ELSEVIER

Biophysical Chemistry 104 (2003) 279–289

Biophysical Chemistry

www.elsevier.com/locate/bpc

DNA probability profiles: examples from the *Treponema pallidum* genome

Douglas Poland*

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

Received 1 August 2002; received in revised form 25 November 2002; accepted 25 November 2002

Abstract

In this paper we apply an algorithm developed by Poland (Biopolymers 13 (1974) 1859) to treat the statistical mechanics of the thermal unwinding of DNA to the genome of *Treponema pallidum*, the syphilis spirochete. We calculate probability profiles (giving the probability that each unit in the molecule is in the helix-state) and other statistical distributions for genes and sequences of genes, the longest containing 100 genes and 107 139 base pairs ($\approx$10% of the genome).

Keywords: DNA statistical mechanics; Sequence distribution functions; Thermal unwinding; Probability profiles

1. Introduction

In 1974 we published an algorithm to calculate exactly the statistical mechanics of the thermal unwinding for a model of DNA [1]. The model treated a finite double-helix of arbitrary length and included the effects of a specific sequence of base pairs, the nucleation of helical sequences, and the long-range correction for the entropy of interior loops. We have previously published the application of this method to artificially constructed base sequences (Poland and Scheraga [2,3]; Poland [4]). Since that time a great deal has been learned about the base sequences of the entire genome for many organisms. In the present paper we apply the algorithm to pieces of the genome of *Treponema pallidum*, the syphilis spirochete (Fraser et al. [5]).

We will basically treat the same model we used previously. The algorithm, however, does not depend critically on the details of the model (e.g. it can incorporate any type of long-range interaction in coil sequences). Our model uses nucleation and growth parameters σ and s analogous to the parameters used in the helix–coil treatment of polypeptides by Zimm and Bragg [6]. The s parameters in DNA give the statistical weight of a given base pair at site n followed by another given base pair at site $n+1$. SantaLucia [7] has published a unified set of the required s parameters and we will use his functions here. He gives the following free energies (of helix doublets of base pairs relative to the random coil) where we follow the order used in his Table 2:

*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 Science B.V. All rights reserved.
PII: S0301-4622(02)00382-4

$$\Delta G_1 = \Delta G(\mathrm{AA/TT}) = -7.9 + 22.2T'$$
$$\Delta G_2 = \Delta G(\mathrm{AT/TA}) = -7.2 + 20.4T'$$
$$\Delta G_3 = \Delta G(\mathrm{TA/AT}) = -7.2 + 21.3T'$$
$$\Delta G_4 = \Delta G(\mathrm{CA/GT}) = -8.5 + 22.7T'$$
$$\Delta G_5 = \Delta G(\mathrm{GT/CA}) = -8.4 + 22.4T'$$
$$\Delta G_6 = \Delta G(\mathrm{CT/GA}) = -7.8 + 21.0T'$$
$$\Delta G_7 = \Delta G(\mathrm{GA/CT}) = -8.2 + 22.2T'$$
$$\Delta G_8 = \Delta G(\mathrm{CG/GC}) = -10.6 + 27.2T'$$
$$\Delta G_9 = \Delta G(\mathrm{GC/CG}) = -9.8 + 24.4T'$$
$$\Delta G_{10} = \Delta G(\mathrm{GG/CC}) = -8.0 + 19.9T' \quad (1)$$

where

$$T' = T/1000 \quad (2)$$

which gives the free energies in kcal/mole. The convention used in representing the successive base pairs is illustrated below:

$$\mathrm{AC/TG} = 5' - \mathrm{AC} - 3'/3' - \mathrm{TG} - 5' \quad (3)$$

which we will abbreviate as follows (giving just the pair in the upper chain):

$$\mathrm{AC} = 5' - \mathrm{AC} - 3' \quad (4)$$

There are a total of 16 possible nearest-neighbor doublets of base pairs of which the 10 given in Eq. (1) are unique. The others are given by symmetry:

$$\Delta G(\mathrm{AA/TT}) = \Delta G(\mathrm{TT/AA})$$
$$\Delta G(\mathrm{CA/GT}) = \Delta G(\mathrm{TG/AC})$$
$$\Delta G(\mathrm{GT/CA}) = \Delta G(\mathrm{AC/TG})$$
$$\Delta G(\mathrm{CT/GA}) = \Delta G(\mathrm{AG/TC})$$
$$\Delta G(\mathrm{GA/CT}) = \Delta G(\mathrm{TC/AG})$$
$$\Delta G(\mathrm{GG/CC}) = \Delta G(\mathrm{CC/GG}) \quad (5)$$

The s parameters are then given by

$$s_m = \exp[-\Delta G_m/RT'] \quad (6)$$

where $R = 1.9872$ to give RT' both in kcal/mole.

