# Segmentation of DNA sequences into twostate regions and melting fork regions

**Eivind Tøstesen**[1,2]**, Geir Kjetil Sandve**[3]**, Fang Liu**[4,5] **and Eivind Hovig**[1,3,4]

[1] Department of Medical Informatics, Norwegian Radium Hospital, N-0310 Oslo, Norway
[2] Department of Mathematics, University of Oslo, N-0316 Oslo, Norway
[3] Department of Informatics, University of Oslo, N-0316 Oslo, Norway
[4] Department of Tumour Biology, Institute for Cancer Research, Norwegian Radium Hospital,
N-0310 Oslo, Norway
[5] PubGene AS, Vinderen, N-0319 Oslo, Norway

E-mail: eivindto@math.uio.no

## Abstract
The accurate prediction and characterization of DNA melting domains by computational tools could facilitate a broad range of biological applications. However, no algorithm for melting domain prediction has been available until now. The main challenges include the difficulty of mathematically mapping a qualitative description of DNA melting domains to quantitative statistical mechanics models, as well as the absence of 'gold standards' and a need for generality. In this paper, we introduce a new approach to identify the twostate regions and melting fork regions along a given DNA sequence. Compared with an ad hoc segmentation used in one of our previous studies, the new algorithm is based on boundary probability profiles, rather than standard melting maps. We demonstrate that a more detailed characterization of the DNA melting domain map can be obtained using our new method, and this approach is independent of the choice of DNA melting model. We expect this work to drive our understanding of DNA melting domains one step further.

(Some figures in this article are in colour only in the electronic version)

## 1. Introduction

The organization of *DNA melting domains* is the main characteristic of DNA melting cooperativity. However, a mathematical definition of melting domains and an algorithm for locating them in a given sequence does not exist.

It has been known since the 1970s that DNA melts stepwise, domain after domain, in a reproducible series of subtransitions as temperature is increased [1]. DNA melting domains have also been called cooperatively melting regions (CMR) [2], thermalites [1], cooperative units and isomelting domains [3]. DNA melting domains are regions of tens to hundreds of base pairs, with locations determined by the sequence. The melting of each domain is approximately a *twostate transition*. In the strict sense, 'twostate' means that only two states ('11111' and '00000') are involved, no intermediate states are populated, all base pairs thus open and close in unison and the melting domains are the smallest units

of melting. Short DNAs and oligonucleotides usually consist of only one domain. The stability and melting temperature of a domain not only depends on its internal sequence composition, but also on the stabilities of adjacent domains.

A distinction must be made between domains and bubbles: the melting domains indicate potential melting events, such as the creation, growth or merging of bubbles. Melting domains are regions with fixed boundaries, while helix–coil boundaries fluctuate and depend on temperature. Various approaches to computing bubbles already exist [4–6].

DNA melting can be predicted by various statistical mechanics models and algorithms [7–9]. The formulation of these models involves microscopic interactions and variables at the base pair level, but the concept of melting domains is absent in these formulations. There is no explicit representation of melting domains in the input or the output of these algorithms. Neither are melting domains simply due to certain input parameters, such as a bubble nucleation barrier, but are rather
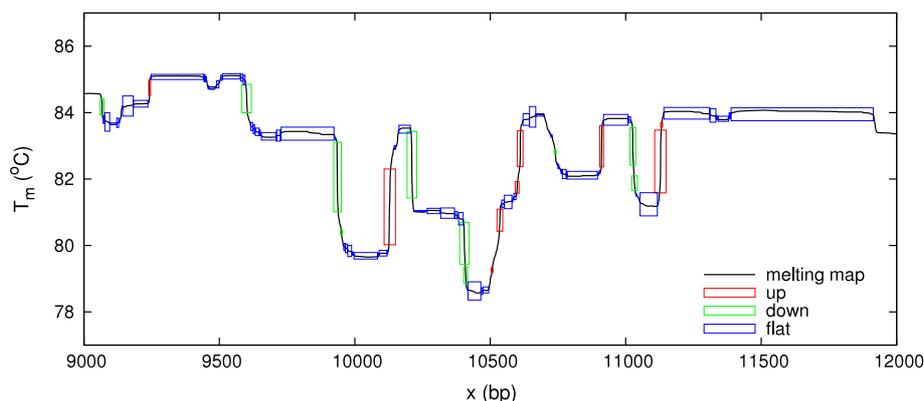
**Figure 1.** The ad hoc segmentation method that was applied to the human genomic melting map [15]. The boxes indicate the up, down and flat regions that are determined by the local slope of the melting map.

emergent properties that depend on all the parameters and the whole sequence. The concept of melting domains comes into play at a postprocessing stage and in the interpretation of the output. For example, plots of melting curves and melting maps usually contain some characteristic features, such as peaks, steps or plateaux, that may be associated with melting domains. In some studies, this association is only qualitative, because a precise definition of melting domains was not made. However, in the analysis and decomposition of optical melting curves, a quantitative relationship between peaks and domains is assumed [10, 11].

The concept of DNA melting domains rests on work by Azbel in the late 1970s [12, 13]. He described a method for determining the ground state of DNA as a function of temperature. His analysis showed that a sequence is partitioned into *ground state domains*. Each ground state domain has a temperature at which the ground state changes from helix to coil. These temperatures define the order in which the ground state changes with increasing temperature. He proposed the ground state versus temperature scenario as a simplified model of the DNA melting process to provide quantitative predictions. The predictive power of his ground state model soon turned out to be limited, with the conclusion that all microstates must be included in the partition function [14, 2]. In spite of that, the ground state model seems to have strongly influenced, if not dictated, the concept of DNA melting domains. In this study, we consider melting domains and ground state domains to be different concepts.

In a previous study, we approached the problem of computing melting domains. In a bioinformatical analysis of the human genomic melting map, we found it useful to extract qualitative features of the curve, such as flat plateaux and steep slopes [15]. We devised an ad hoc method to do such a segmentation of the melting map, illustrated in figure 1, with threshold values that were chosen intuitively. Using this method, we identified melting domains as *flat segments*, based on the assumption that consecutive base pairs melt in unison if they have equal melting temperatures. However, this is not necessarily true. We only know that the reverse implication is true: base pairs share the same melting temperature if they melt together. In this work, we seek a more proper way of doing it.

