Partly melted DNA conformations obtained with a probability peak finding method

Eivind Tøstesen
Department of Tumor Biology, The Norwegian Radium Hospital, N-0310 Oslo, Norway
(Received 19 November 2004; published 28 June 2005)

Peaks in the probabilities of loops or bubbles, helical segments, and unzipping ends in melting DNA are found in this article using a peak finding method that maps the hierarchical structure of certain energy landscapes. The peaks indicate the alternative conformations that coexist in equilibrium and the range of their fluctuations. This yields a representation of the conformational ensemble at a given temperature, which is illustrated in a single diagram called a stitch profile. This article describes the methodology and discusses stitch profiles vs the ordinary probability profiles using the phage lambda genome as an example.

DOI: 10.1103/PhysRevE.71.061922 PACS number(s): 87.14.Gg, 87.15.Ya, 05.70.Fh, 02.70.Rr

I. INTRODUCTION

DNA melts by a stochastic formation and growth of loops [1] and tails (i.e., unzipping ends). Loop formation is also induced in the dense cellular environment and is at the heart of DNA biology [2]. For the past three decades, numerically calculated properties of the DNA melting process have been represented by plotting curves of three types: probability profiles, melting curves, and Tm profiles. Probability profiles [3,4] are plots of the basepairing probability $p_{\mathrm{bp}}(i)$ or the "upside-down" $1-p_{\mathrm{bp}}(i)$ vs sequence position i. Melting curves [5,6] are plots of the helicity Θ or its derivative vs temperature T. Tm profiles are plots of the base pair melting temperature $T_{\mathrm{m}}(i)$ vs sequence position i. Apparently, there has been less interest in calculating other properties, one reason being perhaps that the ordinary plots outline the main features on the experimental side, such as the melting curves from UV spectroscopy and differential scanning calorimetry. However, there are other interesting properties within reach of calculations. For example, what are the sizes and locations of loops and how do they fluctuate? How do distant loops correlate? What are the alternative conformations of a region that coexist when it melts? What events are predominant and what others are rare? In addition to these questions being important *per se*, advances in single molecule techniques [7–9] provide new types of measurement of the micromechanical, dynamical, and structural properties, and motivate predictions beyond the ordinary curves [10–12].

In this article, we turn our attention to stitch profiles. A *stitch profile* is a diagrammatic representation of the alternative DNA conformations that coexist at a given temperature [13]. Figure 1 shows the four types of graphical elements called *stitches* that go into a stitch profile. A stitch represents either a loop, a right tail, a left tail or a helical region, as shown, and indicates its boundary positions and the ranges of fluctuation of these positions. In analogy with sewing, where a thread coming up and down through the fabric forms a row of stitches, any conformation of DNA is an alternating row of blocks of open or closed basepairs. A stitch profile indicates alternative conformations as alternative threads (i.e., paths) through the diagram. The aim of this work is to develop a method for constructing stitch profiles and to discuss stitch profiles vs probability profiles.

In the Poland-Scheraga model [14], a state of the chain molecule is specified by N binary variables, $x_1,\ldots,x_N$, where the jth variable x_j (=0 or 1) indicates if the jth base in the sequence is base paired or not with the complementary strand. While the classical three types of curves are based on calculating the base pairing probabilities related to the state x_j of each base pair, a stitch profile, in contrast, is based on probabilities of blocks of base pairs being in states corresponding to loops, helical segments or tails. A stitch profile made "by hand" was introduced in Ref. [13] (where we referred to it as a loop map) in order to suggest an application of such block probabilities. The article described a DNA melting algorithm with two important features: a speedup based on multiplication of symmetrical left side and right side partition functions; and the statistical weight of a base pair depending rigorously on both of its neighbors. These features allow the block probabilities to be easily calculated as follows.

A loop is a consecutive series of 0's (melted basepairs) bounded by 1's at positions x and y, where $1\leqslant x<y-1<N$. The probability of a loop is calculated by decomposing the chain in three segments,

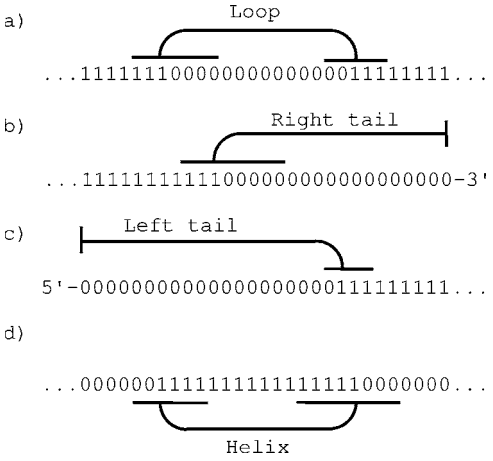

FIG. 1. A stitch profile is composed of four types of stitches: (a) loops, (b) right tails, (c) left tails, and (d) helices. Loop and tail stitches are drawn on the upper side and they span regions of opened base pairs (0's). Helix stitches are drawn on the lower side and they span regions of closed base pairs (1's). The horizontal bars indicate fluctuational ranges of the 0-1 boundaries.

$$p_{\text{loop}}(x,y) = Z_{\text{X10}}(x)\Omega(y-x)Z_{\text{01X}}(y)/\beta Z. \quad (1)$$

$Z_{\text{X10}}(x)$ is a partition function characterizing the segment $[1,x+1]$, $Z_{\text{01X}}(y)$ is a partition function characterizing the segment $[y-1,N]$, Z is the total partition function of the whole chain, $\Omega(y-x)$ is the loop entropy factor (a function of loop size), and β is related to the equilibrium constant of complete dissociation of the two strands. A tail is a block of 0's that extends to the end of the chain. The probability of a right tail from position N (the right chain end) to a bounding 1 at position x is

$$p_{\text{right}}(x) = Z_{\text{X10}}(x)/\beta Z. \quad (2)$$

The probability of a left tail from position 1 (the left chain end) to a bounding 1 at position y is

$$p_{\text{left}}(y) = Z_{\text{01X}}(y)/\beta Z. \quad (3)$$

A helical region is a block of 1's bounded by 0's or by the chain ends. If x is the position of the first 1 in the block and y is the position of the last, where $1 \leqslant x \leqslant y \leqslant N$, then the probability can be written

$$p_{\text{helix}}(x,y) = Z_{\text{X01}}(x)\Xi(x,y)Z_{\text{10X}}(y)/\beta Z, \quad (4)$$

where $Z_{\text{X01}}(x)$ is a partition function characterizing the segment $[1,x]$ and $Z_{\text{10X}}(y)$ is a partition function characterizing the segment $[y,N]$. The *stacking chain* function $\Xi(x,y)$ is the statistical weight of the block of 1's, given as

$$\Xi(x,y) = \begin{cases} s_1(x)s_1(y)\prod_{j=x+1}^{y} s_{11}(j) & \text{for } x < y, \\ s_{010}(x) & \text{for } x = y, \end{cases} \quad (5)$$

where $s_{11}(x)$ is the statistical weight of nearest neighbor basepairs (a pair of 1's), $s_1(x)$ is the statistical weight of a helix-ending basepair (0 on one side, 1 on the other), and $s_{010}(x)$ is the statistical weight of an isolated basepair (0's on both sides). Equations (1)–(4) correspond to Eqs. (12)–(15) in Ref. [13], respectively.