The table of the 16 possible nearest-neighbor pair interactions using the convention of Eq. (4) is then given by

5′\3′	A	T	C	C
A	s_1	s_2	s_5	s_6
T	s_3	s_1	s_7	s_4
C	s_4	s_6	s_{10}	s_8
G	s_7	s_5	s_9	s_{10}

(7)

At 50 °C the values of the entries are

5′\3′	A	T	C	C
A	3.12	2.59	6.14	4.88
T	1.65	3.12	4.97	6.17
C	6.17	4.88	11.58	16.89
G	4.97	6.14	19.87	11.58

(8)

while at 100 °C one has

5′\3′	A	T	C	C
A	0.60	0.58	1.06	0.95
T	0.37	0.60	0.90	1.05
C	1.05	0.96	2.18	1.85
G	0.90	1.06	2.57	2.18

(9)

The s factors in Table (7) take into account the effects of base pair hydrogen bonds (two for A–T and three for C–G) and stacking between neighboring base pairs.

The parameter σ reflects the fact that it is difficult ($\sigma \ll 1$) to nucleate a helix and is assigned to the border between helix and coil sequences. In nucleic acids the analog of the σ nucleation parameter in polypeptides takes into account both the unfavorable free energy associated with a helix–coil boundary and the loss of the entropy of coil sequences due to the fact that interior unwinding in DNA produces closed loops. We follow

Fisher [8] and take the statistical weight for loops as

$$\sigma_{\text{loop}} = \frac{\sigma_0}{(n+d)^{\gamma}} \tag{10}$$

where

$$\gamma = 1.75 \tag{11}$$

Fisher [8] and Poland and Scheraga [9] have pointed out that the loop factor given in Eq. (10) leads to the presence of a true phase transition (singularity in the partition function) in the limit of infinite chains. It is the loop-effect, requiring knowledge of the size of each loop that makes the statistical mechanics of DNA unwinding much more difficult than that for the helix–coil transition in polypeptides.

Gotoh and Tagashira [10] have fitted the parameters in Eq. (10) to the experimental melting curves of long DNA molecules and given the following values of the parameters that give the best fit:

$$\sigma_0 = 1.2 \times 10^{-5}$$
$$d = 450 \tag{12}$$

We will then use the parameters given in Table (7) and Eq. (12) in our calculations of DNA probability profiles.

2. The algorithm

Our algorithm to treat the thermal unwinding of DNA involves two probability arrays. The general element of the first array is the a priori probability that site m in the chain is in the helix-state. To represent helix and coil states we will use the symbols 1 and 0, respectively. Using this notation, the a priori probability for site m is designated by

$$p(1_m) = p(m) \tag{13}$$

The general element of the second array is the conditional probability that given a unit at site m in the chain in the helix-state it is followed by a helix-state at site $m+1$:

$$P(1_m|1_{m+1}) = P(m) \tag{14}$$

Note the shorthand versions introduced above: lower-case $p(m)$ is the a priori probability while upper-case $P(m)$ is the conditional probability.

The algorithm involves two recursion relations containing these probabilities. The use of recursion relations was developed by Lacombe and Simha [11] who applied the method to the nearest-neighbor Ising model in one dimension. The approach was then used to treat linear chains with long-range interactions by Poland [12] and finally was applied to DNA by Poland [1]. The first recursion relation involves $P(m)$ and starts with $P(N-1)$, where N is the chain length, and works down to $P(1)$. The second recursion relation involves $p(m)$ and starts with $p(1)$ and works up to $p(N)$. This procedure is schematically illustrated below:

$$P(1) \leftarrow \ldots \leftarrow P(N-3) \leftarrow P(N-2) \leftarrow P(N-1)$$
Recursion chain of conditional probabilites

$$p(1) \rightarrow p(2) \rightarrow p(3) \rightarrow \ldots \rightarrow p(N)$$
Recursion chain of a priori probabilities (15)

To use these two recursion relations one must prime each relation, the first with the value of $P(N-1)$ and the second with the value of $p(1)$.

The recursion relations and the primers are obtained by using simple conservation of probability relations. In this manner one obtains the following simple results.

