

Research

Open Access

A stitch in time: Efficient computation of genomic DNA melting bubbles

Eivind Tøstesen^{1,2}

Address: ¹Department of Tumor Biology, Norwegian Radium Hospital, N-0310, Oslo, Norway and ²Department of Mathematics, University of Oslo, N-0316, Oslo, Norway

Email: Eivind Tøstesen - eivindto@math.uio.no

Published: 17 July 2008

Received: 1 February 2008

Algorithms for Molecular Biology 2008, **3**:10 doi:10.1186/1748-7188-3-10

Accepted: 17 July 2008

This article is available from: <http://www.almob.org/content/3/1/10>

© 2008 Tøstesen; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: It is of biological interest to make genome-wide predictions of the locations of DNA melting bubbles using statistical mechanics models. Computationally, this poses the challenge that a generic search through all combinations of bubble starts and ends is quadratic.

Results: An efficient algorithm is described, which shows that the time complexity of the task is $O(N \log N)$ rather than quadratic. The algorithm exploits that bubble lengths may be limited, but without a prior assumption of a maximal bubble length. No approximations, such as windowing, have been introduced to reduce the time complexity. More than just finding the bubbles, the algorithm produces a stitch profile, which is a probabilistic graphical model of bubbles and helical regions. The algorithm applies a probability peak finding method based on a hierarchical analysis of the energy barriers in the Poland-Scheraga model.

Conclusion: Exact and fast computation of genomic stitch profiles is thus feasible. Sequences of several megabases have been computed, only limited by computer memory. Possible applications are the genome-wide comparisons of bubbles with promoters, TSS, viral integration sites, and other melting-related regions.

Background

Models of DNA melting make it possible to compute what regions that are single-stranded (ss) and what regions that are double-stranded (ds). Based on statistical mechanics, such model predictions are probabilistic by nature. Bubbles or single-stranded regions play an essential role in fundamental biological processes, such as transcription, replication, viral integration, repair, recombination, and in determining chromatin structure [1,2]. It is therefore interesting to apply DNA melting models to genomic DNA sequences, although the available models so far are limited to *in vitro* knowledge. Genomic applications began around 1980 [3,4], and have been gaining momentum over the years with the increasing availability of

sequences, faster computers, and model development. It has been found that predicted ds/ss boundaries often are located at or very close to exon-intron junctions, the correspondence being stronger in some genomes than others [5-9], which suggested a gene finding method [10]. In the same vein, comparisons of actin cDNA melting maps in animals, plants, and fungi suggested that intron insertion could have target the sites of such melting fork junctions in ancient genes [11,12]. In other studies, bubbles in promoter regions were computed to test the hypothesis that the stability of the double helix contributes to transcriptional regulation [13-18]. The role of TATA bubbles and their lifetimes has been further discussed using a stochastic model of dynamics based on single molecule experi-

ments [19,20]. Bubbles induced by superhelicity have also been found to correlate with replication origins as well as promoters [21-24]. In addition to the testing of specific hypotheses, a strategy has been to provide whole genomes with annotations of their melting properties [25,26]. Combined with all other existing annotations, such melting data allow exploratory data mining and possibly to form new hypotheses [27]. For example, the human genomic melting map was made available, compared to a wide range of other annotations, and was shown to provide more information than the local GC content [26].

In the genomic studies, various melting features have proved to be of particular interest. These include the bubbles and helical regions, bubble nucleation sites, cooperative melting domains, melting fork junctions, breathers, sites of high or low stability, and SIDD sites. Most often we want to know their locations, but additional information is sometimes useful, such as probabilities, dynamics, stabilities, and context. DNA melting models based on statistical mechanics are powerful tools for calculating such properties, especially those models that can be solved by dynamical programming in polynomial time. For many features of interest, however, algorithms remain to be developed to do such predictions. The existing melting algorithms typically produce melting profiles of some numerical quantity for each sequence position. The prototypical example is Poland's probability profile [28], but also profiles of melting temperatures (melting maps), free energies or other quantities are computed per basepair.

The result can be plotted as a curve, while the wanted features often have the format of regions, junctions and other sites. Some genomics data mining tools also require data in these formats rather than curves. As a remedy, melting profiles have been subjected to *ad hoc* post-processing methods to extract the wanted features, such as segmentation algorithms [26], thresholding [25], and relying on the eye through visualization [9,12].

In previous work, we developed an algorithm that identifies regions of four types: helical regions, bubbles (internal loops), and unzipped 5' and 3' end regions (tails) [29-31]. The algorithm produces a *stitch profile*, which is a probabilistic graphical model of DNA's conformational space. A stitch profile contains a set of regions of the four types. Each region is called a *stitch*, because of the way they can be connected in paths. The stitch profile algorithm computes the location (start and end) of each stitch and the probability of that region being in the corresponding state (ds or ss) at the specified temperature. A stitch profile can be plotted in a *stitch profile diagram*, as illustrated in Figure 1. The location of a bubble or helix stitch is not given as a precise coordinate pair (x, y) , but rather as a pair of ds/ss boundaries with fuzzy locations. For each ds/ss boundary, the range of thermal fluctuations is computed and given as an interval. A stitch profile indicates a number of alternative configurations, both optimal and suboptimal, as illustrated in Figure 1. In contrast, a melting map would indicate the single configuration at each