We consider 'twostateness' to be the defining property of melting domains. Previous tradition has claimed that twostateness can be detected at the calorimetric level and by considering the widths of transitions. But counterexamples have shown that those approaches were error-prone [16, 17]. Instead, we consider the preferred locations of melting forks. We describe a segmentation method based on *boundary probability* profiles. Already in his 1974 paper, Poland discussed the probabilities that a 01 or 10 boundary exists at a specific unit [18]. This seems to have had little impact, however. Most studies have used the standard *probability profiles*—even when melting forks were the main interest [19, 20]. The present method produces segmentations in which melting domains do not cover the whole sequence: the result is not simply a partitioning. Instead, we find that there are *twostate regions* interspersed with *melting fork regions*.

Many cellular functions require local single-strandedness of DNA, for example, DNA replication and gene regulation. A number of DNA-associated proteins contribute to the information capacity of the four-base DNA code [21, 22]. Thus, for instance, nuclear matrix proteins and nucleosomes are part of the chromatin level of information, and participate to modulate transcription factor activity. Ultimately, the DNA sequence itself is insufficient to understand the dynamic interplay of the molecules, as indicated by the performance of prediction algorithms for the binding of these types of molecules along the genomic DNA. These algorithms are at present mostly based on the sequence level of information. In order to advance further in the understanding of these aspects, it appears necessary to incorporate the physical properties of DNA in a comprehensive way, and we here attempt to precisely define the extent of the genomic features of DNA melting.

## 2. Methods

There are a number of alternative DNA melting models being used by various research groups. The present work aims at generality: the concepts and methods should be applicable to each of the DNA melting models. Our approach is formulated solely in terms of equilibrium probabilities that should be computable with any model, without reference to quantities

or properties pertaining to one model only. In some DNA melting models, the state of a base pair is a binary variable (0 or 1), where 0 is the melted/coil/open state and 1 is the bound/helix/closed state, while in other DNA melting models, a state can be a distance and/or angle. In the latter models, however, a criterion for distinguishing single-stranded (ss) and double-stranded (ds) is sometimes applied.

We assume that the bases in the sequence are numbered $1, \ldots, N$ and that each conformation (microstate) may be represented as a chain of 0 and 1 values. At each base position $x$, we define the probability of a 0 or 1:

$$p_0(x) = P(\ldots X \overset{x}{0} X \ldots) \tag{1a}$$

$$p_1(x) = P(\ldots X \overset{x}{1} X \ldots). \tag{1b}$$

In these equations, 1 is a bound base pair (ds), 0 is a melted base pair (ss), X indicates either 0 or 1, and the sequence position $x$ is indicated. At each specific temperature, we can calculate a *probability profile*, which is the $p_1(x)$ value for $1 \leqslant x \leqslant N$. A *melting map* can be interpolated from several probability profiles and indicates at each position the temperature at which $p_1(x) = \frac{1}{2}$. Probability profiles and melting maps have been standard tools in biological applications. The present approach employs another set of probabilities. Each segment of two nearest-neighbour base positions $[x - 1, x]$ has four possible conformations. Accordingly, we consider the probabilities

$$p_{00}(x) = P(\ldots X0 \overset{x}{0} X \ldots) \tag{2a}$$

$$p_{01}(x) = P(\ldots X0 \overset{x}{1} X \ldots) \tag{2b}$$

$$p_{10}(x) = P(\ldots X1 \overset{x}{0} X \ldots) \tag{2c}$$

$$p_{11}(x) = P(\ldots X1 \overset{x}{1} X \ldots) \tag{2d}$$

for $2 \leqslant x \leqslant N$. $p_{01}(x)$ and $p_{10}(x)$ are called *boundary probabilities*. We assume that they can be calculated to obtain *boundary probability profiles* at each temperature. Figure 2 shows an example of 01 and 10 boundary probability profiles compared with the probability profile. By basic probability laws, we have

$$p_0(x) + p_1(x) = 1 \tag{3}$$

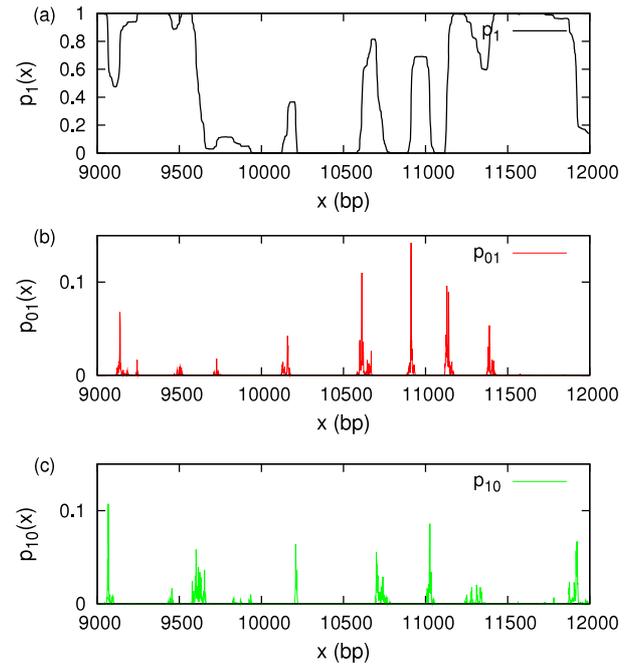$$p_{00}(x) + p_{01}(x) + p_{10}(x) + p_{11}(x) = 1 \tag{4}$$

and

$$p_{00}(x) = p_0(x - 1) p_{0|0}(x) \tag{5a}$$

$$p_{01}(x) = p_0(x - 1) p_{1|0}(x) \tag{5b}$$

$$p_{10}(x) = p_1(x - 1) p_{0|1}(x) \tag{5c}$$

$$p_{11}(x) = p_1(x - 1) p_{1|1}(x). \tag{5d}$$

Here we have introduced the *conditional probabilities*, for example, the probability $p_{1|0}(x)$ that there is a 1 at position $x$ given that there is a 0 at position $x - 1$.