(A) Prime conditional chain:

$$\bar{P}(N-1) = \frac{r(N-1)}{1+r(N)} \tag{16}$$

(B) Conditional chain:

$$\bar{P}(m) = r(m)\left[1 + \sigma \sum_{n=1}^{N-m+1} n^{-\lambda} \prod_{k=m+1}^{m+n} \bar{P}(k) + r(N) \prod_{k=m+1}^{m+n} \bar{P}(k)\right]^{-1}$$
(for $m = N-2$ to 1 in steps of -1) (17)

(C) Prime a priori chain:

$$p(1)=\left[1+\sum_{n=1}^{N-1}\prod_{k=1}^{n}\bar{P}(k)+r(N)\prod_{k=1}^{N-1}\bar{P}(k)\right]^{-1} \tag{18}$$

(D) A priori chain:

$$p(m)=\frac{p(m-1)\bar{P}(m-1)}{r(m-1)}+\sigma\sum_{j=1}^{m-2}p(j)\,(m-j-1)^{-\gamma}\prod_{k=j}^{m-1}\bar{P}(k)+p(1)\prod_{k=1}^{m-1}\bar{P}(k)$$
$$(\text{for } m=2 \text{ to } N \text{ in steps of } +1) \tag{19}$$

where

$$r(m)=1/s(m) \tag{20}$$

and

$$\bar{P}(m)=r(m)P(m) \tag{21}$$

which is a combination of the conditional probability $P(m)$ and the s parameters of Eq. (1). Notice that the only parameters appearing in Eqs. (16)–(19) are the s parameters, σ_0, and γ.

We note that the partition function for this DNA model can also be evaluated as a matrix product [2–4]. However, the size of the square matrix required is $(N+1)\times(N+1)$ where N is the chain length. For short chains one can check that the matrix method and the algorithm of Eqs. (16)–(19) give exactly the same results. Another check on the algorithm is obtained when one sets $\gamma=0$ in which case the model reduces to that for a specific-sequence polypeptide which one can treat by other methods [1]. For more complicated models with loops on loops such as found in β-sheets and t-RNA the matrix formalism becomes even more complex [13].

3. Sequence probabilities

From the application of the algorithm of Eqs. (16)–(19) one obtains the complete arrays $p(m)$ and $P(m)$. The array $p(m)$ gives directly the probability that unit m in the chain is in the helix-state, thus giving the probability profile for the entire molecule. In addition to this quantity, we have shown [1] that given the conditional probabilities, $P(m)$, one can calculate the probabilities of arbitrary sequences of helix and coil and, indeed, the probability of any specific conformation in the molecule. We review these relations below.

The first relation we quote is for the probability of an interior loop starting at unit m and containing n base pairs. This probability is given in terms of the $\bar{P}(m)$ given in Eq. (21):

$$p_1(m,n)=\frac{\sigma}{n^{\gamma}}p(m)\prod_{k=m}^{m+n}\bar{P}(k) \tag{22}$$

The average number of interior loops in the chain having n base pairs is simply the sum over m of the probabilities given in Eq. (22):

$$N_{\mathrm{I}}(n)=\sum_{m=1}^{N-n-1}p_{\mathrm{I}}(m,n) \tag{23}$$

The probability that one has n coil states unwound on the left-end of the molecule is given by

$$p_{\mathrm{L}}(n)=p(1)\prod_{k=1}^{n}\bar{P}(k) \tag{24}$$

while the net average number of coil states resulting from unwinding from the left-end is the sum over n of these probabilities:

$$\langle N_{\mathrm{cL}}\rangle=\sum_{k=1}^{N-1}np_{\mathrm{L}}(n) \tag{25}$$

Similarly, the probability that one has n coil states unwound on the right-end of the molecule is given by

$$p_{\mathrm{R}}(n)=p(N-n)\frac{s(N-n)}{s(N)}\prod_{k=N-n}^{N-1}\bar{P}(k) \tag{26}$$

with the analog of Eq. (25) giving the net amount of unwinding from the right-end:

$$\langle N_{\mathrm{cR}}\rangle = \sum_{n=1}^{N-1} n p_{\mathrm{R}}(n) \tag{27}$$

The average number of helix-states is given simply in terms of the a priori probabilities $p(m)$:

$$\langle N_{\mathrm{h}}\rangle = \sum_{m=1}^{N} p(m) \tag{28}$$

Given this quantity we immediately obtain the average number of coil states from the relation

$$\langle N_{\mathrm{c}}\rangle = N - \langle N_{\mathrm{h}}\rangle \tag{29}$$

The net amount of unwinding from the ends of the double-helix is given by Eqs. (25) and (27). Combining these relations with Eq. (29) we obtain the amount of unwinding resulting from interior loops:

$$\langle N_{\mathrm{cI}}\rangle = \langle N_{\mathrm{c}}\rangle - \langle N_{\mathrm{cL}}\rangle - \langle N_{\mathrm{cR}}\rangle \tag{30}$$