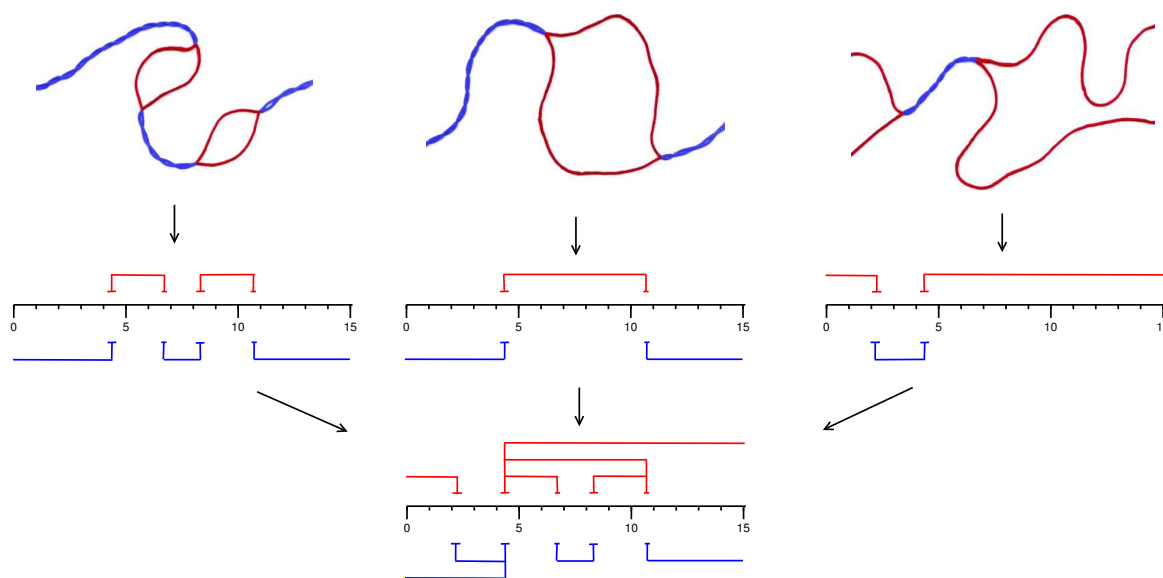


Figure 1

What is a stitch profile diagram? At the top are sketched three alternative DNA conformations at the same temperature. In the middle diagrams, the sequence location of each helical region (blue) and each bubble or single-stranded region (red) is represented by a stitch. At the bottom, the three "rows of stitches" are merged into a stitch profile diagram.

temperature, in which each basepair is in its most probable state.

A stitch profile thus provides some features, e.g. bubbles, that would be of interest in genomic analyses. However, the previously described algorithm for computing stitch profiles [29] has time complexity $O(N^2)$. Genomics studies often require faster algorithms, both to compute long sequences and to compute many sequences. In this paper, therefore, an efficient stitch profile algorithm with time complexity $O(N \log N)$ is described, and the prospects of computing genomic stitch profiles are discussed. The original algorithm [29] is referred to as Algorithm 1, while the new algorithm is referred to as Algorithm 2.

The reduction in time complexity has been achieved without introducing any approximation or simplification such as windowing. The usual tradeoff between speed and precision is therefore not involved here. The output of Algorithm 2 is not of a lower quality, but identical to Algorithm 1's output. Algorithm 1 was simply inefficient. However, it was not obvious that this problem has time complexity $O(N \log N)$, which is the same as computing melting profiles with the Poland-Fixman-Freire algorithm [32]. It would appear that the stitch profile had greater complexity, for example, that the search for all bubble starts and ends would be quadratic. On the other hand, we know that bubbles may be small compared to the sequence length. Algorithm 2 detects such circumstances in an adaptive way, without assuming a maximal bubble length.

Methods

The proper way of computing DNA conformations, as well as other macromolecular structures, is to consider a *rugged landscape* [33,34]. As an abstract mathematical function, a landscape applies to widely different complex systems, for example, fitness landscapes in evolutionary biology for defining populations and species. The ruggedness implies many local maxima and minima on many levels. In optimization, the task would be to avoid all the "false" local optima and find the global optimum. That is not what we want. On the contrary, we would prefer to include most of them.

A local optimum corresponds to an instantaneous conformation or *microstate* that is more fit or stable than its immediate neighbors. However, fluctuations over time cover a larger area in the landscape around the local optimum, which is defined as a *macrostate*. A macrostate can not simply be associated with a local optimum, because it usually covers many local optima. On the other hand, a local optimum may be part of different macrostates. Fluctuations are biologically important, as they represent stability and robustness, rather than noise and uncertainty

[35]. Conformations are properly represented by macrostates, not microstates. We want to characterize the whole landscape of DNA conformations by a set of macrostates.

More specifically, this article considers certain probability landscapes, in which the probability peaks are the macrostates. The algorithmic task is to find a set of peaks. Automatic peak detecting is applied in various kinds of spectroscopy (NMR), spectrometry (mass-spec), and image segmentation (e.g. in astronomy), but these algorithms usually do not consider any hierarchical aspects. Hierarchical peak finding is analogous to hierarchical clustering, which is widely used in bioinformatics. However, our approach is closely related to the hierarchical analyses of energy landscapes and their barriers in studies of dynamics, metastability, and timescales [36-39]. The algorithm uses a subroutine for finding hierarchical probability peaks in one dimension, described in the next section.

1D peaks

This section briefly revisits the 1D peak finding method and the use of a nonstandard pedigree terminology [29]. Here is a generic formulation of the problem: Let $p(x)$ be some probabilities (possibly marginal) defined for $x = 1, \dots, N$. What are the peaks in $p(x)$? The computational task is divided into two steps. The first step is to construct a discrete tree of possible peaks, and the second step is to select peaks by searching the tree.

To simplify the presentation, we assume that $p(x_1) \neq p(x_2)$ if $x_1 \neq x_2$. Let Ψ be the set of x -values, where $p(x)$ has local minima and maxima. We associate a possible peak with each element $a \in \Psi$. If a is a local minimum, the peak is defined as illustrated in Figure 2. The *peak location* is the extent on the x -axis, $L(a) = [x_{\text{start}}(a), x_{\text{end}}(a)]$, defined as the largest interval including a in which $p(x) \geq p(a)$. The *peak width* is the size of $L(a)$, $p_w(a) = x_{\text{end}}(a) - x_{\text{start}}(a) + 1$. The *peak volume* is the probability summed over the location, $p_v(a) = \sum_{x \in L(a)} p(x)$. The peak's *bottom* (or *mode*) $\beta a = \arg \max_{x \in L(a)} p(x)$ is the x -value where p attains its maximum. (The term "bottom" originates from the corresponding energy landscape picture, but it is the position of the peak's top.) The *peak height* is $p_h(a) = p(\beta a)$. The peak's *depth* is $D(a) = \log_{10} \frac{p(\beta a)}{p(a)}$. We also associate a possible peak with each local maximum $a \in \Psi$, namely the spike itself: $L(a) = [a, a]$, $p_w(a) = 1$, $\beta a = a$, $p_v(a) = p_h(a) = p(a)$, and $D(a) = 0$.