The block probabilities p_{loop}, p_{right}, p_{left}, and p_{helix} depend on precisely located boundaries x and/or y. But thermal motion causes the boundaries to fluctuate. These fluctuations are represented by fluctuation bars in a stitch profile. They are not merely attributes like "error bars," but rather an essential ingredient. Each stitch represents not a single conformation of a region, but a grouping of conformations that are supposed to be related via fluctuations. In a plot of any of the block probabilities as a function of x and/or y, such a grouping will appear as a broad peak, and the fluctuation bar(s) indicate the extent of the peak. A stitch profile is simply a representation of the peaks in the four block probability functions, and the problem of constructing a stitch profile is basically a peak finding problem.

The peak finding problem is important in data and signal analysis in diverse areas of science, for example, in various types of spectroscopy and image analysis. Peak finding has also been used in statistical mechanics to define macroscopic states of RNA secondary structure: native, intermediate, molten, and denatured states [15]. The following issues apply to our case here: A peak's size is given by its volume rather than its height. If a peak is broad enough, it may have a higher volume than another peak, even if it has a lower height. Probability peak heights can be very low, so we can not use a height cutoff for detecting peaks. There is no erroneous noise in our calculated probabilities, so smoothing should not be used. Actually, the shapes of peaks are quite irregular with "peaks within peaks," so the problem is *hierarchical* peak finding (analogous to hierarchical clustering vs clustering). There may be limited space in a stitch profile, so it must be chosen which peaks to represent.

The right and left tail stitches are found by 1D peak finding in p_{right} and p_{left}, respectively, and the loop and helix stitches are found by 2D peak finding in p_{loop} and p_{helix}, respectively. The challenge is not so much finding a peak as deciding its extent. In 1D we use an interval—the fluctuation bar—to delimit a peak. In 2D we use a *frame*, that is, the Cartesian product of the two fluctuation bars on the x axis and the y axis, to represent a peak by "framing" it. Let the *peak volume* p_{v} be defined by the probability summed over the interval in 1D or the frame in 2D.

This article describes a probability peak finding method in 1D based on a detailed mapping of the hierarchical structure. The 2D case is solved by combining the 1D results for x and y. For the extent of a peak, the main idea is to find where the probability has dropped to a certain fraction relative to the peak maximum value. This fraction is controlled by a parameter to the algorithm and it determines the widths of the fluctuation bars. These widths, in turn, determine the peak volumes that can be used for choosing if stitches are included or not in the stitch profile.

II. THE METHOD

Minus the logarithm of a probability is an energylike quantity. Using this, we transform probability peak finding into finding the wells or lakes in a (pseudo)energy landscape. A peak with a certain ratio between the probability at the maximum and the probabilities at the edges corresponds to a lake with a certain depth in an energy landscape. The analogy to mountain landscapes, lakes, ponds, etc., is standard in statistical mechanics [16]. We use it here to redefine the stitch profile problem of peak finding to be a lake finding problem in four energy landscapes—two 1D landscapes:

$$E_1(x) = -\log_{10} p_{\text{right}}(x), \quad (6)$$

$$E_2(y) = -\log_{10} p_{\text{left}}(y), \quad (7)$$

and two 2D landscapes:

$$E_3(x,y) = -\log_{10} p_{\text{loop}}(x,y), \quad (8)$$

$$E_4(x,y) = -\log_{10} p_{\text{helix}}(x,y). \quad (9)$$

A. Peak finding method in 1D

The 1D method is described here using $E_1(x)$ as an example, but $E_2(y)$ is treated the same way. Of all the possible lakes that can be created by filling water into the various wells, we restrict ourselves to considering only a finite set of representative lakes. Let Ψ_1 be the set of sequence positions

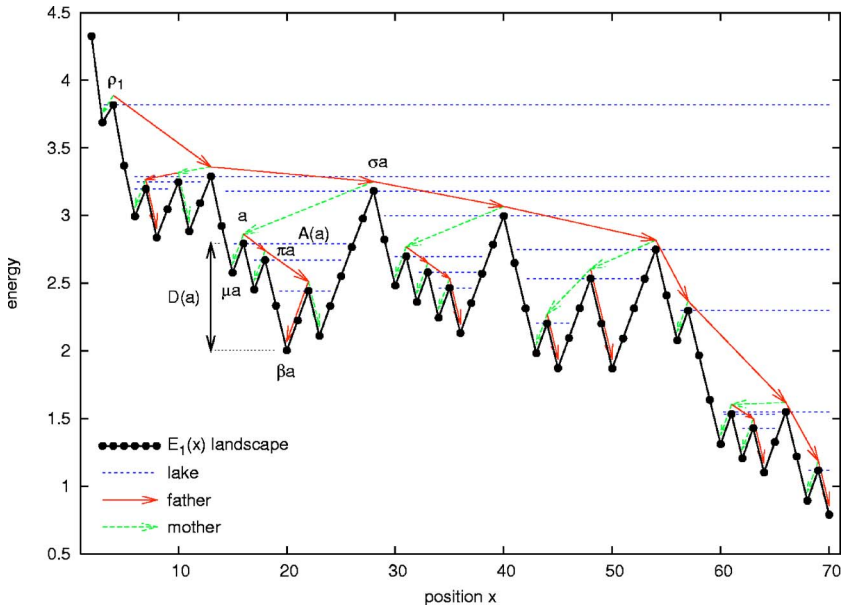

FIG. 2. (Color online) $E_1(x)$ is plotted for a 70 bp sequence to illustrate the pedigree ordering of lakes in an energy landscape. Lakes corresponding to each local maximum are shown as horizontal dashed (blue) lines. Arrows indicate the binary tree starting from the root $\rho_1=4$. Paternal lines are connected series of fathers shown as solid (red) arrows. The node $a=16$, for example, has the lake $L(a)=[15,28]$, bottom $\beta a=20$, depth $D(a)\simeq 0.8$, father $\pi a=18$, mother $\mu a=15$, and successor $\sigma a=28$, as shown.

a at which E_1 has an extremum. Minima and maxima in Ψ_1 are alternating along the x axis. To each element $a\in\Psi_1$ we associate a *lake* $L(a)$ in the landscape, see Fig. 2. The altitude of the surface (water level) is $E_1(a)$ and the lake surface spans the interval $L(a)=[L_L(a),L_R(a)]$ given by

$$a \in L(a), \tag{10a}$$

$$\forall x \in L(a): E_1(x) \leqslant E_1(a), \tag{10b}$$

$$E_1(L_L(a)-1) > E_1(a) \text{ or } L_L(a)=1, \tag{10c}$$

$$E_1(L_R(a)+1) > E_1(a) \text{ or } L_R(a)=N. \tag{10d}$$

When a is a minimum, in most cases $L(a)=\{a\}$. When a is a maximum, the lake $L(a)$ is nearly split in two adjacent lakes, divided at position a where the local depth becomes zero. However the corresponding probability peak is not split in two by a zero probability, so we consider $L(a)$ as one lake. The *bottom* βa of a lake $L(a)$ is defined as the position with lowest energy in the lake,

$$\beta a = \arg \min_{x\in L(a)} E_1(x). \tag{11}$$

The *depth* $D(a)$ of a lake $L(a)$ is defined as the energy difference between the surface and the bottom: $D(a)=E_1(a)-E_1(\beta a)$. Some lakes are contained inside deeper lakes: $L(a)\subset L(b)$. This partial ordering of lakes in Ψ_1 defines a hierarchical structure [17]. Assuming that both the leftmost and the rightmost (on the x axis) elements in Ψ_1 are minima (terminal maxima could just be excluded), the elements in Ψ_1 can be considered as the nodes of a binary tree. The root ρ_1 of the tree is the global maximum and its lake $L(\rho_1)$ spans the entire sequence (or almost).