The net probability that unit m is in the c-state as a result of unwinding from the left-end is

$$p_0(m;\mathrm{L}) = \sum_{n=m}^{N-1} p_{\mathrm{L}}(n) \tag{31}$$

while the same quantity for unwinding from the right-end is

$$p_0(m;\mathrm{R}) = \sum_{n=N-m+1}^{N-1} p_{\mathrm{R}}(n) \tag{32}$$

It is of interest to know the probability that a border exists between helix and coil sequences at a given location in the chain. We can calculate the probability that a (1–0) border exists at a specific unit in the chain as follows. From the algorithm of Eqs. (16)–(19) we obtain the conditional probability that given state-1 at unit m it is followed by state-1 at unit $m+1$. Now state-1 at unit m must be followed by either state-1 or state-0 and thus we have

$$P(1_m|0_{m+1}) = 1 - P(1_m|1_{m+1}) = 1 - P(m) \tag{33}$$

We introduce the border probability

$$p(1_m 0_{m+1}) = p_{10}(m) \tag{34}$$

which can also be expressed as

$$p(1_m 0_{m+1}) = p(1_m) P(1_m|0_{m+1}) \tag{35}$$

Combining the above relations we obtain

$$P_{10}(m) = p(m)[1 - P(m)] \tag{36}$$

which gives the probability that at unit m in the chain there is the right-hand boundary between a helix-sequence and a coil-sequence.

To obtain the probability of a (0–1) border, designated as $p_{01}(m) = p(0_{m-1}1_m)$, we can use the procedure just employed for calculating the (1–0) border probability, but with the sequence of the molecule reversed (thinking of the sequence as running from left to right is clearly arbitrary).

The probability of a sequence of contiguous helix-states starting at site m and running for n states to site $m+n-1$ is given by

$$P_{\mathrm{H}}(m,n) = p_{01}(m)\left(\prod_{k=m}^{m+n-2} P(k)\right)[1 - P(m+n-1)] \tag{37}$$

while the number of helix-sequences containing n base pairs is then obtained by summing the probability given above over the index

$$N_{\mathrm{H}}(n) = \sum_{m=1}^{N-n+1} p_{\mathrm{H}}(m,n) \tag{38}$$

Finally, we show that one can calculate the probability of any specific combination of helix and coil states in the molecule. To this end we need the conditional probabilities for coil and helix sequences. These are given by the following simple relations (where again, lower-case p represents

an a priori probability and upper-case P represents a conditional probability):

$$P_L(n)=p_L(n)/p(1) \quad \text{and}$$
$$P_R(n)=p_R(n)/p(n)$$
$$P_I(m,n)=p_I(m,n)/p(m) \quad \text{and}$$
$$P_H(m,n)=p_H(m,n)/p(m) \tag{39}$$

where the a priori probabilities are given in Eqs. (22), (24), (26) and (37).

As a specific example we treat the following conformation:

$$\overbrace{0\,0\,0\,0\,0}^{5}\,\overbrace{1}^{1}\,\overbrace{1\,0}^{2}\,\overbrace{0\,0\,0}^{3}\,\overbrace{1}^{1}\,\overbrace{1\,0}^{2}\,\overbrace{0\,0\,0}^{3} \tag{40}$$

The probability of this conformation expressed in terms of the probabilities given in Eq. (39) is then

$$p=[1-p(1)]P_L(5)P(6)P(7)P_I(8,4)P(13)P_R(3) \tag{41}$$

We have already shown that the end and interior-loop conditional probabilities can be expressed in terms of the $P(m)$, thus giving the probability of the above conformation in terms solely of known quantities.

4. Application to the genome of *Treponema pallidum*

The complete genome sequence of *Treponema pallidum*, the spirochete that causes syphilis, has been determined by Fraser et al. [5]. This organism contains a circular chromosome of slightly over 1000 kilobase pairs which makes it one of the smallest prokaryotic genomes and hence a good example for us to examine. Fraser et al. found that the genome consists of precisely 1 138 006 base pairs divided into 1041 genes giving the average gene size as approximately 1000 base pairs.

We will begin our statistical mechanical investigation of this genome by examining the behavior of a sequence of eight genes taking the sequence

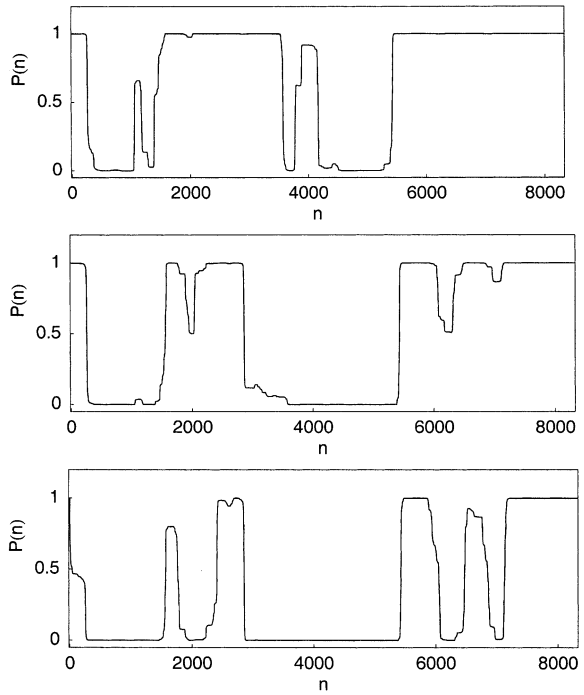

Fig. 1. Probability profiles for genes 516–523 of *Treponema pallidum*. The curves from top to bottom are for $T=371$, 372 and 373 K.

gene-516–523. In the chromosome there are spacers of variable length between successive genes and occasionally genes overlap to a small extent. For this sequence the genes are almost contiguous, and as an example we treat them as being exactly so, giving a gene sequence containing 8314 base pairs.