**Figure 2.** The (a) probability profile (in black), (b) 01 boundary probability profile (in red) and (c) 10 boundary probability profile (in green), of the human mitochondrion sequence calculated at $T = 83.678\,°\mathrm{C}$ and $[\mathrm{Na}^+] = 0.075\,\mathrm{M}$ are viewed in the window 9000–12 000 bp as an example. (Colour mnemonic: green = bubble start, red = bubble stop.)

### 2.1. Twostate regions

In this section, we propose a formal definition of twostateness. Usually, the adjective *twostate* indicates a property of transitions, but we will use it to indicate a property of sequence regions. Literally, *twostate* means that only two possible conformations of a region $[x, y]$, the all 1's and all 0's, can exist at any instant and at any temperature. This 'all-or-none' situation may be expressed as

$$P(\ldots X \underbrace{\overset{x}{1} \cdots \overset{y}{1}}_{\text{all 1's}} X \ldots) + P(\ldots X \underbrace{\overset{x}{0} \cdots \overset{y}{0}}_{\text{all 0's}} X \ldots) = 1, \tag{6}$$

or, equivalently, that there is zero probability at all temperatures of the region containing any ds/ss boundaries. However, regions that are strictly twostate do not exist. We expect that there is always a nonzero probability of partially melted intermediates. A more realistic definition must incorporate a small probability of thermally induced 'defects' in a twostate region. We achieve this by means of a threshold $\epsilon > 0$: a region $[x, y]$ is called *globally twostate* if

$$P(\ldots X \underbrace{\overset{x}{1} \cdots \overset{y}{1}}_{\text{all 1's}} X \ldots) + P(\ldots X \underbrace{\overset{x}{0} \cdots \overset{y}{0}}_{\text{all 0's}} X \ldots) \geqslant 1 - \epsilon \tag{7}$$

at all temperatures. The term *globally* refers to the above two probabilities, in which the conformation of the whole region $[x, y]$ is specified. Such probabilities may be computationally challenging. We expect that computing globally twostate regions could involve high algorithmic complexities. For this
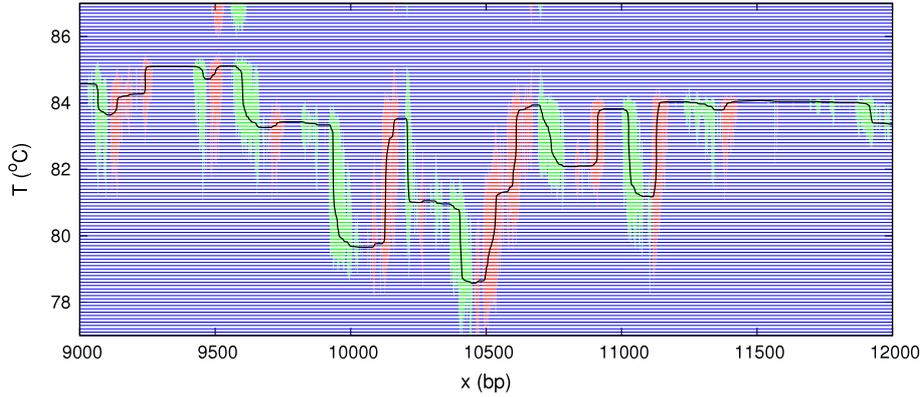
**Figure 3.** At each temperature (in steps of 0.1 °C) and at each position $x$ in the sequence, the colours red, green and blue indicate the presence of a 01 boundary ($p_{01}(x) > \epsilon$ and $p_{10}(x) \leqslant \epsilon$), 10 boundary ($p_{01}(x) \leqslant \epsilon$ and $p_{10}(x) > \epsilon$) or no boundary ($p_{01}(x) \leqslant \epsilon$ and $p_{10}(x) \leqslant \epsilon$) in the parameter-free case: $\epsilon = \epsilon_{\min}$. Boundaries tend to 'live' during temperature intervals up to 4 °C wide. On top of the colour map is plotted the melting map (black curve). Note that the two were calculated independently from each other.

practical reason, we will instead study regions that are *locally* twostate. Equation (7) applied locally to a nearest-neighbour segment at position $x$ can be written

$$p_{11}(x) + p_{00}(x) \geqslant 1 - \epsilon \qquad (8)$$

or, equivalently,

$$p_{01}(x) + p_{10}(x) \leqslant \epsilon. \qquad (9)$$

However, in the following definition we will require each of $p_{01}(x)$ and $p_{10}(x)$, rather than their sum, to be smaller than $\epsilon$. A region $[x_{\text{start}}, x_{\text{end}}]$ is *locally twostate* if for each $x \in [x_{\text{start}} + 1, x_{\text{end}}]$

$$p_{01}(x) \leqslant \epsilon \qquad (10a)$$

$$p_{10}(x) \leqslant \epsilon \qquad (10b)$$

at all temperatures. A region being globally twostate implies that it is locally twostate, but not vice versa. However, if region $[x, y]$ is locally twostate, then (by Boole's inequality)

$$P(\ldots X \underbrace{\overset{x}{1} \cdots \overset{y}{1}}_{\text{all 1's}} X \ldots) + P(\ldots X \underbrace{\overset{x}{0} \cdots \overset{y}{0}}_{\text{all 0's}} X \ldots) \geqslant 1 - 2n\epsilon$$

(11)

at all temperatures, where $n = x_{\text{end}} - x_{\text{start}}$, that is, the region is globally twostate with a less strict threshold value.