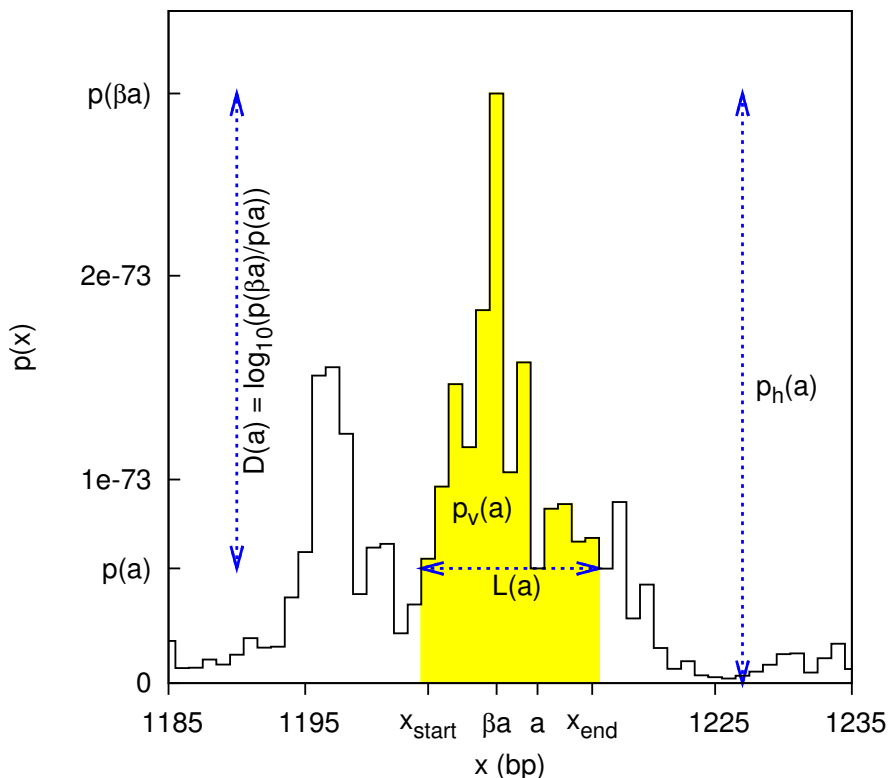


Figure 2
Example of a 1D peak. This peak in $p(x)$ has peak volume (yellow area) $p_v(a) = 1.5 \times 10^{-72}$, while the peak height is $p_h(a) = 2.9 \times 10^{-73}$, which is the maximum probability attained at $\beta a = 1209$. The peak location $L(a)$ is the extent from $x_{start} = 1204$ to $x_{end} = 1216$, which corresponds to the local minimum attained at $a = 1212$. The depth is $D(a) = 0.711$.

While peaks may be high, it is a more defining characteristic that they are wide. A peak is produced by the fluctuations in x , rather than disturbed by them. For each local maximum, there are many possible peaks. Therefore, a peak can not be identified with its bottom. Instead, we use the elements in Ψ as unique identifiers of peaks. The location of a peak is $L(a)$, not the bottom position βa , and the size of a peak is the peak volume, not the peak height. However, for the second type of peaks (the maxima), the peak location reduces to the bottom and the peak volume reduces to the peak height.

The set Ψ of possible peaks is hierarchically ordered. A binary tree is defined by the set inclusion order on the set of peak locations. For each pair $a, a' \in \Psi$, either $L(a) \subseteq L(a')$, or $L(a) \supseteq L(a')$, or they are disjoint. The branching corresponds to each local minimum a dividing the peak into two subpeaks, see Figure 2, just as a barrier or a watershed or a saddle point divides two valleys or lakes in a landscape [36,38,39]. The global minimum is the root node ρ of the tree. The local maxima are the leaf nodes of the tree. Each $a \in \Psi$ has at most three edges, one towards

the root and two away from the root. Each $a \neq \rho$ has an edge towards the root that connects to the successor σa . Each successor has an increased depth: $D(\sigma a) \geq D(a)$. And each local minimum a has two edges away from the root that connect to two ancestors. The highest peak of the two ancestors is the father πa and the other is the mother μa , i.e., they are distinguished by $p_h(\pi a) > p_h(\mu a)$. A left-right distinction between the two is not used. The notation $\sigma^n a$ means the successor taken $n \geq 0$ times, where $\sigma^0 a = a$. Each a has a set of successors $\Sigma(a)$ defined as the path from a to the root: $a, \sigma a, \sigma^2 a, \dots, \rho$. Each a also has a set of ancestors $\Delta(a)$ defined by $a' \in \Delta(a) \Leftrightarrow a \in \Sigma(a')$. The set $\Delta(a)$ is the subtree that has a as its root node. A bottom is typically shared by several peaks. For example, a peak has the same bottom as its father, $\beta a = \beta \pi a$, but not the same as its mother, $\beta a \neq \beta \mu a$. Each a has a paternal line $\Pi(a)$, defined as the set of all nodes that share a 's bottom. $\Pi(a)$ is also the path including a connected by fathers that ends at βa . The beginning of the path, called the full node ϕa , is either a mother or the root. The paternal lines establish a one-to-one correspondence between the set of maxima (i.e. bottoms) and the set of mothers including the root.

Having established a hierarchy Ψ of possible peaks, the second step is to select among them. The selection applies two independent criteria, each controlled by an input parameter: the *maximum depth* D_{max} and the *probability cutoff* p_c . The first criterion is that a is a *1D peak* according to the following definition.

Definition 1. Let D_{max} be the maximum depth of peaks. Then $a \in \Psi$ is a 1D peak if

$$(i) D(a) < D_{max}$$

$$(ii) D(\sigma a) \geq D_{max} \text{ or } a = \rho.$$

The second criterion is that $p_v(a) \geq p_c$. The first criterion is invoked by using the MAXDEEP subroutine [29], which returns the set P of all 1D peaks. The second criterion is subsequently invoked by calculating the peak volume of each $a \in P$ and comparing with the probability cutoff.