For the purpose of calculation we clamp the ends of this sequence into the helix-state. The probability profiles for this sequence are shown in Fig. 1 for three different values of the temperature $T=371$, 372 and 373 K from top to bottom, respectively. One sees that the probability profiles consist of distinct regions that are mostly either all-helix or all-coil and that the average length of helix or coil sequences is hundreds of base pairs long (i.e. a result of the very small value of σ as given by Eq. (10)).

In the upper graph in Fig. 2 we show the average s values at $T=372$ K for blocks of units M base pairs long where the block size taken is

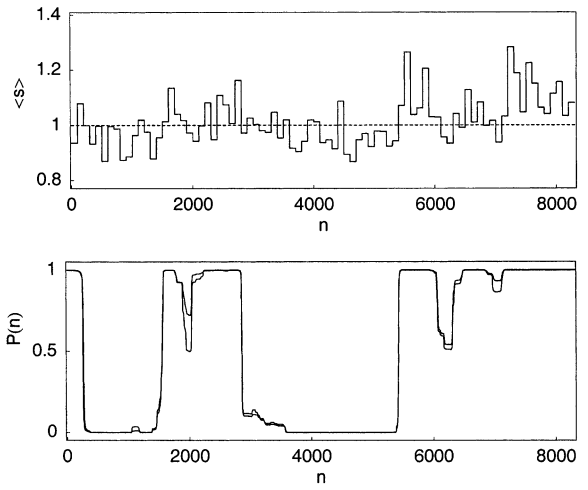

Fig. 2. The upper curve shows the average s value over blocks of 100 base pairs for the eight-gene sequence used in Fig. 1 for $T=372$ K. The lower curve repeats the exact probability profile from Fig. 1 for $T=372$ K and superimposes the probability profile calculated using the average s values for blocks of 10 base pairs.

$M=100$ units where the dashed curve indicates the locus of $s=1$. One sees that this average s function correlates well with the structure of the probability profile which is reproduced from Fig. 1 for $T=372$ K in the graph below it: regions where $\langle s\rangle>1$ tend to be all-helix while regions where $\langle s\rangle<1$ tend to be all-coil. The other curve superimposed on the lower graph in Fig. 2 will be discussed shortly.

We see in Fig. 2 that the profile of the average value of s over blocks of 100 units correlates very well with the structure of the probability profile. We now raise the question of how similar this profile of the average blocked-s values is to what would be expected for a random sequence of base pairs having the same overall base composition. Of course, the base sequence is not random since it contains the information for the construction of the proteins required for the existence of the organism. But from the point of view of the thermal stability of the double-helix it could well be that there is little difference from the behavior given by a random sequence. We explore this question using our eight-gene sequence. For this piece of the molecule we have counted the number of sequences containing n contiguous A or T units or n contiguous C or G units as a function of n. Notice that we lump A and T together and C and G together. Now, if two different kinds of unit are placed at random in a linear array with one type of unit having probability p and the other probability $(1-p)$, then for a chain having a total of N units the following relation gives the average number of sequences expected containing n contiguous units of one type:

$$f(n)=N(1-p)^2p^n \tag{42}$$

with an analogous equation for the other type of unit with p replaced by $(1-p)$. In Fig. 3 we compare the actual sequence statistics found in the