### 2.2. Boundary probability segmentation

So far, we have proposed global and local variants of twostate regions, but we have not indicated how to determine the start and end positions of such regions. For globally twostate regions, we leave this as an open question. For locally twostate regions, the starts and ends follow naturally when evaluating at each position from 2 to $N$ whether or not each of the two boundary probabilities is small enough at all temperatures. This yields four possible cases at each position $x$:

(i) $\forall T: p_{01}(x) \leqslant \epsilon$ and $\forall T: p_{10}(x) \leqslant \epsilon$ (twostate)
(ii) $\exists T: p_{01}(x) > \epsilon$ and $\forall T: p_{10}(x) \leqslant \epsilon$ (01 fork)
(iii) $\forall T: p_{01}(x) \leqslant \epsilon$ and $\exists T: p_{10}(x) > \epsilon$ (10 fork)

(iv) $\exists T: p_{01}(x) > \epsilon$ and $\exists T: p_{10}(x) > \epsilon$ (double fork).

By thus classifying each nearest-neighbour segment $[x - 1, x]$, we obtain a segmentation that divides the sequence into four types of regions. For each region $[x_{\text{start}}, x_{\text{end}}]$, one of cases (i)–(iv) is true at all internal positions $x \in [x_{\text{start}} + 1, x_{\text{end}}]$, but not true at the flanking positions $x_{\text{start}}$ and $x_{\text{end}} + 1$. A case (i) region is locally twostate, hereafter referred to as a *twostate region*. In a case (ii) region, there are large probabilities of 01 boundaries at some temperatures, while the 10 boundary probabilities always remain small. This is called a *01 fork region*. Likewise, in a case (iii) region, there are large probabilities of 10 boundaries at some temperatures, while the 01 boundary probabilities remain small always. This is called a *10 fork region*. The latter two types of regions are called *melting fork regions*. The last case, (iv), is a region in which there may be large probabilities of both 01 and 10 boundaries. This is called a *double fork region*.

Figure 3 illustrates that each point $(x, T)$ in the plane can be classified and given a colour according to the comparison of the two boundary probabilities with $\epsilon$. In this plot, twostate regions correspond to entirely blue vertical columns, while melting fork regions are columns that contain red or green 'islands'. Figure 3 also shows how some of the islands correspond to the steep slopes on the melting map.

### 2.3. Sampling algorithm

An algorithm for finding twostate regions and melting fork regions must evaluate statements of the type 'for all temperatures'. An approximate method is to sample only a finite, but representative, set of temperatures, for example, by scanning a range from $T_{\text{low}}$ to $T_{\text{high}}$ with incremental steps $\Delta T$. We use values $T_{\text{low}}$ and $T_{\text{high}}$ corresponding to helicities $\theta = 0.999$ and $0.001$, so as to cover most of the subtransitions. This usually means $T_{\text{high}} - T_{\text{low}} \approx 20$ °C. The choice of step size is a trade-off. The smaller the $\Delta T$, the more temperatures are taken into account and the more accurate is the segmentation, but at the cost of longer computation time. Melting forks only exist over a temperature interval (figure 3), so the resolution
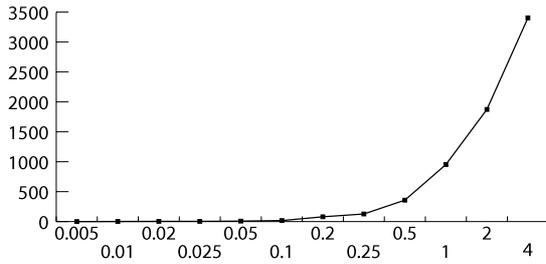
**Figure 4.** Convergence test for the sampling algorithm, using a segmentation based on a step size of 0.005 °C as a reference. Different choices of step size are given on the *x* axis, with the *y* axis denoting the Hamming distance between resulting segmentations and the reference segmentation.

must be high enough to detect the shortest of such temperature intervals.

Figure 4 shows how the segmentation depends on sampling resolution. A segmentation based on a very small step size of 0.005 °C is used as a reference and the figure shows the Hamming distance between this reference and segmentations based on larger step size. The Hamming distance measures the number of positions that differs with respect to the class (twostate and melting forks) between two compared segmentations.

As can be seen from the figure, the use of step sizes above 1 °C alters the resulting segmentation substantially, while the segmentation converges for step sizes below 0.1 °C. Based on this analysis we have chosen $\Delta T = 0.1$ °C as standard, since this gives a segmentation almost identical to segmentations achieved with smaller step sizes and longer computational time.

### 2.4. Determining the $\epsilon$ value

The parameter $\epsilon$ is the threshold used for distinguishing large from small boundary probabilities. There are two ways of determining its value: (1) as an arbitrary input to the algorithm and (2) as a parameter-free output of the algorithm.

Thermal fluctuations create boundaries anywhere in the sequence, which imposes a background noise level in the boundary probability profiles. On the other hand, the sequence encodes the preference for certain boundary locations, which gives rise to high peaks above the background noise (figure 2). The parameter-free $\epsilon$ value separates the peaks from the noise. Figure 5 shows how the segmentation depends on $\epsilon$. At $\epsilon = 0.1$, the boundary probabilities are considered small almost everywhere, and there are no melting fork regions. At the other extreme, $\epsilon < 0.00001$, the boundary probabilities are considered large almost everywhere, there are no twostate regions and most of the sequence is considered to be a double fork. We interpret the double fork regions as a reflection of the background noise. We define the parameter-free $\epsilon_{\min}$ as the value above which the double fork regions disappear. The $\epsilon_{\min}$ depends on the sequence and can be written

$$\epsilon_{\min} = \max_T[\max_x(\min\{p_{01}(x), p_{10}(x)\})]. \quad (12)$$
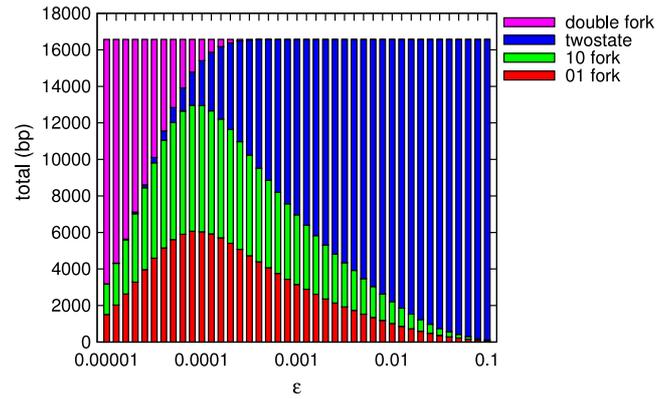


**Figure 5.** Stacked histogram showing the total amounts of twostate regions (blue), 01 fork regions (red), 10 fork regions (green) and double fork regions (magenta) in a 16571 bp sequence, plotted over a range of $\epsilon$ values (logarithmic axis). The double fork regions disappear at $\epsilon_{\min} \approx 0.0006$.