Bubbles and helical regions

The stitch profile algorithm is separate from the statistical mechanical DNA melting model. The only interface to the underlying model is by calling the following probability functions:

$$p_{right}(x) = P(\dots XX \underbrace{1 \ 0 \dots 0}_{unzipped} - 3'), \tag{1}$$

$$p_{left}(y) = P(5' - \underbrace{0 \dots 0}_{unzipped} \ 1 \ XX \dots), \tag{2}$$

$$p_{bubble}(x, y) = P(\dots XX \underbrace{1 \ 0 \dots 0}_{bubble} \ 1 \ XX \dots), \tag{3}$$

$$p_{helix}(x, y) = P(\dots XX \underbrace{0 \ 1 \dots 1}_{helix} \ 0 \ XX \dots), \tag{4}$$

$$p_{helix}(x, N) = P(\dots XX \underbrace{0 \ 1 \dots 1}_{zipped} - 3'), \tag{5}$$

$$p_{helix}(1, y) = P(5' - \underbrace{1 \dots 1}_{zipped} \ 0 \ XX \dots). \tag{6}$$

In these equations, 1 is a bound basepair (helix), 0 is a melted basepair (coil), X is either 0 or 1, and the sequence positions x and/or y are indicated.

In addition to these, the stitch profile algorithm calls methods for adding these probabilities (peak volumes) and for computing upper bounds on such probability sums. This means that it is easy to change or replace the

underlying model. In this article, the Poland-Scheraga model with Fixman-Freire loop entropies is used [30], but in principle, other DNA melting models could be used, or even models that include secondary structure [40].

This article discusses how to efficiently compute bubble stitches and helix stitches only. The 5' and 3' tail stitches are efficiently computed as in Algorithm 1 [29]. Each bubble stitch corresponds to a peak in the bubble probability function in Eq. (3). And each helix stitch corresponds to a peak in the helix probability function in Eq. (4). These two probability functions and their peaks are two dimensional, so the 1D peak finding method does not directly apply. However, the 1D peak analysis can be performed for each of the other four probability functions [Eqs. (1), (2), (5), and (6)]. Using Eq. (1), a binary tree Ψ_x and a set of 1D peaks P_x is computed, and using Eq. (2), a binary tree Ψ_y and a set of 1D peaks P_y is computed. The probability cutoff is not invoked here. These two tree structures with their 1D peaks are then further processed, as described in the following two sections, to obtain the bubble stitches. Likewise, using Eq. (5), a binary tree Ψ_x and a set of 1D peaks P_x is computed, and using Eq. (6), a binary tree Ψ_y and a set of 1D peaks P_y is computed. These are used similarly to obtain the helix stitches. This division of labor also indicates an obvious parallelization of the algorithm using two or four processors. Parallelism was not implemented in this study, however.

2D peaks

The goal of this section is to define 2D peaks and to prove the key result that some 2D peaks are simply the Cartesian product of two 1D peaks. But not all 2D peaks have this property, making it a nontrivial result. This is expressed in Theorem 2.

Theorem 2 also indicates a convenient way of computing all 2D peaks, on which Algorithm 2 is directly based. Theorem 2 shows that Algorithm 2's computation of stitch profiles is exact, that is, complying strictly with the mathematical definition of 2D peaks. The proof is therefore important for the validation of Algorithm 2. While Theorem 2 is the primary goal, we also prove Theorem 1 which similarly provides validation of Algorithm 1. But more importantly, a comparison of the two theorems gives more insight in both algorithms.

A *frame* is a pair $(a, b) \in \Psi_x \times \Psi_y$. A frame also refers to the corresponding box $L(a) \times L(b)$ in the xy -plane. A frame (a, b) is *contained inside* another frame (a', b') , if $L(a) \times L(b) \subset L(a') \times L(b')$, that is, if $a' \in \Sigma(a)$ and $b' \in \Sigma(b)$. The *root frame* is (ρ_x, ρ_y) . A frame (a, b) is *nonroot* if $(a, b) \neq (\rho_x, \rho_y)$. A frame (a, b) is a *bottom frame* if $(a, b) = (\beta a, \beta b)$ and it is *nonbottom* if $(a, b) \neq (\beta a, \beta b)$. The *depth* of a frame (a, b) is

$D(a, b) = \max\{D(a), D(b)\}$. From this definition, we immediately get

$$D(a, b) < D_{\max} \Leftrightarrow D(a) < D_{\max} \text{ and } D(b) < D_{\max}. \quad (7)$$

To simplify the presentation, we assume that for all frames: $D(a) \neq D(b)$.

Definition 2. The successor of a nonroot frame (a, b) is

$$\sigma(a, b) = \begin{cases} (\sigma a, b) & \text{if } D(\sigma b) > D(\sigma a) \text{ or } b = \rho_y \\ (a, \sigma b) & \text{if } D(\sigma a) > D(\sigma b) \text{ or } a = \rho_x \end{cases} \quad (8)$$

A successor of the root frame does not exist.

Having defined the depth and the successor, what is the depth of a successor?

Proposition 1. For every nonroot (a, b) , $D(\sigma(a, b)) \geq D(a, b)$.

Proof. For $\sigma(a, b) = (\sigma a, b)$, $\max\{D(\sigma a), D(b)\} \geq \max\{D(a), D(b)\}$ because $D(\sigma a) \geq D(a)$. Likewise for $\sigma(a, b) = (a, \sigma b)$. \square

Definition 3. A frame (a, b) is σ -above if

$$(i) \ D(\sigma a) > D(b) \text{ or } a = \rho_x,$$

$$(ii) \ D(\sigma b) > D(a) \text{ or } b = \rho_y.$$

The term " σ -above" is a mnemonic for the two inequalities in the definition. The set of all frames that are σ -above is called *the frame tree*. While Prop. 1 only sets a lower bound on the depth of a successor, we can write the actual value for σ -above frames:

Proposition 2. If (a, b) is nonroot and σ -above, then

$$D(\sigma(a, b)) = \begin{cases} D(\sigma a) & \text{if } \sigma(a, b) = (\sigma a, b) \\ D(\sigma b) & \text{if } \sigma(a, b) = (a, \sigma b). \end{cases} \quad (9)$$

Furthermore, $D(\sigma(a, b)) = \min\{D(\sigma a), D(\sigma b)\}$ if both $a \neq \rho_x$ and $b \neq \rho_y$.

