 \end{aligned} \tag{A3}$$

We take $\alpha>1$ to increase the correlation of diagonal terms and $\beta<1$ to decrease the correlation of off-diagonal terms. To retain normalization of the total distribution we require (for our choice of correlations given in Eq. (A3))

$$\alpha+\beta=2 \tag{A4}$$

As a numerical example we take $f_0=0.3$, $f_1=0.4$ and $f_2=0.3$ as the probabilities for the $m=2$ block. The random distribution based on these numbers as given in Eq. (A2) is shown in the upper graph in Fig. 21. Next we turn to the correlation distribution using the above numbers in addition to the values $\alpha=1.4$ and $\beta=0.6$ in Eq.

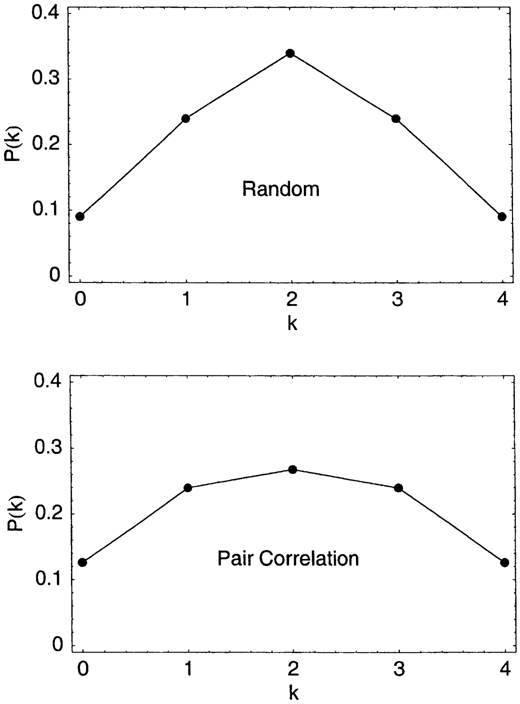


Fig. 21. Illustration of broadening of distributions caused by pair correlations for the case of $m = 4$ given in Appendix A The upper graph gives the random distribution given by Eq. (A2), while the lower graph gives the distribution when correlation is included as given by Eq. (A3).

(A3). On comparing the graphs in Fig. 21, one sees that the effect of correlation (increasing the probability that like follows like) causes a marked broadening of the *c*-content distribution.

References

[1] S.G. Andersson, A. Zomorodipour, J.O. Andersson, et al., The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, Nature 396 (6707) (1998) 133–140.

[2] The annotated genome sequence of *Rickettsia prowazekii* is available on the World Wide Web at www.tigr.org.

[3] J. SantaLucia, A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA 95 (1998) 1460–1465.

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

[5] D. Poland, Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations D. Poland, Biopolymers 13 (1974) 1859–1871.

[6] News release from PubGene Inc. This is available from: www.pubgene.com.

[7] B.H. Zimm, J.K. Bragg, Theory of the phase transition between helix and random coil in polypeptide chains, J. Chem. Phys. 31 (1959) 526–535.

[8] B.H. Zimm, Theory of melting of the helical form in double chains of the DNA type, J. Chem. Phys. 33 (1960) 1349–1356.

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

[10] D. Poland, Ligand binding distributions in biopolymers, J. Chem. Phys. 113 (2000) 4774–4784.

[11] D. Poland, Enthalpy distributions in proteins, Biopolymers 58 (2001) 89–105.

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

[13] D. Poland, Protein-binding polynomials, D. Poland, J. Protein Chem. 20 (2001) 91–97.

[14] D. Poland, Free energy distributions in proteins, Proteins: Structure, Function Genetics 45 (2001) 325–336.

[15] D. Poland, Maximum-entropy determination of self-association distribution functions: daunorubicin and ATP, Biophys. Chem. 94 (2002) 185–199.

[16] D. Poland, Maximum-entropy calculation of free energy distributions in two forms of myoglobin, J. Protein Chem. 21 (2002) 187–194.

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

[18] D. Poland, Protein denaturant binding polynomials, J. Protein Chem. 21 (2002) 477–485.

[19] D. Poland, Free energy of proton binding in proteins, Biopolymers, 69 (2003) 60–71.

[20] M.W. Springgate, D. Poland, Moments of distribution functions for linear systems using Toeplitz matrices, J. Chem. Phys. 62 (1975) 675–679.

[21] M.V. Springgate, D. Poland, Lattice statistics using Toeplitz matrices, J. Chem. Phys. 62 (1975) 680–686.