To avoid detecting thermal noise, one should use an $\epsilon \geqslant \epsilon_{\min}$. In parameter-free segmentations, we do a sampling over a temperature interval $[T_{\text{low}}, T_{\text{high}}]$ to determine the $\epsilon_{\min}$ by (12). Figure 5 shows that the amounts of twostate, 01 fork and 10 fork regions are about a third each at $\epsilon_{\min}$.

### 2.5. Approximations derived from probability profiles and melting maps

Figure 2 shows a correspondence between a probability profile and the two boundary probability profiles: steep increases in $p_1(x)$ coincide with peaks in $p_{01}(x)$, while steep decreases coincide with peaks in $p_{10}(x)$. This suggests that the twostate, 01 fork and 10 fork regions can be estimated approximately from standard probability profiles alone.

For $2 \leqslant x \leqslant N$ we can define the *slope* of a probability profile as the difference:

$$\Delta p_1(x) = p_1(x) - p_1(x-1). \quad (13)$$

By

$$p_1(x) = p_{01}(x) + p_{11}(x) \quad (14a)$$

$$p_1(x-1) = p_{10}(x) + p_{11}(x) \quad (14b)$$

we derive

$$\Delta p_1(x) = p_{01}(x) - p_{10}(x). \quad (15)$$

This equation explains the correspondence in figure 2 between the slope in the probability profile and the peaks in the boundary probabilities. It also shows that there is a loss of information when two boundary probability profiles are 'collapsed' into the probability profile. Instead of classifying each nearest-neighbour segment $[x - 1, x]$ according to cases (i)–(iv) above, we could do the following classification:

$$\text{class}(x) = \begin{cases} \text{up} & \text{if } \exists T : \Delta p_1(x) > \epsilon \\ \text{flat} & \text{if } \forall T : |\Delta p_1(x)| \leqslant \epsilon \quad (16) \\ \text{down} & \text{if } \exists T : \Delta p_1(x) < -\epsilon. \end{cases}$$

(Note that this segmentation of the probability profiles into up, down and flat regions is not the same as the segmentation of

the melting map.) From (16) and (15) it follows that each twostate region is contained inside a flat region. Furthermore, if $\epsilon \geqslant \epsilon_{\min}$ (no double forks), then each up region is contained inside a 01 fork region and each down region is contained inside a 10 fork region. Used as an approximation, (16) would misclassify some (parts of) the melting fork regions as being twostate.

We have not analytically solved how these results translate to the melting map segmentation [15], which has the form

$$\text{class}(x) = \begin{cases} \text{up} & \text{if } \Delta T_{\mathrm{m}}(x) > \epsilon_1 \\ \text{flat} & \text{if } |\Delta T_{\mathrm{m}}(x)| \leqslant \epsilon_2 \\ \text{down} & \text{if } \Delta T_{\mathrm{m}}(x) < -\epsilon_1 \\ \text{none} & \text{otherwise} \end{cases} \quad (17)$$

where $\Delta T_{\mathrm{m}}(x) = T_{\mathrm{m}}(x) - T_{\mathrm{m}}(x-1)$. Note that this has two thresholds: $\epsilon_1$ and $\epsilon_2$. The melting map segmentation with $\epsilon_1 = 0.13$ and $\epsilon_2 = 0.01$ is illustrated in figure 1.

### 2.6. Subtransitions and fork–fork types

Once a segmentation is produced, it is possible to distinguish different types of twostate regions, called *fork–fork* types, by identifying the nearest flanking or bracketing melting fork region on each side. If a twostate region is close to the sequence end (5′ or 3′), then it is possible that no melting fork regions exist before that end. On each side, therefore, there are three possibilities (01, 10, either 5′ or 3′), which combine into nine possible fork–fork types of twostate region. We write the fork–fork types in the notation 01_10, 10_10, 5′_10, 01_3′, etc. The nine fork–fork types are listed in table 1.

In the human genomic melting map, we divided the flat segments into four types called *top, bottom, upstair and downstair* [15]: we compared the melting map temperature averaged over a flat segment with the melting map temperatures averaged over the two flanking regions of equal length. The flat segment is a top if both flanking temperatures are lower, it is a bottom if both are higher, it is an upstair if the left-hand side is lower and the right-hand side is higher, and downstair otherwise. Table 1 shows the correspondence between the four melting map types and four of the fork–fork types. Because of the lengths of the chromosomes, we did not consider the other five types of flat segment at the sequence ends. We found that the type of a flat segment (top and bottom) has a strong effect on its melting temperature, that is, the stability depends on neighbouring domains, and not only on the internal GC content [15].

These findings were in accordance with Azbel's prediction of the deviation of a domain's $T_{\mathrm{m}}$ from the GC%-based Marmur–Doty prediction [12, 10, 2, 1]. He distinguished five types of subtransition of a melting domain by the number of boundaries they create: (I) nucleation of a bubble, (II) unzipping from an end, (III) internal growth of a bubble, (IV) merging of a bubble and an unzipping end, and (V) merging of two bubbles.