Proof. If $\sigma(a, b) = (\sigma a, b)$, then $a \neq \rho_x$ and $\max\{D(\sigma a), D(b)\} = D(\sigma a)$ by Def. 3. If, furthermore, $b \neq \rho_y$, then $D(\sigma(a, b)) = D(\sigma a) < D(\sigma b)$ by Def. 2. Likewise if $\sigma(a, b) = (a, \sigma b)$. \square

By repeatedly taking the successor, we eventually end up at the root frame in, say, R steps. $\Sigma(a, b)$ is the sequence of successors of (a, b) , i.e., the sequence $\{\sigma^n(a, b)\}_0^R$ that begins at (a, b) and ends at the root frame. Alternatively,

$\Sigma(a, b)$ is defined as the *set of successors*, i.e., the set of such sequence elements. What if we want to exclude (a, b) from $\Sigma(a, b)$? That can be written as $\Sigma(\sigma(a, b))$.

If (a, b) is not σ -above, then its sequence of successors takes the shortest path to a σ -above frame, or put another way:

Proposition 3. If $a' \in \Sigma(a)$, $b' \in \Sigma(b)$ and (a', b') is σ -above, then $(a', b') \in \Sigma(a, b)$.

Proof. All elements in both $\Sigma(a)$ and $\Sigma(b)$ are visited by the sequence $\Sigma(a, b)$ on its climb to the root frame. Assume $(a', b') \notin \Sigma(a, b)$. Then either a' is passed before b' is reached, or viceversa, and we can assume that a' comes first. In other words, $a' \neq \rho_x$ and there is a $b'' \neq b'$ such that $b' \in \Sigma(b'')$ and $\sigma(a', b'') = (\sigma a', b'')$. Then $D(b') \geq D(\sigma b'')$. By Def. 2, we see that $D(\sigma b'') > D(\sigma a')$. (a', b') is σ -above, so by Def. 3, we see that $D(\sigma a') > D(b')$. We arrive at the contradiction $D(b') > D(b')$. \square

Each frame is the successor of at most four frames. If $(a, b) = \sigma(a', b')$ then (a', b') is either $(\pi a, b)$, $(a, \pi b)$, $(\mu a, b)$, or $(a, \mu b)$. Two of these are defined as *ancestors*:

Definition 4. The father of a nonbottom frame (a, b) is

$$\pi(a, b) = \begin{cases} (\pi a, b) & \text{if } D(a) > D(b) \\ (a, \pi b) & \text{if } D(a) < D(b) \end{cases} \quad (10)$$

The mother of a nonbottom frame (a, b) is

$$\mu(a, b) = \begin{cases} (\mu a, b) & \text{if } D(a) > D(b) \\ (a, \mu b) & \text{if } D(a) < D(b). \end{cases} \quad (11)$$

Fathers and mothers of bottom frames do not exist.

Each father or mother can have its own father and mother, and so on. The *set of ancestors* $\Delta(a, b)$ is the binary subtree defined recursively by: (1) $(a, b) \in \Delta(a, b)$. (2) If nonbottom $(a', b') \in \Delta(a, b)$ then $\pi(a', b') \in \Delta(a, b)$ and $\mu(a', b') \in \Delta(a, b)$.

The next proposition shows that being σ -above is propagated by σ , π , and μ :

Proposition 4. Let (a, b) be σ -above.

(i) If $(a', b') \in \Sigma(a, b)$ then (a', b') is σ -above.

(ii) If $(a', b') \in \Delta(a, b)$ then (a', b') is σ -above.

Proof. (i): First, we show that $\sigma(a, b)$ is σ -above: If $\sigma(a, b) = (\sigma a, b)$, then Def. 2 implies the second condition: $D(\sigma b)$

$> D(\sigma a)$ or $b = \rho_y$. And (a, b) is σ -above which by Def. 3 implies the first condition: $D(\sigma^2 a) > D(\sigma a) > D(b)$ or $\sigma a = \rho_x$. Similarly, $\sigma(a, b) = (a, \sigma b)$ is shown to be σ -above. The proof is completed by induction.

(ii): First, we show that $\pi(a, b)$ is σ -above: If $\pi(a, b) = (\pi a, b)$, then Eq. (10) implies the first condition: $D(\sigma \pi a) = D(a) > D(b)$ or $\pi a = \rho_x$. And (a, b) is σ -above which by Def. 3 implies the second condition: $D(\sigma b) > D(a) > D(\pi a)$ or $b = \rho_y$. Similarly, $\pi(a, b) = (a, \pi b)$ and $\mu(a, b)$ are shown to be σ -above. The proof is completed by induction. \square

Successors are the inverse of fathers and/or mothers for σ -above frames only:

Proposition 5. *If (a, b) is nonbottom and nonroot, the following statements are equivalent:*

(i) (a, b) is σ -above

(ii) $\sigma\pi(a, b) = (a, b)$

(iii) $\sigma\mu(a, b) = (a, b)$

(iv) $\pi\sigma(a, b) = (a, b)$ or $\mu\sigma(a, b) = (a, b)$

Proof. (i) \Leftrightarrow (ii): If $\pi(a, b) = (\pi a, b)$, then Eq. (10) implies the first condition that (a, b) is σ -above: $D(\sigma a) > D(a) > D(b)$ or $a = \rho_x$. Then (a, b) is σ -above $\stackrel{\text{Def.3}}{\Leftrightarrow} D(\sigma b) > D(a) = D(\sigma \pi a)$ or $b = \rho_y \stackrel{\text{Def.2}}{=} \Leftrightarrow \sigma(\pi a, b) = (\sigma \pi a, b) \Leftrightarrow \sigma\pi(a, b) = (a, b)$. If $\pi(a, b) = (a, \pi b)$, the equivalence is shown similarly.

(i) \Leftrightarrow (iii): Replace π by μ in the above.