1. Pedigree ordering

Imagine a walk along the branches of the binary tree. In order to orient itself, the walker needs "road signs" at each node a that point the directions to the root ρ_1 and the bottom βa. This imposes a structure similar to a pedigree (i.e., tree of ancestors) as illustrated in Fig. 2. Each node $a\neq\rho_1$ is connected upwards in the direction of the root to a unique node σa called the *successor* of a. This means $L(a)\subset L(\sigma a)$. Each node a corresponding to a maximum is also connected downwards to two parent nodes: a *father* node πa in the direction of the bottom; and a *mother* node μa in the other direction. This means that the father is the parent with the lowest bottom, $E_1(\beta\pi a)<E_1(\beta\mu a)$, and that $\beta\pi a=\beta a$. But it does *not* imply that $D(\pi a)>D(\mu a)$. In contrast to an offspring tree, parents are located in the direction away from the root, rather than the reverse. Define the set of successors of a node a:

$$\Sigma(a)=\{a,\sigma a,\sigma^2 a,\sigma^3 a,\ldots,\rho_1\}. \tag{12}$$

The set $\Sigma(a)$ traces a path from a up to the root. Define the set of ancestors of a node a:

$$\Delta(a)=\{b\in\Psi_1 | a\in\Sigma(b)\}. \tag{13}$$

The set $\Delta(a)$ is the subtree that has a as its root or top node. Each node $a\in\Psi_1$ belongs to a unique *paternal line*,

$$\Pi(a)=\{\varphi a,\ldots,a,\pi a,\ldots,\beta a\}, \tag{14}$$

which is a maximal set of nodes that are related through a series of fathers. The series ends at their common bottom node βa and, oppositely, it begins at a node, called the *full* node φa, which is not itself a father. Each node corresponding to a minimum is the bottom of its paternal line. And each node which is either a mother or the root ρ_1 is the full node of its paternal line. For a node a that is both a minimum and a mother, $\Pi(a)=\{a\}$. The term "full" stems from filling water into a bottom: The paternal line indicates successively deeper lakes until the full lake is reached. The successor of the full lake belongs to another paternal line and corresponds to another bottom being filled.

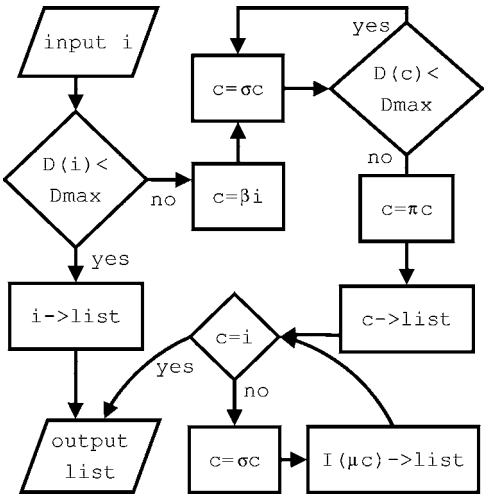

FIG. 3. The MAXDEEP algorithm that finds stitches with depths below $D_{\max}$. For a given input node i, it returns a list of those nodes in $\Delta(i)$ that fulfill the criterion in Eq. (15). c is the tree climber, $\rightarrow$ means push (i.e., "put on a list"), and $I(\mu c)$ is the output of a recursive call to the algorithm itself with μc as the input node. The diagram can be read in conjunction with Fig. 2.

2. The MAXDEEP algorithm

The lake finding task at hand is to search through the set of lakes corresponding to the set Ψ_1, and select the lakes to be represented in a stitch profile. A possible solution is the MAXDEEP algorithm. With a given parameter $D_{\max}$, the algorithm finds all nodes $a \in \Psi_1$ where

$$D(a) < D_{\max}, \tag{15a}$$

$$D(\sigma a) \geqslant D_{\max} \text{ or } a = \rho_1. \tag{15b}$$

In words, this means that they should be as deep as possible without exceeding the maximum depth. In many cases, it can be expected that the depth increases in just a small step from a node to its successor, but not for full nodes that often have successors that are much deeper than themselves. Therefore, some of the selected lakes will have depths close to $D_{\max}$, while others (the full ones) can be more shallow. The MAXDEEP algorithm is shown in Fig. 3. It is not necessary to evaluate all $a \in \Psi_1$. Instead, the MAXDEEP algorithm involves a "tree climber" c that basically climbs up the paternal line $\Pi(\rho_1)$ of the input node $i=\rho_1$, starting from the bottom $\beta\rho_1$, until it exceeds $D_{\max}$, and then takes one step down again. At that point, c fulfills the criterion in Eq. (15), while all other nodes in $\Sigma(c)$ and $\Delta(c)$ do not. Subsequently, the algorithm calls itself recursively with mothers of $\Sigma(c)$ as input nodes, in order to explore other paternal lines.

The output of the MAXDEEP algorithm is illustrated in Fig. 4. Lakes have depths between zero and $D_{\max}$ and they cover a large fraction of the landscape. The figure shows that the effect of increasing $D_{\max}$ is that some lakes become wider and deeper, some lakes merge, and some lakes (corresponding to full nodes) remain unchanged.

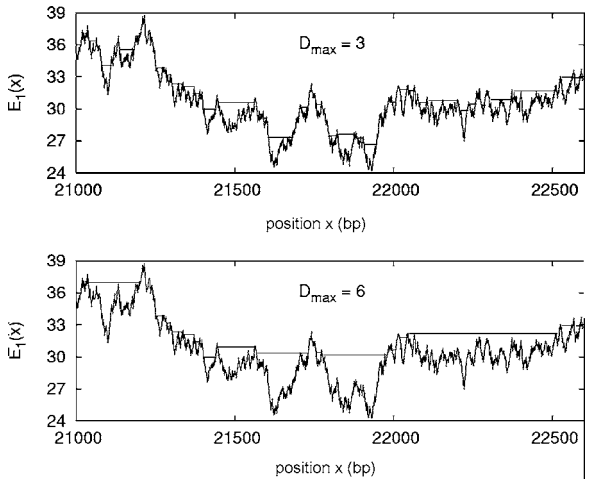

FIG. 4. Lakes in 1D found by the MAXDEEP algorithm. The two plots show the same region of an energy landscape. The first plot shows the lakes found by the algorithm using $D_{\max}=3$ and the second plot shows the lakes with $D_{\max}=6$.