Table 1 shows that the fork–fork classification corresponds to distinguishing the left and right variants (mirror images) of the subtransition types II, III and IV. The 5′_3′ type corresponds

**Table 1.** Twostate regions are divided into nine *fork–fork* types. Four of these correspond to *meltmap* types of flat regions: top, bottom, upstair and downstair [15]. The *subtransition* indicates the most likely melting process of each region [10, 23], based on Azbel's distinction according to the change $n_{\mathrm{b}}$ in the number of boundaries [12]. The *effect on* $T_{\mathrm{m}}$ is the deviation from the expectation based on GC% content [12].

| Fork–fork | Meltmap | Subtransition | $n_{\mathrm{b}}$ | Effect on $T_{\mathrm{m}}$ |
|-----------|---------|---------------|------|------------------|
| 01_10 | Top | V | $-2$ | $T_{\mathrm{m}} < T_{\mathrm{GC\%}}$ |
| 10_01 | Bottom | I | $+2$ | $T_{\mathrm{m}} > T_{\mathrm{GC\%}}$ |
| 01_01 | Upstair | III | $0$ | $T_{\mathrm{m}} \approx T_{\mathrm{GC\%}}$ |
| 10_10 | Downstair | III | $0$ | $T_{\mathrm{m}} \approx T_{\mathrm{GC\%}}$ |
| 5′_10 | — | IV | $-1$ | $T_{\mathrm{m}} < T_{\mathrm{GC\%}}$ |
| 5′_01 | — | II | $+1$ | $T_{\mathrm{m}} \approx T_{\mathrm{GC\%}}$ |
| 01_3′ | — | IV | $-1$ | $T_{\mathrm{m}} < T_{\mathrm{GC\%}}$ |
| 10_3′ | — | II | $+1$ | $T_{\mathrm{m}} \approx T_{\mathrm{GC\%}}$ |
| 5′_3′ | — | — | $0$ | $T_{\mathrm{m}} \approx T_{\mathrm{GC\%}}$ |

to oligonucleotides that do not contain any melting fork regions, and melt and dissociate in a single step. The table summarizes the expected relationships between the different types of regions and transitions. For example, when a 01_3′ region melts, it will most likely open to merge with an existing bubble to the left and with the 3′ end to the right. Otherwise, it would create a 10 fork to the left and/or a 01 fork to the right, in contradiction with its type being 01_3′. The 01_3′ region is destabilized by the flanking bubble on the left, lowering its $T_{\mathrm{m}}$.

## 3. Results

The methodology for determining melting domains is in itself the main result of this study. Although melting domains have been used as a concept for a long time, this paper presents a principled approach to calculation of the domains. With the melting segmentation produced by the methodology as a proposed reference, we here analyse melting segmentation of biological DNA sequences and compare our proposed reference to the previous ad hoc method for the prediction of melting domains.

We first give a direct visualization of melting domains in a selected portion of the mitochondrion, and then provide some statistical results from the full human chromosome 21 segmentation.

Figure 6 shows the segmentation from position 8250 to 8750 of the mitochondrial genome. Black bars denote twostate regions, while areas between bars denote melting fork regions. As can be seen from the figure, the segmentation typically forms several tight clusters of twostate regions. Very short melting fork regions are typically occurring between twostate regions within each such cluster, while one or a few long melting fork regions are separating the clusters. Also, there are a few very short, isolated twostate regions.

The segmentation is further compared to the melting map segmentation [15]. Flat regions of this segmentation should correspond to twostate regions of the new segmentation and are denoted in black. Non-flat regions correspond to melting fork regions, with 'up' regions corresponding to '01' forks and 'down' regions corresponding to '10' forks. As can be seen from the figure, the ad hoc procedure used in the previous
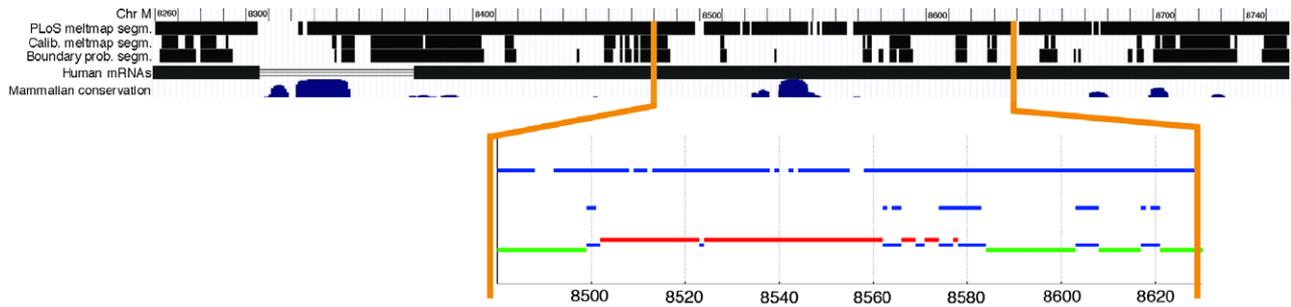
**Figure 6.** Visualization of twostate and melting fork regions in the UCSC genome browser. Position 8250–8750 of the mitochondrial genome is shown. Also shown is a genomic type track, in this case evolutionary conservation of the genomic sequence. Although there is no direct information to be obtained from this small segment of DNA, these two features, or other genomic features, may be compared to the information provided by the segmentations, in order to uncover hidden dependences. A zoomed-in version is also shown for a part of the segmentation, with twostate regions denoted in blue, and '01' and '10' forks further specified in red and green, respectively.
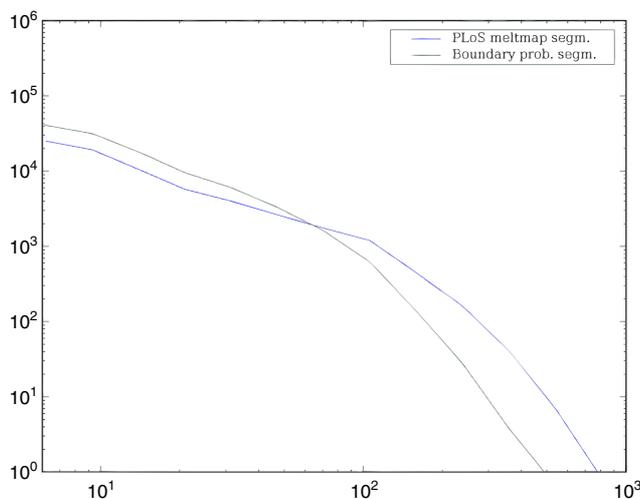