(i) \Leftrightarrow (iv): If $\sigma(a, b) = (\sigma a, b)$, then Def. 2 implies the second condition that (a, b) is σ -above: $D(\sigma b) > D(\sigma a) > D(a)$ or $b = \rho_y$. Then (a, b) is σ -above $\stackrel{\text{Def.3}}{\Leftrightarrow} D(\sigma a) > D(b) \stackrel{\text{Def.4}}{\Leftrightarrow} \pi(\sigma a, b) = (\pi \sigma a, b)$ or $\mu(\sigma a, b) = (\mu \sigma a, b) \Leftrightarrow \pi(\sigma a, b) = (\pi \sigma a, b)$ or $\mu(\sigma a, b) = (\mu \sigma a, b) \Leftrightarrow \pi\sigma(a, b) = (a, b)$ or $\mu\sigma(a, b) = (a, b)$. If $\sigma(a, b) = (a, \sigma b)$, the equivalence is shown similarly. \square

Accordingly, there is an "inverse" relationship between the sets of successors and ancestors:

Proposition 6. *(a', b') is σ -above and $(a, b) \in \Sigma(a', b')$ iff (a, b) is σ -above and $(a', b') \in \Delta(a, b)$.*

Proof. $(a, b) \in \Sigma(a', b')$ implies a path of successors from (a', b') to (a, b) . Prop. 4 shows that all elements in the path are σ -above. Prop. 5(iv) applied to each step in the path gives an opposite path of ancestors.

Conversely, $(a', b') \in \Sigma(a, b)$ implies a path of ancestors from (a, b) to (a', b') . Prop. 4 shows that all elements in the path are σ -above. Prop. 5(ii) and (iii) applied to each step in the path gives an opposite path of successors. \square

It follows from Prop. 6 that the frame tree is equal to the binary tree $\Delta(\rho_x, \rho_y)$, because $(\rho_x, \rho_y) \in \Sigma(a', b')$ for any (a', b') . It has the same pedigree properties as Ψ , such as paternal lines and $\beta\pi(a, b) = \beta(a, b)$. So far, we have covered ground that was already implicit in [29], but augmented here with proofs. The next concept is new, however, namely the Cartesian products of 1D peaks.

Definition 5. *(a, b) is a grid frame if a and b are 1D peaks.*

The set of all grid frames is $G = P_x \times P_y$. As Figure 3 shows, G has a grid-like ordering in the xy -plane. All 1D peaks $a \in P_x$ have disjoint peak locations $L(a) = [x_{\text{start}}(a), x_{\text{end}}(a)]$. They can be indexed by $i = 1, 2, 3, \dots$ according to their ordering from 5' to 3' on the sequence, such that $x_{\text{end}}(a_i) < x_{\text{start}}(a_{i+1})$. Likewise, the 1D peaks $b \in P_y$ can be indexed by j . Then the grid frames form a matrix G with elements $[G]_{ij} = (a_i, b_j)$. We use the symbol G for both the set and the matrix.

Proposition 7. *Every grid frame (a, b) is σ -above.*

Proof. If $a \neq \rho_x$, then $D(\sigma a) \geq D_{\text{max}}$ because a is a 1D peak and $D_{\text{max}} > D(b)$ because b is a 1D peak (see Def. 1), thus showing Def. 3(i). Similarly, we show Def. 3(ii). \square

The following two lemmas show that grid frames inherit some properties from 1D peaks.

Lemma 1. *(a, b) is a grid frame iff*

(i) (a, b) is σ -above,

(ii) $D(a, b) < D_{\text{max}}$

(iii) $D(\sigma(a, b)) \geq D_{\text{max}}$ or (a, b) is the root frame.

Proof. If (a, b) is a grid frame, then it is σ -above by Prop. 7 and Eq. (7) implies $D(a, b) < D_{\text{max}}$. For nonroot (a, b) , $D(\sigma(a, b))$ equals either $D(\sigma a)$ or $D(\sigma b)$ (Prop. 2), which is $\geq D_{\text{max}}$ because a and b are 1D peaks.

Conversely, Eq. (7) implies $D(a) < D_{\text{max}}$. For $a = \rho_x$, a is then a 1D peak. For $a \neq \rho_x$, Prop. 2 gives $D(\sigma a) \geq D(\sigma a,$

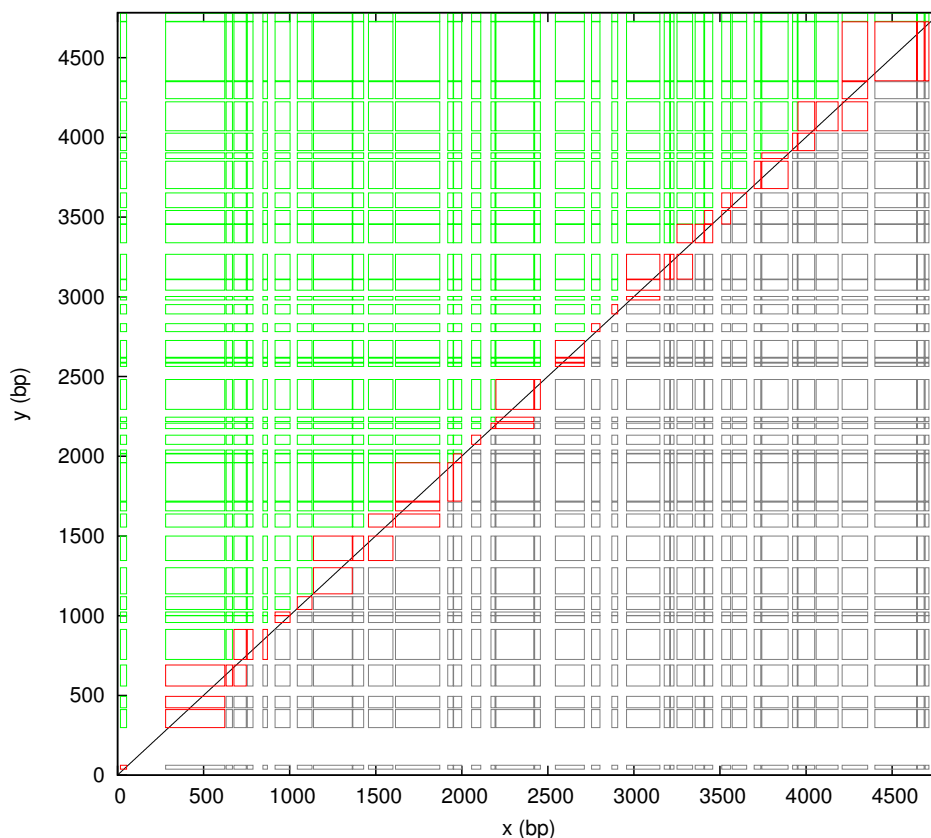


Figure 3
The set $G = P_x \times P_y$ of all grid frames plotted in the xy -plane. The grid frames are colored to distinguish those that are above the diagonal (green), crossing the diagonal (red), and below the diagonal (grey), thus illustrating the subsets G_a , G_c and G_b , respectively. Frames with side lengths below 20 bp are not shown to unclutter the figure.