3. The largest peaks

Some of the lakes found by the MAXDEEP algorithm are probably not very significant, having low depths and low peak volumes at the same time. A second selection process is required before we finally get the stitches for the profile. Since the MAXDEEP algorithm does not consider the peak volumes p_v, these can be considered in the second selection. If we want to select the largest peaks, for example, this can be achieved by having a *cutoff* as a second parameter p_c and select only nodes a with peak volume $p_v(a) \geqslant p_c$.

B. Peak finding method in 2D

While a 1D lake is completely described as an interval $[L_L(a), L_R(a)]$, lakes in the 2D landscapes E_3 and E_4 have complicated contours and perhaps islands in the interior [17]. Fortunately, we only need to know a 2D lake's extents on the x axis and the y axis, in order to enclose it in a frame, which is needed in a stitch profile (cf. Sec. I).

1. The helix landscape

Consider the lake finding problem in the helix landscape E_4. Although isolated base pairs are allowed [$x=y$ in Eq. (4)] we will ignore those instances for convenience, and require $1 \leqslant x < y \leqslant N$ in the landscape. It follows from Eqs. (4) and (5) that

$$p_{\text{helix}}(1,y)p_{\text{helix}}(x,N) = p_{\text{helix}}(x,y)p_{\text{helix}}(1,N). \tag{16}$$

Defining two 1D landscapes,

$$E_5(x) = -\log_{10} p_{\text{helix}}(x,N), \tag{17}$$

$$E_6(y) = -\log_{10} p_{\text{helix}}(1,y), \tag{18}$$

we can write $E_4(x,y)=E_5(x)+E_6(y)+\text{const}$. This decoupling of x and y allows us to analyze the lakes in E_4, based on 1D analysis of E_5 and E_6 and their binary trees Ψ_5 and Ψ_6.

Let $a \in \Psi_5$ and $b \in \Psi_6$ and consider the frame $L(a) \times L(b)$. Does such a frame enclose a 2D lake in E_4? At least, the frame should cover a region of E_4. The frame is said to be *above the diagonal* if $L_R(a) < L_L(b)$, such that all $(x,y) \in L(a) \times L(b)$ are points in the helix landscape. Because of the decoupling, the minimum in a frame is

$$\arg \min_{(x,y) \in L(a) \times L(b)} E_4(x,y) = (\beta a, \beta b). \tag{19}$$

Consider the energy barriers seen from $(\beta a, \beta b)$:

$$\begin{aligned} \Delta E_4(x,y) &= E_4(x,y) - E_4(\beta a, \beta b) \\ &= E_5(x) - E_5(\beta a) + E_6(y) - E_6(\beta b) \\ &= \Delta E_5(x) + \Delta E_6(y). \end{aligned} \tag{20}$$

Using Eq. (10b), we find that $\Delta E_4(x, \beta b) \leqslant D(a)$ for all $x \in L(a)$, and using Eqs. (10c) and (10d), we find that $\Delta E_4(x,y) > D(a)$ just outside the west and east side of the frame. This means that $L(a)$ is the extent on the x axis of a 2D lake with depth $D(a)$ and bottom $(\beta a, \beta b)$. And vice versa, $L(b)$ is the extent on the y axis of a 2D lake with depth $D(b)$ and bottom $(\beta a, \beta b)$. If $D(a) \neq D(b)$, then a 2D lake with bottom $(\beta a, \beta b)$ must have depth $D \leqslant \min\{D(a), D(b)\}$ to be confined by the frame, and it will not extend to all four frame sides. A lake with depth $D > \min\{D(a), D(b)\}$ is not confined and may not even have its bottom inside the frame. If $D(a) = D(b)$ then the frame $L(a) \times L(b)$ is exactly the extent in both dimensions of a 2D lake with that depth.

Unfortunately, we can not expect to find a's and b's with equal depths. In order to best approximate 2D lakes, we want instead $D(a)$ and $D(b)$ to be as close as possible, that is, none of πa, σa, πb or σb should have depths in between $D(a)$ and $D(b)$. This can be formulated as two conditions: we say that (a,b) is "σ above" if

$$D(\sigma b) > D(a) \text{ or } b = \rho_6, \tag{21a}$$

$$D(\sigma a) > D(b) \text{ or } a = \rho_5, \tag{21b}$$

and we say that (a,b) is "π below" if

$$D(\pi b) < D(a) \text{ or } b = \beta b, \tag{22a}$$

$$D(\pi a) < D(b) \text{ or } a = \beta a. \tag{22b}$$

Frames that are σ above, π below, and above the diagonal are good representations of lakes with depth $\min\{D(a), D(b)\}$. Let us examine the three conditions one by one. First, define the set Γ_4 of frames that are σ above,

$$\Gamma_4 = \{(a,b) \in \Psi_5 \times \Psi_6 | (a,b) \text{ is } \sigma \text{ above}\}. \tag{23}$$

Equation (21) shows that $(\rho_5, \rho_6) \in \Gamma_4$. And $(a,b) \in \Gamma_4$ if a and b both correspond to minima, because Eq. (21) is true with $D(a)=0$ and $D(b)=0$. This means $(\beta a, \beta b) \in \Gamma_4$ for any $a \in \Psi_5$ and $b \in \Psi_6$. Just as 1D lakes are hierarchically ordered, so are frames. A frame is contained in another frame, $L(a) \times L(b) \subset L(c) \times L(d)$, if $a \in \Delta(c)$ and $b \in \Delta(d)$. It turns out that the elements of Γ_4 are the nodes of a binary tree with (ρ_5, ρ_6) as its root. We call Γ_4 the *frame tree*, and it is a kind of *product tree* between the trees Ψ_5 and Ψ_6. The successor of a node $(a,b) \neq (\rho_5, \rho_6)$ is

$$\sigma(a,b) = \begin{cases} (\sigma a, b) \text{ if } D(\sigma b) > D(\sigma a) \text{ or } b = \rho_6, \\ (a, \sigma b) \text{ if } D(\sigma a) > D(\sigma b) \text{ or } a = \rho_5. \end{cases} \tag{24}$$

Each node $(a,b) \in \Gamma_4$, where a and b are not both minima, has two parent nodes. We define the bottom of a node as $\beta(a,b) = (\beta a, \beta b)$, which enables us to distinguish the two parent nodes as a father,

$$\pi(a,b) = \begin{cases} (\pi a, b) \text{ if } D(a) > D(b), \\ (a, \pi b) \text{ if } D(a) < D(b), \end{cases} \tag{25}$$

with $\beta\pi(a,b) = \beta(a,b)$, and a mother

$$\mu(a,b) = \begin{cases} (\mu a, b) \text{ if } D(a) > D(b), \\ (a, \mu b) \text{ if } D(a) < D(b), \end{cases} \tag{26}$$

with $\beta\mu(a,b) \neq \beta(a,b)$. This gives the frame tree a pedigree ordering with paternal lines, etc., just as in 1D.