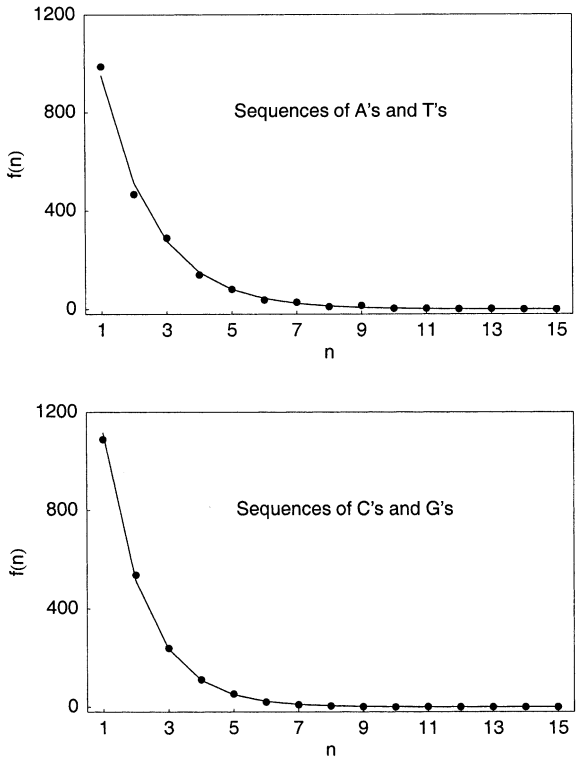

Fig. 3. Base pair sequence statistics for the gene sequence 516–523 of *Treponema pallidum*. The upper and lower curves give the results for sequences of A–T and C–G base pairs, respectively. The solid dots give the actual number of sequences found while the solid lines are the expected results given by Eq. (42) for a random sequence with 54.0% A–T and 46.0% G–C pairs.

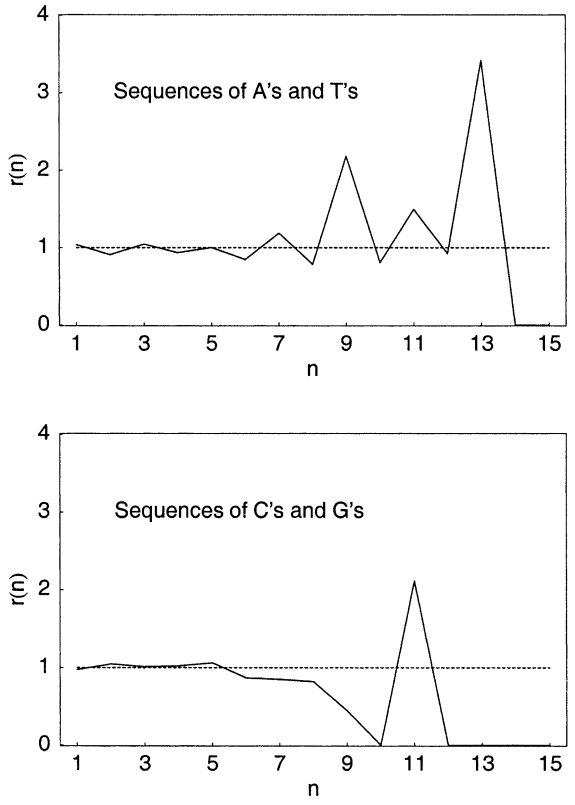

Fig. 4. The same data shown in Fig. 3 but plotted as the ratio of the actual sequence statistics found (the solid dots in Fig. 3) to the expected statistics for random occurrence (the solid lines in Fig. 3). The dashed line is the unit line for reference.

gene sequence 516–523 with those expected for a random sequence as given by Eq. (42). The results for A–T sequences are shown in the upper graph in Fig. 3 where the solid dots give the actual number of A–T sequences containing n contiguous units found and the solid line is the result given by Eq. (42) for a random chain having a total of 8314 units with 54.0% A–T and 46.0% C–G. Analogous results are shown in the lower graph for sequences of C–G. One sees that with respect to the occurrence of sequences of A–T and C–G there is no difference from the results expected for a random sequence. A finer comparison is given in Fig. 4 using the data given in Fig. 3. Here we plot the ratio of the actual sequence statistics found (the solid dots in Fig. 3) to the results expected for random array (the solid lines in Fig. 3); the dashed line gives the unit line for comparison. One now sees that there is some difference between the DNA statistics and the results expected for a random chain, especially for longer sequences of A–T. From this exercise we see that from the point of view of the thermal stability of the double-helix there is essentially no difference from what one would expect from a random array of A–T and C–G base pairs. From the point of view of information content one would expect that certain triplets corresponding to more common amino acids would occur more often. It is an interesting question as to whether there is any biological benefit from having certain regions in the double-helix more thermally stable than others.

The fact that the statistics of the frequency of occurrence of weak helix formers (A and T) and of strong helix formers (C and G) shows little difference from that expected for random occurrence, as illustrated in Figs. 3 and 4, means that plots of $\langle s \rangle$ for a random sequence look qualitatively the same as those shown in Fig. 2. Thus a random sequence gives a probability profile that is qualitatively the same as that found in the above figures with distinct helix and coil regions extending over hundreds of base pairs. The essence of the thermal equilibrium in DNA is that regions that are statistically rich in strong helix formers (with essentially a random occurrence) are correlated by the nucleation parameter σ and the loop-effect of Eq. (10) to give conformational features that cover hundreds of base pairs.