**Figure 7.** Length distribution of twostate regions in human chromosome 21. This is shown using logarithmic scales, with segment length between 10 and 1000 on the *x* axis, and the count of segments per exact length value on the *y* axis. Smoothed curves are shown for both the PLoS meltmap segmentation and the boundary probability segmentation.
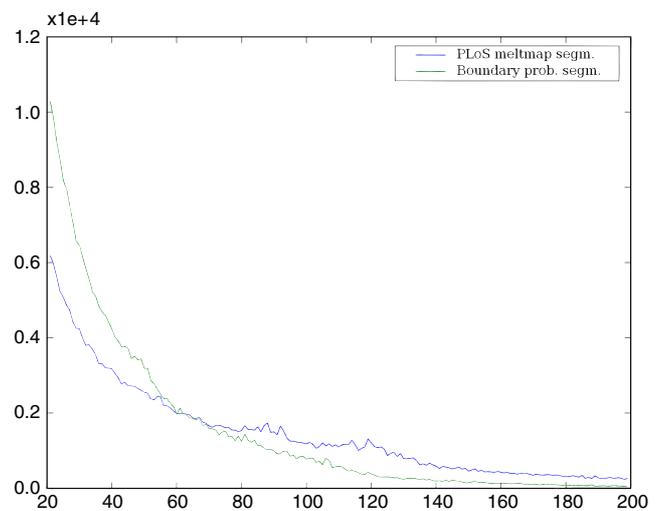


**Figure 8.** Length distribution of twostate regions in human chromosome 21. This is shown using plain scales, with segment length between 20 and 200 on the *x* axis, and the count of segments per exact length value on the *y* axis. Curves are shown for both the PLoS meltmap segmentation and the boundary probability segmentation.

paper gave rise to a much larger amount of flat regions. The ad hoc segmentation typically gave very long flat regions, only separated by short non-flat regions.

As twostate (flat) regions covered a much larger part of the genome according to the previous segmentation procedure, we also constructed a third segmentation. This segmentation was based on the melting map segmentation (17), but with parameter values $\epsilon_1 = 0.13$ and $\epsilon_2 = 0.0002$ calibrated to give similar global amounts of twostate regions and flat segments. This calibrated melting map segmentation resembles the new segmentation, although some minor differences are noticeable. For instance, where one segmentation assigns a continuous twostate region, the other segmentation sometimes assigns multiple clustered regions.

The statistical properties of the segmentations are further analysed on human chromosome 21. Figure 7 shows a log–log plot of the length distributions of twostate/flat regions for both the new and the previously proposed segmentation method.

Most regions are short, with relatively few regions being longer than 100 bp. For lengths of around 1000 bp, there is on average less than one region of each exact length value.

Figure 8 further shows the detailed distribution of twostate regions of length 20–200 using plain scales. From both figures it is apparent that the twostate regions of the new segmentation are generally shorter than for the previously proposed melting map segmentation.

For the full chromosome 21, there are 1.3 million twostate regions according to the new segmentation, with an average length of 18. This compares with 0.5 million flat (twostate) regions for the previous melting map segmentation, with an average length of 55.

## 4. Discussion and conclusion

In this work, we have developed a segmentation of DNA sequences that is based on detection of local twostateness.

Azbel's ground state segmentation produced domains that partitioned or 'tiled' the sequence [12]. His approach ignores any excited states. In contrast, our approach takes the ensemble into account, with the consequence that melting fork regions appear in between the twostate regions. Melting fork regions occupy a quite large fraction of the sequence (figure 5). We found that the average lengths of twostate regions are 18 bp in human chromosome 21, while the average length of flat regions in the melting map was 55 bp [15]. This is an order of magnitude smaller than previously reported lengths of melting domains [24].

These findings may have important implications for how to utilize DNA melting in a genomic context, in defining the extent of local DNA single-strandedness of a genome, and how it may influence the informational context upon which other molecules act.

A by-product of the methodology is the characterization of the fluctuational noise level in terms of the sequence-dependent $\epsilon_{\min}$. From a few sequences studied so far, we have computed values in the range $10^3$–$10^4$. Furthermore, we have found this noise level to vary by less than one order of magnitude over the melting range of temperatures, being roughly the same at high and low temperatures.

Melting fork regions have a preferred direction: either 01 or 10. Although a few double fork regions may actually be physically meaningful, we suggest avoiding double fork regions by using $\epsilon \geqslant \epsilon_{\min}$, in order not to mistake noise for a signal.

We have aimed at generality with respect to the statistical mechanical DNA melting model. But so far we have only applied the methodology to the Poland–Scheraga model. It remains to be investigated how results would depend on using other models, and if it would be computationally viable at all. It could be a way of comparing different DNA melting models to compute a segmentation with each.

The generality criterion imposed some algorithmic restrictions, for example, that we could not exploit properties of the Poland–Scheraga partition functions or the nearest-neighbour stability parametrization. It is possible that a specialized treatment would be able to exploit such detailed knowledge to obtain much more efficient segmentation algorithms.

The temperature scanning algorithm is simple and has algorithmic complexity $O(N)$, but we found that small step sizes ($\Delta T = 0.1\,°C$) are needed to obtain sufficient accuracy. Unfortunately, this means long computation times for human genomic sequences. In this work, we have not concentrated further on algorithm optimization.

As a possible application, the boundary probability segmentation may provide annotations of all the twostate regions and melting fork regions in the human genome, as we did in chromosome 21. *Regions* is a type of annotation that is well accommodated by existing genome browsers. For such applications, it would be useful to add an attribute to the twostate regions indicating their (average) melting temperature. Compared with a standard melting map, such a *melting domain map* emphasizes more the essential features of cooperativity, while reflecting to a lesser degree the variation in GC content.