Some frames in Γ_4 are not above the diagonal, such as the root frame (ρ_5, ρ_6) that spans almost the entire sequence in both dimensions. Next, define the set of frames that are σ above *and* above the diagonal:

$$\Psi_4 = \{(a,b) \in \Gamma_4 | L_R(a) < L_L(b)\}. \tag{27}$$

Ψ_4 is organized as a number of disjoint binary trees, $\Psi_4 = \cup_j \Delta(a_j, b_j)$, each one being a subtree of Γ_4. The top node (a_j, b_j) of the jth subtree is above the diagonal, while its successor $\sigma(a_j, b_j)$ crosses the diagonal. And lastly, define the set of frames that are σ above, π below, and above the diagonal:

$$\Upsilon_4 = \{(a,b) \in \Psi_4 | (a,b) \text{ is } \pi \text{ below}\}. \tag{28}$$

This set also consists of disjoint trees, but they are not binary, nodes can have more than two parents.

If we do not require π below and consider the larger set Ψ_4, then such frames are still good representations of lakes. A computational advantage of this is that each subtree $\Delta(a_j, b_j)$ in Ψ_4 can be searched with the MAXDEEP algorithm, by using its top node (a_j, b_j) as the input i. The E_4 lake finding problem is solved by finding frames in Ψ_4 using the MAXDEEP algorithm in this manner, followed by a second selection based on the cutoff p_c, which yields the helix stitches for a stitch profile. For the MAXDEEP algorithm we must define the depth of a frame. A possible depth definition is $D(a,b) = \max\{D(a), D(b)\}$. Using this, the MAXDEEP algorithm finds frames (a,b) with $D(a) < D_{max}$ and $D(b) < D_{max}$.

2. The loop landscape

Consider the lake finding problem in the loop landscape E_3. It follows from Eqs. (1)–(3) that for $1 \leqslant x < y - 1 < N$,

$$E_3(x,y) = E_1(x) + E_2(y) - \log_{10} \Omega(y-x) + \text{const}. \tag{29}$$

x and y do not decouple in E_3. But $\log_{10} \Omega(y-x)$ varies slowly enough to be considered constant as an approximation. With this assumption, it turns out that an analysis parallel to the one for the helix landscape gives reasonable stitch

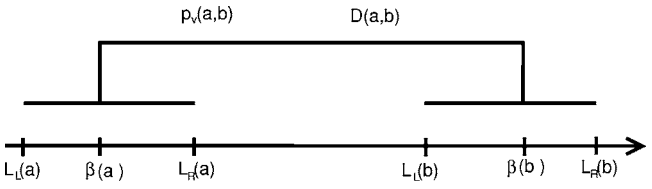

FIG. 5. Eight quantities characterize a loop stitch.

profiles. There are only minor differences between the loop and helix cases. For example, a loop frame is said to be *above the diagonal* if $L_R(a)+1<L_L(b)$. Here is a brief outline: We define the frame tree

$$\Gamma_3=\{(a,b)\in\Psi_1\times\Psi_2|(a,b)\text{ is }\sigma\text{ above}\}. \quad (30)$$

The root of Γ_3 is (ρ_1,ρ_2). By replacing ρ_5 with ρ_1 and ρ_6 with ρ_2 in Eqs. (21) and (24)–(26) we define σ above, successors, fathers, and mothers in Γ_3. The set of σ-above frames that are also above the diagonal is

$$\Psi_3=\{(a,b)\in\Gamma_3|L_R(a)+1<L_L(b)\}. \quad (31)$$

And we search Ψ_3 for frames using the top nodes of the subtrees and the MAXDEEP algorithm. Subsequently, the loop stitches are found by a selection based on the cutoff p_c.

C. Stitch profile data

A mere calculation of the block probabilities [cf. Eqs. (1)–(4)] gives $O(N^2)$ small numbers which are information in a fragmented form. In principle, these block probabilities can be obtained as functions of x and/or y using the Poland-Scheraga model [14], the Peyrard-Bishop model [18], or something else. The peak finding method is applied for putting this information in the more useful form of a stitch profile, which is data of size $O(N)$ only. A stitch profile is a set of stitches of the four types in Fig. 1. As we have seen, each fluctuation bar corresponds to a lake in a 1D landscape. Figure 5 shows how the two fluctuation bars of a loop stitch (a,b) span the lake intervals $L(a)$ and $L(b)$, and how the diagram also indicates the position of the lake bottom $(\beta a,\beta b)$, where the probability peak has its maximum. Two additional quantities are associated with a stitch: the depth $D(a,b)$ and the peak volume $p_v(a,b)$. These quantities can also be illustrated, for example, by labeling the stitch. Thus, eight quantities are associated with each loop or helix stitch, but only five quantities for left or right tail stitches that only have one fluctuation bar.

III. DISCUSSION

This section discusses different aspects of stitch profiles, what information they represent, and the choice of parameters and algorithm. The 48 kbp phage lambda genome (GenBank accession number NC_001416) is used as a test sequence, to illustrate stitch profiles and probability profiles rather than the melting behavior vs biology of lambda [19]. All stitch profiles were calculated for the whole 48 kbp sequence, but the interesting features are viewed in windows of length 1–20 kbp. The stitch profiles in full length are better viewed on a computer than in print [20]. Partition functions were calculated in the Poland-Scheraga model using the algorithm in Ref. [13] with $\beta=1$. The parameter set of Blake and Delcourt [5] at $[Na^+]=0.075$ M was applied, with the loop entropy factor $\Omega(y-x)=\sigma[2(y-x)+1]^{-\alpha}$ reparametrized by Blossey and Carlon [21] using $\alpha=2.15$ and $\sigma=1.26\times10^4$.

A. Alternative conformations

As previously stated, a stitch profile shows a number of alternative conformations that coexist in equilibrium at a given temperature. It does so in two ways: (1) each stitch represents a fluctuational variation of its boundaries, and (2) alternative rows of stitches represent alternative series of loops and helices along the chain. Figure 6 shows a short sequence window of a stitch profile, in which there are nine loop stitches and nine helix stitches. Note how the fluctuation bars and the lake bottoms of loops often coincide with those of helices, although they were calculated independently of each other. This is expected because a helix must begin where a loop or tail ends, of course, so coinciding fluctuation bars reflect the same boundary fluctuation. A *row of stitches* is a series of stitches connected in a chain by coinciding boundaries, and in the diagram, it forms a continuous path or thread, which alternates between the upper and lower side. There are often several stitches to choose among at a given boundary, resulting in a combinatorial number of alternative rows of stitches. Each row of stitches corresponds to a specific conformation (apart from the fluctuational variation) of a region that is much longer than the regions specified by the individual stitches. The stitch profile in Fig. 6 is aligned with a schematic list of the seven alternative conformations corresponding to the possible rows of stitches. These alternatives do not represent the only possible conformations in that window—strictly speaking, any conformation has a nonzero probability—but they represent the most stable conformations in terms of probability peak volume and depth.

Note that Fig. 6 also illustrates that stitches are sorted and stacked vertically in the diagram according to their lengths $\beta b-\beta a$. This is for aesthetic reasons only. There is no quantity associated with the vertical axis. Fluctuation bars are also placed on different levels to avoid overlap.