$b)) \geq D_{\max}$, so a is a 1D peak. Similarly, b is shown to be a 1D peak. \square

Lemma 2. Let D_{\max} be the maximum depth of peaks.

(i) For each a with $D(a) < D_{\max}$ there is exactly one 1D peak $a' \in \Sigma(a)$.

(ii) For each (a, b) with $D(a, b) < D_{\max}$ there is exactly one grid frame $(a', b') \in \Sigma(a, b)$.

Proof. (i): The depth increases monotonically in the sequence $\Sigma(a)$ of successors ($\forall n : D(\sigma^n a) \leq D(\sigma^{n+1} a)$). For $D(\rho_x) \geq D_{\max}$ there is therefore a unique element $a' \neq \rho_x$ with $D(a') < D_{\max}$ and $D(\sigma a') \geq D_{\max}$. For $D(\rho_x) < D_{\max}$, $a' = \rho_x$ is a 1D peak and no other element in $\Sigma(a)$ can fulfill Def. 1(ii).

(ii): Eq. (7) gives $D(a) < D_{\max}$ and $D(b) < D_{\max}$. By applying (i) to a and b , we obtain a unique grid frame (a', b') where

$a' \in \Sigma(a)$ and $b' \in \Sigma(b)$. (a', b') is σ -above by Prop. 7, so $(a', b') \in \Sigma(a, b)$ by Prop. 3. \square

How do we define 2D peaks? A straightforward way would be to generalize 1D peaks by simply rewriting Def. 1 in the frame tree context. The result would be the grid frames, as we see by Lemma 1. However, there is more to the picture than the frame tree, due to a further constraint to be discussed next, which requires a more elaborate definition of 2D peaks.

In genomic annotations, a region is specified by coordinates x, y , where by convention $x < y$, i.e., x is the 5' end and y is the 3' end. We adopt the same constraint for our notation (x, y) of the instantaneous location of a bubble or helix. In the xy -plane, helices are only defined for (x, y) above the diagonal line $y = x$. Bubbles have at least one melted basepair in between x and y , so they are only defined for (x, y) above the diagonal line $y = x + 1$. Accordingly, we require that frames are above the diagonal line, as defined in the following.

Definition 6. A frame (a, b) is above the diagonal if

$$x_{end}(a) + 1 < y_{start}(b) \text{ for bubbles,} \tag{12a}$$

$$x_{end}(a) < y_{start}(b) \text{ for helices.} \tag{12b}$$

A frame (a, b) is below the diagonal if

$$x_{start}(a) + 1 \geq y_{end}(b) \text{ for bubbles,} \tag{13a}$$

$$x_{start}(a) \geq y_{end}(b) \text{ for helices.} \tag{13b}$$

A frame (a, b) is crossing the diagonal if it is neither above the diagonal nor below the diagonal.

Note: A frame that is crossing the diagonal contains at least one point (x, γ) above the diagonal line, while a frame that is below the diagonal contains no points above the diagonal line, but its upper left corner may be on the

diagonal line. Figure 3 illustrates frames that are above, crossing and below the diagonal.

The requirement that a frame is above the diagonal puts a constraint on its size. This is embodied in the next concept.

Definition 7. The root frame is a fractal frame if it is above the diagonal. A nonroot frame (a, b) is a fractal frame if

- (i) (a, b) is above the diagonal,
- (ii) $\sigma(a, b)$ is crossing the diagonal,
- (iii) (a, b) is σ -above.

The set of all fractal frames is denoted F . As Figure 4 shows, fractal frames tend to be smaller the closer they are to the diagonal, thus resembling a fractal. For a typical fractal frame, the fluctuations in x and γ are comparable in size to the length $\gamma - x$ of the bubble or helix itself. Indeed,