## Appendix. Calculation of boundary probabilities in the Poland–Scheraga model

The two boundary probabilities can be computed from partition functions in the Poland–Scheraga model as follows:

$$p_{01}(x) = \frac{Z_{X01}(x)Z_{01X}(x)}{Z} \tag{A.1}$$

$$p_{10}(x) = \frac{Z_{X10}(x-1)Z_{10X}(x-1)}{Z}. \tag{A.2}$$

$Z_{X01}(x)$ and $Z_{X10}(x-1)$ are partition functions of the segment $[1, x]$, $Z_{01X}(x)$, $Z_{10X}(x-1)$ are partition functions of the segment $[x-1, N]$ and $Z$ is the total partition function of the whole chain [25]. All these partition functions are computed in our previously described algorithm [26]. In the language of [26]

$$p_{01}(x) = \frac{U_{01}^{LR}(x)V_{10}^{RL}(N+2-x)}{\beta Q_{\text{total}}} \tag{A.3}$$

$$p_{10}(x) = \frac{V_{10}^{LR}(x)U_{01}^{RL}(N+2-x)}{\beta Q_{\text{total}}}. \tag{A.4}$$

Alternatively, one can use the Poland–Fixman–Freire algorithm [18, 27]. This algorithm is highly optimized for the purpose of computing $p_1(x)$ profiles. The PFF algorithm first computes all the conditional probabilities $p_{1|1}(x)$ (forward sweep) and then all the $p_1(x)$ (backward sweep). From these two arrays, we readily obtain the boundary probabilities. By (5c) we derive

$$p_{10}(x) = p_1(x-1)p_{0|1}(x) \tag{A.5}$$

$$= p_1(x-1) - p_1(x-1)p_{1|1}(x). \tag{A.6}$$

By (15) and (A.6) we derive

$$p_{01}(x) = p_1(x) - p_1(x-1)p_{1|1}(x). \tag{A.7}$$

## References

[1] Wartell R M and Benight A S 1985 Thermal denaturation of DNA molecules: a comparison of theory with experiment *Phys. Rep.* **126** 67–107

[2] Gotoh O 1983 Prediction of melting profiles and local helix stability for sequenced DNA *Adv. Biophys.* **16** 1–52

[3] Ekstrøm P O, Bjørheim J and Thilly W G 2007 Technology to accelerate pangenomic scanning for unknown point mutations in exonic sequences: cycling temperature capillary electrophoresis (CTCE) *BMC Genet.* **8** 54

[4] Tøstesen E 2008 A stitch in time: efficient computation of genomic DNA melting bubbles *Algorithms Mol. Biol.* **3** 10

[5] Ambjörnsson T, Banik S K, Krichevsky O and Metzler R 2007 Breathing dynamics in heteropolymer DNA *Biophys. J.* **92** 2674–84

[6] van Erp T S, Cuesta-Lopez S, Hagmann J-G and Peyrard M 2005 Can one predict DNA transcription start sites by studying bubbles? *Phys. Rev. Lett.* **95** 218104

[7] Poland D and Scheraga H A 1970 *Theory of Helix-coil Transitions in Biopolymers* (New York: Academic)

[8] Dauxois T, Peyrard M and Bishop A R 1993 Entropy-driven DNA denaturation *Phys. Rev.* E **47** R44–7

[9] Benham C J and Bi C 2004 The analysis of stress-induced duplex destabilization in long genomic DNA sequences *J. Comput. Biol.* **11** 519–43

[10] Yen W S and Blake R D 1980 Analysis of high-resolution melting (thermal dispersion) of DNA *Methods Biopolym.* **19** 681–700

[11] Blake R D and Delcourt S G 1998 Thermal stability of DNA *Nucleic Acids Res.* **26** 3323–32

[12] Azbel M Y 1980 DNA sequencing and helix-coil transition. I. Theory of DNA melting *Biopolymers* **19** 61–80

[13] Azbel M Y 1980 DNA sequencing and helix-coil transition. II. Loop entropy and DNA melting *Biopolymers* **19** 81–93

[14] Lyubchenko Yu L, Kalambet Yu A, Lyamichev V I and Borovik A S 1982 A comparison of experimental and theoretical melting maps for replicative form of $\varphi$X174 DNA *Nucleic Acids Res.* **10** 1867–76

[15] Liu F, Tøstesen E, Sundet J K, Jenssen T-K, Bock C, Jerstad G I, Thilly W G and Hovig E 2007 The human genomic melting map *PLoS Comput. Biol.* **3** e93

[16] Zhou Y, Hall C K and Karplus M 1999 The calorimetric criterion for a two-state process revisited *Protein Sci.* **8** 1064–74

[17] Bakk A 2004 Is it always possible to distinguish two- and three-state systems by evaluating the van't Hoff enthalpy? *Phys. Biol.* **1** 152–8

[18] Poland D 1974 Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations *Biopolymers* **13** 1859–71

[19] Carlon E, Malki M L and Blossey R 2005 Exons, introns, and DNA thermodynamics *Phys. Rev. Lett.* **94** 178101

[20] Carlon E, Dkhissi A, Malki M L and Blossey R 2007 Stability domains of actin genes and genomic evolution *Phys. Rev.* E **76** 051916

[21] von Hippel P H 2007 From simple DNA–protein interactions to the macromolecular machines of gene expression *Annu. Rev. Biophys. Biomol. Struct.* **36** 79–105

[22] Chaires J B 2008 Allostery: DNA does it, too *ACS Chem. Biol.* **3** 207–9

[23] Blake R D, Bizzaro J W, Blake J D, Day G R, Delcourt S G, Knowles J, Marx K A and SantaLucia J Jr 1999 Statistical mechanical simulation of polymeric DNA melting with MELTSIM *Bioinformatics* **15** 370–5

[24] Blake R D 1987 Cooperative lengths of DNA during melting *Biopolymers* **26** 1063–74

[25] Tøstesen E 2005 Partly melted DNA conformations obtained with a probability peak finding method *Phys. Rev.* E **71** 061922

[26] Tøstesen E, Liu F, Jenssen T-K and Hovig E 2003 Speed-up of DNA melting algorithm with complete nearest neighbor properties *Biopolymers* **70** 364–76

[27] Fixman M and Freire J J 1977 Theory of DNA melting curves *Biopolymers* **16** 2693–704