B. Correlations and cooperativity

DNA cooperativity [14] is the presence of certain long-range *correlations*, that should be distinguished from the long-range *interactions* embedded in the loop entropy factor $\Omega(y-x)$. In a probability profile, cooperativity appear as the characteristic plateaus that indicate the tendency of blocks of basepairs to "melt as one"—being either all 0's or all 1's. Base pairs within such a cooperative region are strongly correlated. This aspect of cooperativity is also prominent in a stitch profile, where the block organization is shown more specifically. For comparison, Fig. 7 shows a stitch profile and a probability profile of the same piece of DNA. Stitches in the diagram are labeled with their peak volumes. As can be

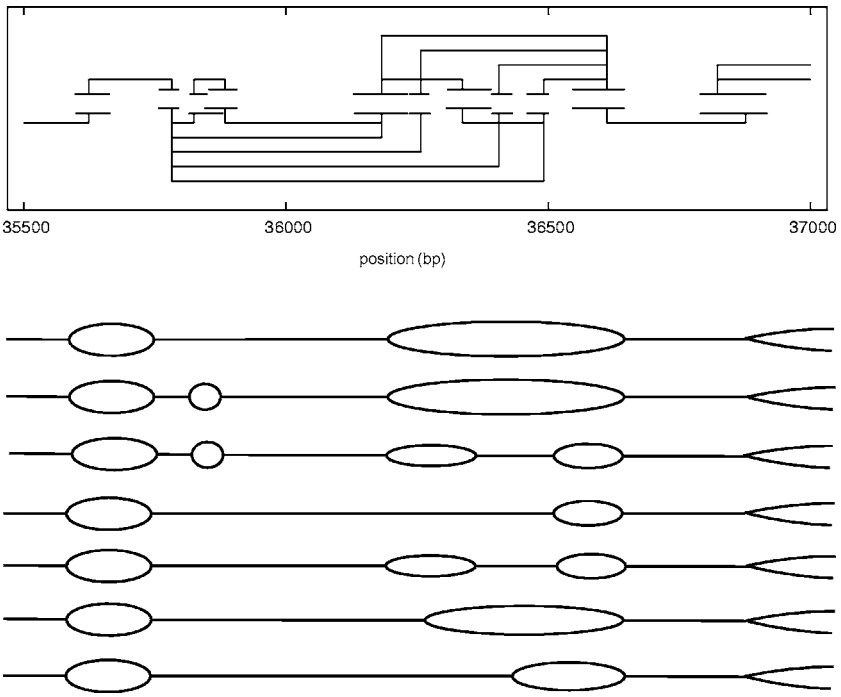

FIG. 6. A stitch profile (in the box) represents a number of alternative conformations. In this example, there are seven possible conformations listed schematically below the box. The parameters are $D_{\max}=3$, $p_c=0.02$, and $T=81.9\ ^\circ\mathrm{C}$ and the sequence window is 35.5–37 kbp.

seen, plateaus in the probability profile can be identified with one or more stitches where there are correspondences in position and between the peak volumes and the plateau height relative to surrounding plateaus.

In addition to the strong correlations inside a cooperative region, there are weaker correlations over longer distances. For example, a certain loop at one site can influence what loops that can exist at a distant site. Information about such correlations across one or more stitch boundaries may be derived from the alternative rows of stitches that indicate the possible multiloop conformations. However, this has yet to be developed formally.

It can be expected that stitch profiles represent DNA cooperativity better than probability profiles, when considering the types of probabilities involved: Correlations between base pairs x_i and x_j can be formulated using conditional probabilities $p(x_i|x_j)$. Conditional probabilities cannot be derived from a probability profile $p(x_i)$ alone, but they can be derived using block probabilities $p(x_i \ldots x_j)$.

Figure 7 also illustrates that some stitches have a *dead end*, that is, a boundary that does not coincide with other stitches' boundaries. A row of stitches can not be continued at a dead end. Stitches with dead ends typically have low peak volumes. Dead ends exist because the continuation of a row splits up in several stitches that all have peak volumes below the cutoff and are therefore not included. It is possible to make stitch profiles without dead ends by replacing the simple cutoff selection with some other appropriate method.

C. The parameters $D_{\max}$ and p_c

We have seen in Fig. 4 how $D_{\max}$ controls the lakes found by the MAXDEEP algorithm. An effect of increasing $D_{\max}$ is to

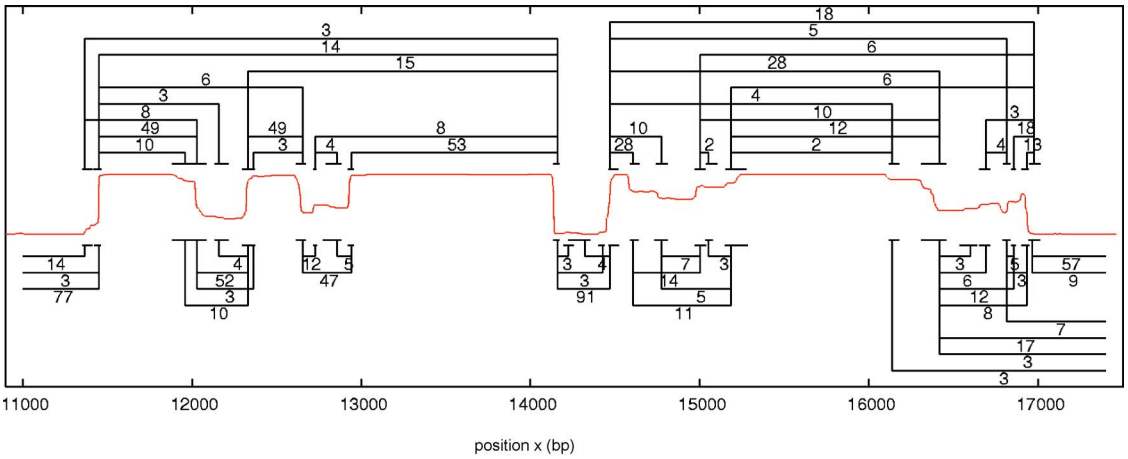

FIG. 7. (Color online) Comparison of a stitch profile and a probability profile, both calculated at $T\simeq 90\ ^\circ\mathrm{C}$ where the helicity is $\Theta=0.1$. The curve in the middle (in red) is the probability profile $1-p_{\mathrm{bp}}(x)$ and it varies between 0 and 1 (vertical axis not shown). Each stitch is labeled with its peak volume p_v in percent. The parameters are $D_{\max}=3$ and $p_c=0.02$ and the sequence window is 11–17.4 kbp.