The computer time required by our algorithm given in Eqs. (16)–(19) increases with the square of the number of base pairs in the sequence being treated. For $N=1000$ the calculation of the probability profile takes about a minute on a standard desktop computer. Thus a calculation for a sequence containing $N=10^5$ base pairs is possible, but time consuming. There is a simple way to greatly reduce the computer time required and that is suggested by the correlation we have found between the $\langle s \rangle$ profile, such as given in Fig. 2, and the probability profile giving the probability that a given base pair is in the helix-state. The idea is that if the values, where the average is taken over blocks of, say, 50 units, correlates well

with the occurrence of helix and coil states, then one can use blocks of base pairs as the unit rather than individual base pairs. This approximation has been described by Crothers and Kallenbach [14] and has been applied to the thermal transition in DNA by Poland and Scheraga [2,3]. By using blocks of base pairs rather than individual base pairs as the unit one forces helix and coil sequences to grow in multiples of the block size and thus constrains the system, lowering the combinatorial entropy of these sequences. This effect can be compensated for, by using an approximate correction that is quite accurate. For the units at the border between helix and coil sequences one includes a statistical weight that reintroduces a variation in the location of the border. There are several ways one can do this and we pick one of the simplest. We consider a block of M units in the helix-state bordering a block of coil states. If we simply consider blocks as units then this block would contribute a factor s^M to the partition function where s is the average value over the particular block. The approximation we use is to replace this single term with the sum

$$\sum(s) = s + s^2 + \ldots + s^M = \frac{s^M - 1}{1 - 1/s} \tag{43}$$

where again, s is the average value of this parameter over the particular block. When $s = 1$ one has $\sum = M$ and one sees that this reintroduces some combinatorial fuzziness (the size of the block) at the borders between helix and coil sequences. Note that in Eq. (43) we have omitted the term with zero helix-states present in the block since this would reduce it to a coil block and that would result in an overcount of such blocks.

Such a calculation is shown in the lower graph of Fig. 2 where it is superimposed on the exact profile for the case of $T = 372$ K. The block size used is $M = 10$ and one sees that the profile calculated using the blocking approximation reproduces all of the major features of the exact profile.

We conclude our application of statistical mechanics to the thermal behavior of the genome of *Treponema pallidum* by treating a 100-gene sequence containing 107 139 base pairs (almost 10% of the total genome). The genes we treat are

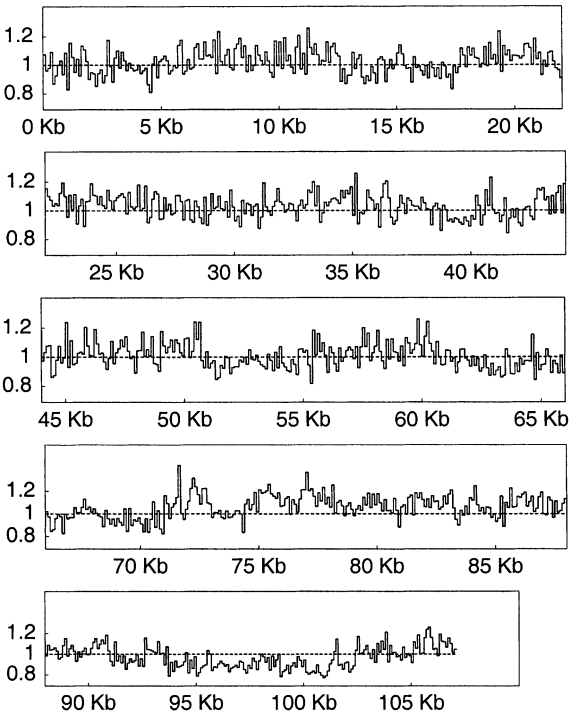

Fig. 5. The average s values over blocks containing 100 base pairs at $T = 375$ K for the 100 gene sequence 790–889 containing 107 139 base pairs. The dashed line indicates the locus of $s = 1$.

in the sequence 790–889. The profile of $\langle s \rangle$ at $T = 375$ K for this gene sequence, again using blocks containing 100 base pairs, is shown in Fig. 5; the dashed line indicates the locus of $s = 1$.

Using the blocking approximation described above, with a block size of $M = 50$, we obtain the probability profile for our 100-gene sequence at T_0 as shown in Fig. 6. The ticks on the top of the frame of the graph indicate the gene boundaries. As with most of our previous sequences, we clamp the first and last units in the sequence in the helix-state. Comparing Figs. 5 and 6 one sees that on a very large scale the plot of the average s over blocks of 100 base pairs correlates very well with the helix probability profile.

An important property of DNA probability profiles, such as those shown in Fig. 6 is that they exhibit distinct plateaus where the helix probability is essentially one over a range of several hundred

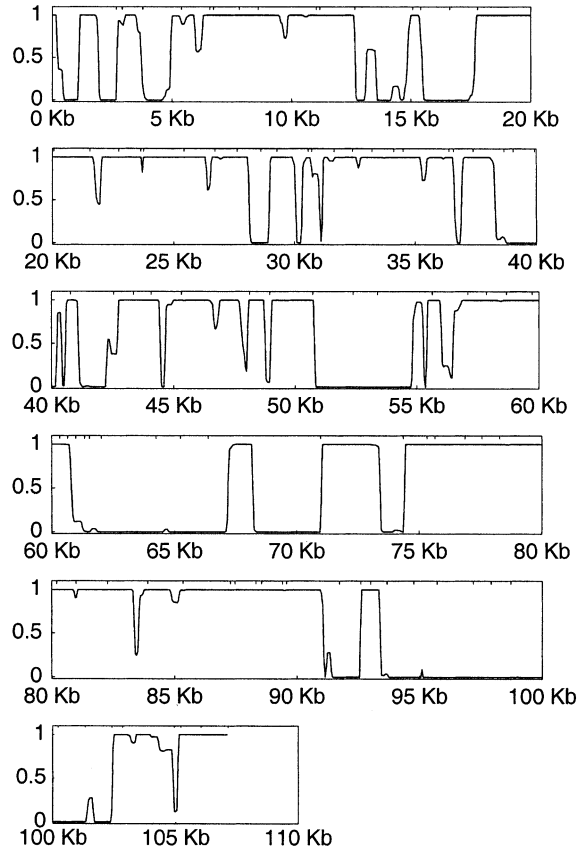

Fig. 6. The probability profile for the 100-gene sequence used in Fig. 5. The profile was calculated using the block-approximation with blocks of 50 base pairs. The gene sequence contains 107 139 base pairs.

base pairs. Thus, a practical approach to the statistical mechanics of an entire genome is to use the block-approximation together with the correction of Eq. (43) to locate the helix-plateaus. Using this approximation one can easily treat consecutive regions containing approximately 10^5 base pairs in the genome. Once the suitable helix-plateaus, separated by approximately 10^4 base pairs are located, one can then apply the algorithm of Eqs. (16)–(19) to the region between the plateaus without using any approximation. Because there is a helix-plateau at each end of the region treated, the boundary conditions for the application of the algorithm are that a unit in the center of each bordering plateau has unit probability of being in the helix-state.

One has thus divided the total genome up into independent manageable units of approximately 10^4 base pairs each. The key to this approach is the use of the block-approximation, which is a very good approximation to the exact statistical mechanics, to locate the helix-plateaus. At each value of the temperature one must relocate the helix-plateaus used since some of them will melt as the temperature is increased. In this manner one can very accurately map out the thermal behavior of the entire genome of an organism and thereby obtain a picture of the thermodynamic stability of helix in the genome as a function of position in the sequence.

References

[1] D. Poland, Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations, Biopolymers 13 (1974) 1859–1871.
[2] D. Poland, H.A. Scheraga, Equilibrium unwinding in finite chains of DNA, Physiol. Chem. Phys. 1 (1969) 389–446.
[3] D. Poland, H.A. Scheraga, Theory of Helix–Coil Transitions in Biopolymers, Academic Press, New York, 1970.
[4] D. Poland, Cooperative Equilibria in Physical Biochemistry, Oxford University Press, Oxford, 1978.
[5] C.M. Fraser, S.J. Norris, G.M. Weinstock, et al., Complete genome sequence of *Treponema pallidum*, the syphilis spirochete, Science 281 (1998) 375–388, The annotated genome sequence is available on the World Wide Web at www.tigr.org.
[6] 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.
[7] J. SantaLucia, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl Acad. Sci. USA 95 (1998) 1460–1465.
[8] M.E. Fisher, Effect of excluded volume on phase transitions in biopolymers, J. Chem. Phys. 45 (1966) 1469–1473.
[9] D. Poland, H.A. Scheraga, Occurrence of a phase transition in nucleic acid models, J. Chem. Phys. 45 (1966) 1464–1468.
[10] O. Gotoh, Y. Tagashira, Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles, Biopolymers 20 (1981) 1033–1042.

[11] R.H. Lacombe, R. Simha, Probability profiles using conditional probabilities, J. Chem. Phys. 58 (1973) 1043–1053.
[12] D. Poland, Evaluation of linear chain partition functions by consideration of sequence conditional probabilities, J. Chem. Phys. 60 (1974) 808–812.
[13] S.J. Chen, K.A. Dill, Theory for the conformational changes of double-stranded molecules, J. Chem. Phys. 109 (1998) 4602–4616.
[14] D.M. Crothers, N.R. Kallenbach, On the helix–coil transition in heterogeneous polymers, J. Chem. Phys. 45 (1966) 917–927.