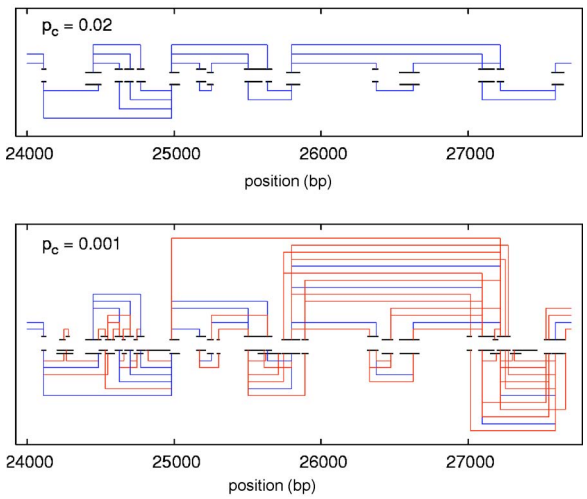

FIG. 8. (Color online) Extra stitches are included when the cutoff p_c is lowered. The first stitch profile ($p_c=0.02$) contains stitches with peak volumes $p_v \geqslant 0.02$. The second stitch profile ($p_c=0.001$) contains the same stitches as the first one (in blue), plus an extra set of stitches with peak volumes $0.001 \leqslant p_v < 0.02$ (in red). The other parameters are $D_{max}=3$ and $T \simeq 80.6\,^{\circ}\mathrm{C}$ and the sequence window is 24–27.7 kbp.

increase the widths of the fluctuation bars (i.e., lakes) in a stitch profile. Another effect has to do with the hierarchical merging of lakes: In Fig. 7, the fluctuation bars correspond to the sloping parts of the probability profile. However, these sloping parts contain smaller plateaus. Decreasing D_{max} can reveal this internal structure, by splitting a stitch into several stitches with smaller fluctuation bars.

Figure 8 illustrates the role of p_c. When p_c is lowered, an extra set of stitches is added to the stitch profile. p_c does not modify the stitches as found by the MAXDEEP algorithm, it only controls how many of them are included in the diagram. The extra stitches represent rare events (low probabilities), and they provide a more finegrained picture like higher order terms in an expansion. The number of stitches in a stitch profile directly depends on p_c, but it also depends on D_{max} indirectly, because the peak volumes depend on the widths of lakes and frames.

The stitch profiles made using this article's methods depend on the values of D_{max} and p_c. The choice of these values depends on what we want to see. One could seek for alternative methods that are parameterfree or only use one parameter by applying, for example, an optimization scheme and some optimality criterion. But in my opinion, reducing the number of parameters would only hide the fact that the peak finding task involves two different types of choice: (1) lumping together related events into a peak and (2) selecting what peaks to include. The cutoff selection method using p_c is simple to program, requires only little computer power, and can reuse data from a stitch profile that has a lower cutoff. Therefore, p_c can be used in practice for "fine-tuning" by trying out different values iteratively. In this way, we can make a stitch profile with a certain total number of stitches. Or a stitch profile with a certain maximum height of the stackings of stitches in the diagram, which would limit the visual complexity (cf. Fig. 8).

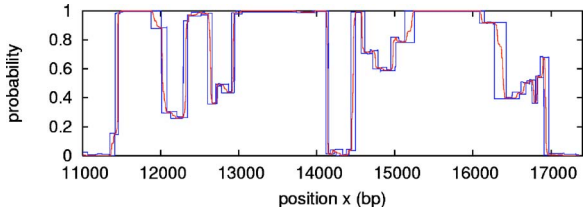

FIG. 9. (Color online) The probability profile $1-p_{bp}(x)$ of Fig. 7 is plotted here (in red) with its upper and lower bounds (in blue), $1-p_{low}(x)$ and $1-p_{up}(x)$, obtained from the stitch profile.

D. Ensemble representation

The alternative conformations represented by a stitch profile are few in numbers compared to the total number 2^N of possible conformations. But they may constitute a considerable fraction of the ensemble in terms of statistical weight. Is this fraction big or small? A direct answer would be obtained by comparing the total partition function Z with the partition function restricted to the stitch profile conformations. Instead, we take a graphical approach that relates peak volumes to base pairing probabilities. A stitch profile provides *upper* and *lower bounds* of the corresponding probability profile. For example, the presence of a loop stitch as in Fig. 5 implies that base pairs in that region are melted with probability greater than the peak volume. Or more precisely, $1-p_{bp}(x) \geqslant p_v(a,b)$ for $L_R(a) < x < L_L(b)$. Stitches that overlap are mutually exclusive, so we can sum over stitches to obtain bounds as follows. Recall that an indicator function is defined as

$$I_{[i,j]}(x) = \begin{cases} 1 \text{ for } x \in [i,j], \\ 0 \text{ otherwise}. \end{cases} \tag{32}$$

Then $p_{low}(x) \leqslant p_{bp}(x) \leqslant p_{up}(x)$, where

$$p_{up}(x) = 1 - \sum_{\text{left tail } a} I_{[1,L_L(a)-1]}(x) p_v(a) - \sum_{\text{loop } (a,b)} I_{[L_R(a)+1,L_L(b)-1]}(x) p_v(a,b) - \sum_{\text{right tail } a} I_{[L_R(a)+1,N]}(x) p_v(a), \tag{33}$$

and

$$p_{low}(x) = \sum_{\text{helix } (a,b)} I_{[L_R(a),L_L(b)]}(x) p_v(a,b). \tag{34}$$

Figure 9 shows the probability profile $1-p_{bp}(x)$ from Fig. 7, together with its two bounds calculated using the stitch profile in Fig. 7 and Eqs. (33) and (34). The three curves are quite close. The two bounding curves consist of vertical and horizontal lines because of the block nature of the stitch profile. Given this constraint, they almost come as close as possible to the probability profile. The probability profile can apparently be reproduced from the stitch profile with only a small error. This suggests that the conformations represented by the stitch profile constitute a majority of the ensemble in terms of statistical weight.

E. Predicting metastable conformations

A depth between zero and $D_{\max}$ is associated with each stitch in a stitch profile. A depth relates to the landscape picture, but it actually characterizes the probabilities of boundary positions: The probability of a boundary in the fluctuation interval $L(a)$ to be located at βa is $10^{D(a)}$ times greater than the probability of being at $L_{\mathrm{L}}(a)$ or $L_{\mathrm{R}}(a)$.

The landscape picture of lakes confined by barriers has been borrowed in this article for the purpose of probability peak finding. This should be distinguished from the usual purpose of describing the nonequilibrium behavior of systems that are "trapped" in metastable states [16]. Nevertheless, an idea of dynamics is implicit when talking about fluctuations. Are the fluctuation bars related to actual fluctuations over time?

In a dynamics interpretation, a 1D lake $L(a)$ describes the ranges of the diffusion of a boundary position on timescale $\tau \propto 10^{D(a)}$. A stitch profile would then predict the ranges of fluctuations that can be observed in an experiment during time $\tau \propto 10^{D_{\max}}$ with an empirical constant of proportionality. However, this timescales interpretation is preliminary until the rates of nucleation events creating new loops, tails, and helices are accounted for. If the depths are related to timescales, a possible application of stitch profiles is to predict metastable conformations that could play an important role in intracellular DNA or other nonequilibrium situations. Stitches that have large depths (i.e., long-lived) and small peak volumes (i.e., rare) are expected to indicate metastable conformations. They can be easily found using a slightly modified cutoff selection method. It is more difficult to detect such conformations in a probability profile because of their low probabilities.

F. Applications

In conclusion, peaks in the block probabilities [Eqs. (1)–(4)] can be found and represented as stitches in a stitch profile. A stitch profile indicates the sizes and locations of loops, tails and helical regions, their probabilities and "depths," and how they fluctuate. Multi-loop conformations can be derived from the alternative rows of stitches, which may show correlations between distant stitches. A stitch profile thus predicts the conformations of partly melted DNA and can account for a majority of the conformational ensemble in terms of the base pairing probabilities.

Stitch profiles are motivated by the general idea that a better prediction of DNA's conformational behavior may contribute to a better understanding of DNA's functional behavior, when there is a structure-function relationship. Strand separation is at the heart of various processes that occur in chromatin and chromosomes [2]. The question is as follows: What role does the sequence dependence of loop stabilities as predicted for DNA melting play in biology? It is reasonable to believe that the low stability of AT-rich regions is important in origins of replication [2] and in transcription initiation [22]. Furthermore, bioinformatic evidence points at more extensive and not yet explained correlations in some genomes between the predicted melting properties and the organization along the sequence of exons, introns, and other genetic elements [19,23–26]. There are different hypotheses. One view is that DNA mainly is digital information storage and that such correlations is a secondary effect reflecting, for example, varying compositions of proteins [27]. Another view is that DNA is also analog and that loop stabilities and/or other sequence dependent biophysical properties of DNA contribute actively in different biological mechanisms. Some more or less speculative examples are: recombination and crossover, sister chromatid adhesion, DNA-protein interactions, and intron insertion [28]. DNA conformational changes in a cell are not driven by temperature changes, but rather by molecular forces and interactions. Why then are DNA melting predictions relevant? The Poland-Scheraga model deals with *in vitro* conditions that are far from the conditions in chromatin: crowding [29] is not accounted for in the loop entropy factor, condensation and protein interactions are missing, and topology and chromosomal geography is not accounted for. Fortunately, the correlations found by Yeramian and others suggest a robustness of the predicted melting properties. Stitch profiles may also apply to intracellular DNA.

As mentioned in the Introduction, the development of stitch profiles is also motivated by new single molecule techniques, in which some of the predicted properties could be measured. For example, it would be interesting to compare with measurements of bubble sizes and their statistical weights [9], positions and stabilities of "tails" [8], and bubble lifetimes [7]. As explained in the previous section, it is an open question how the depths of stitches relate to the lifetimes of the corresponding conformational features.

Stitch profiles may supplement the use of ordinary melting profiles in the design and interpretation of *in vitro* experiments such as gel electrophoresis [30,31] and in the design of probes and primers for PCR and microarrays [32]. For example, stitch profiles emphasize the ensemble aspect and may thereby predict some features of gel experimental data. For short DNAs, however, it is relevant to consider also secondary structure, slippage and mismatches [33,34], that are not accounted for in the Poland-Scheraga model.

A web server for computing stitch profiles has been made available at http://stitchprofiles.uio.no [35].

ACKNOWLEDGMENTS

I thank Edouard Yeramian, Einar Rødland, Eivind Hovig, Enrico Carlon, Fang Liu, Geir Ivar Jerstad, and Tor-Kristian Jenssen.

[1] Loops are here synonymous to *bubbles* and should be distinguished from looping in which the helix as a whole bends back on itself.
[2] A. T. Sumner, *Chromosomes. Organization and Function* (Blackwell, Oxford, 2003).
[3] D. Poland, Biopolymers **13**, 1859 (1974).
[4] E. Yeramian and L. Jones, Nucleic Acids Res. **31**, 3843 (2003).
[5] R. D. Blake and S. G. Delcourt, Nucleic Acids Res. **26**, 3323 (1998).
[6] R. D. Blake, J. W. Bizzaro, J. D. Blake, G. R. Day, S. G. Delcourt, J. Knowles, K. A. Marx, and J. SantaLucia, Bioinformatics **15**, 370 (1999).
[7] G. Altan-Bonnet, A. Libchaber, and O. Krichevsky, Phys. Rev. Lett. **90**, 138101 (2003).
[8] C. Danilowicz, V. W. Coljee, C. Bouzigues, D. K. Lubensky, D. R. Nelson, and M. Prentiss, Proc. Natl. Acad. Sci. U.S.A. **100**, 1694 (2003).
[9] Y. Zeng, A. Montrichok, and G. Zocchi, Phys. Rev. Lett. **91**, 148101 (2003).
[10] A. Hanke and R. Metzler, J. Phys. A **36**, L473 (2003).
[11] T. Hwa, E. Marinari, K. Sneppen, and L. h. Tang, Proc. Natl. Acad. Sci. U.S.A. **100**, 4411 (2003).
[12] M. Peyrard, Nonlinearity **17**, R1 (2004).
[13] E. Tøstesen, F. Liu, T.-K. Jenssen, and E. Hovig, Biopolymers **70**, 364 (2003).
[14] D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers* (Academic Press, New York, 1970).
[15] E. Tøstesen, S.-J. Chen, and K. A. Dill, J. Phys. Chem. B **105**, 1618 (2001).
[16] D. L. Stein and C. M. Newman, Phys. Rev. E **51**, 5228 (1995).
[17] K. H. Hoffmann and P. Sibani, Phys. Rev. A **38**, 4261 (1988).
[18] T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, R44 (1993).
[19] A. Wada and A. Suyama, J. Biomol. Struct. Dyn. **2**, 573 (1984).
[20] See EPAPS Document No. E-PLEEE8-71-148506 for an online view of the stitch profiles of lambda DNA at different temperatures. This document can be reached via a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).
[21] R. Blossey and E. Carlon, Phys. Rev. E **68**, 061911 (2003).
[22] C. H. Choi, G. Kalosakas, K. Ø. Rasmussen, M. Hiromura, A. R. Bishop, and A. Usheva, Nucleic Acids Res. **32**, 1584 (2004).
[23] E. Yeramian, Gene **255**, 139 (2000).
[24] E. Yeramian, Gene **255**, 151 (2000).
[25] E. Yeramian, S. Bonnefoy, and G. Langsley, Bioinformatics **18**, 190 (2002).
[26] C. Bi and C. J. Benham, Bioinformatics **20**, 1477 (2004).
[27] D. Poland, Biopolymers **73**, 216 (2004).
[28] E. Carlon, M. L. Malki, and R. Blossey, q-bio.BM/0409034.
[29] R. J. Ellis, Curr. Opin. Struct. Biol. **11**, 114 (2001).
[30] L. S. Lerman and K. Silverstein, Methods Enzymol. **155**, 482 (1987).
[31] G. Steger, Nucleic Acids Res. **22**, 2760 (1994).
[32] A. Halperin, A. Buhot, and E. B. Zhulina, Biophys. J. **86**, 718 (2004).
[33] R. A. Dimitrov and M. Zuker, Biophys. J. **87**, 215 (2004).
[34] W. Zhang and S.-J. Chen, J. Chem. Phys. **114**, 4253 (2001).
[35] E. Tøstesen, G. I. Jerstad, and E. Hovig, Nucleic Acids Res. (to be published).