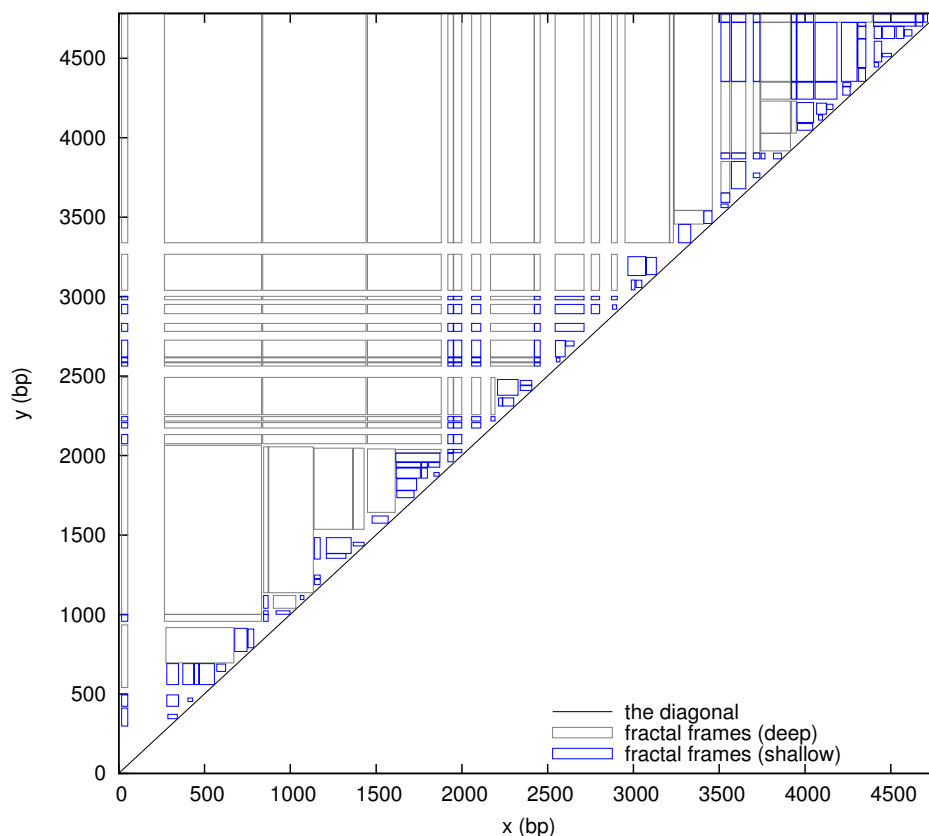


Figure 4
The set F of all fractal frames plotted in the xy -plane. The fractal frames $(a, b) \in F$ are colored to distinguish those with depths $D(a, b) \geq D_{max}$ (grey) and $D(a, b) < D_{max}$ (blue), thus illustrating the subsets F_d and F_s , respectively. Frames with side lengths below 20 bp are not shown to unclutter the figure.

the two peak locations $L(a)$ and $L(b)$ are as wide as possible, while not overlapping each other (because the successor is crossing the diagonal). In contrast, the fluctuations for grid frames are relatively small on average and independent of the bubble or helix length.

Lemma 3. *For each σ -above and above the diagonal (a, b) , there is exactly one fractal frame $(a', b') \in \Sigma(a, b)$.*

Proof. Let $(a', b') = \sigma^n(a, b)$, where n is the largest number for which $\sigma^n(a, b)$ is above the diagonal. (a', b') is σ -above by Prop. 4. For all $m > n$, frames $\sigma^m(a, b)$ (if they exist) are not above the diagonal, nor below the diagonal because they contain (a, b) , hence they are crossing the diagonal. Therefore (a', b') is a fractal frame. For all $m < n$, frames $\sigma^m(a, b)$ (if they exist) are above the diagonal, because they are contained in (a', b') . Therefore (a', b') is the only fractal frame in $\Sigma(a, b)$. \square

Lemma 3 is similar to Lemma 2. By Prop. 6, we can express both lemmas in terms of ancestors Δ instead of successors Σ . The lemmas then say that certain kinds of frames are organized as forests. A *forest* is a set of disjoint trees. The sets F and G generate two forests: $\cup_{(a,b) \in G} \Delta(a, b)$ consists of the subtrees having grid frames as root nodes. $\cup_{(a,b) \in F} \Delta(a, b)$ consists of the subtrees having fractal frames as root nodes. By these forests, we generate from G the set of all σ -above frames with $D(a, b) < D_{\max}$ and we generate from F the set of all σ -above frames above the diagonal.

All the necessary concepts are now in place for the definition of 2D peaks. We will not repeat the "derivation" of 2D peaks given in [29], but just recall that 2D peaks are defined with a purpose: They must capture the extent of the actual peaks in the probability functions $p_{\text{bubble}}(x, y)$ and $p_{\text{helix}}(x, y)$. And they must have an interpretation in terms of fluctuations on a given timescale. The following definition is equivalent to the formulation in [29].

Definition 8. *Let D_{\max} be the maximum depth of peaks. A frame (a, b) is a 2D peak if*

- (i) (a, b) is above the diagonal,
- (ii) (a, b) is σ -above,
- (iii) $D(a, b) < D_{\max}$
- (iv) $D(\sigma(a, b)) \geq D_{\max}$ or (a, b) is a fractal frame.

Note: the *or* in the definition is not an *exclusive or*. A 2D peak (a, b) can both be a fractal frame and have $D(\sigma(a, b)) \geq D_{\max}$. The set of all 2D peaks is denoted P and is illustrated in Figure 5.

Comparing Def. 8 and Lemma 1, we see that the difference between 2D peaks and grid frames is due to the diagonal constraint: First, the requirement that 2D peaks are above the diagonal, and second, the possible exemption from the second inequality, which for grid frames is being the root frame, while for 2D peaks it is being a fractal frame. Unlike grid frames, 2D peaks can capture events close to the diagonal by adapting their size.

Computing the 2D peaks is at the core of the stitch profile methodology. The following two theorems provide characterizations of 2D peaks that may be translated into computer programs.

Theorem 1. *We divide 2D peaks into two types, being fractal frames or not, that can be distinctly characterized as follows.*

- (i) (a, b) is a 2D peak and a fractal frame iff (a, b) is a fractal frame and $D(a, b) < D_{\max}$.
- (ii) (a, b) is a 2D peak and not a fractal frame iff (a, b) is a grid frame and there is a fractal frame (a', b') with $D(a', b') \geq D_{\max}$ such that $(a', b') \in \Sigma(a, b)$.

Proof. (i): Immediate by Defs. 7 and 8.

(ii): If a 2D peak (a, b) is not a fractal frame, then $D(\sigma(a, b)) \geq D_{\max}$ by Def. 8, so (a, b) is a grid frame by Lemma 1. Applying Lemma 3, there is a fractal frame $(a', b') \in \Sigma(a, b)$. $(a, b) \neq (a', b')$ because one is a fractal frame, the other is not, so $(a', b') \in \Sigma(\sigma(a, b))$, which by Prop. 1 implies $D(a', b') \geq D_{\max}$.

Conversely, (a, b) is above the diagonal because it is contained in a fractal frame. $(a, b) \neq (a', b')$ because $D(a, b) < D_{\max}$ and $D(a', b') \geq D_{\max}$ implying that (a, b) is not a fractal frame (uniqueness by Lemma 3) and not the root frame. The other requirements for a 2D peak are established by Lemma 1. \square

Theorem 1 characterizes all 2D peaks by their relationship to fractal frames. This is applied in Algorithm 1, that derives all 2D peaks from fractal frames. However, the next theorem shows that some 2D peaks can be characterized without referring to fractal frames.

Theorem 2. *A nonroot 2D peak has a successor, the depth of which is either greater or less than D_{\max} . We thus divide 2D peaks into two types, that can be distinctly characterized as follows. Let (a, b) be nonroot. Then*

- (i) (a, b) is a 2D peak and $D(\sigma(a, b)) \geq D_{\max}$ iff (a, b) is a grid frame that is above the diagonal.

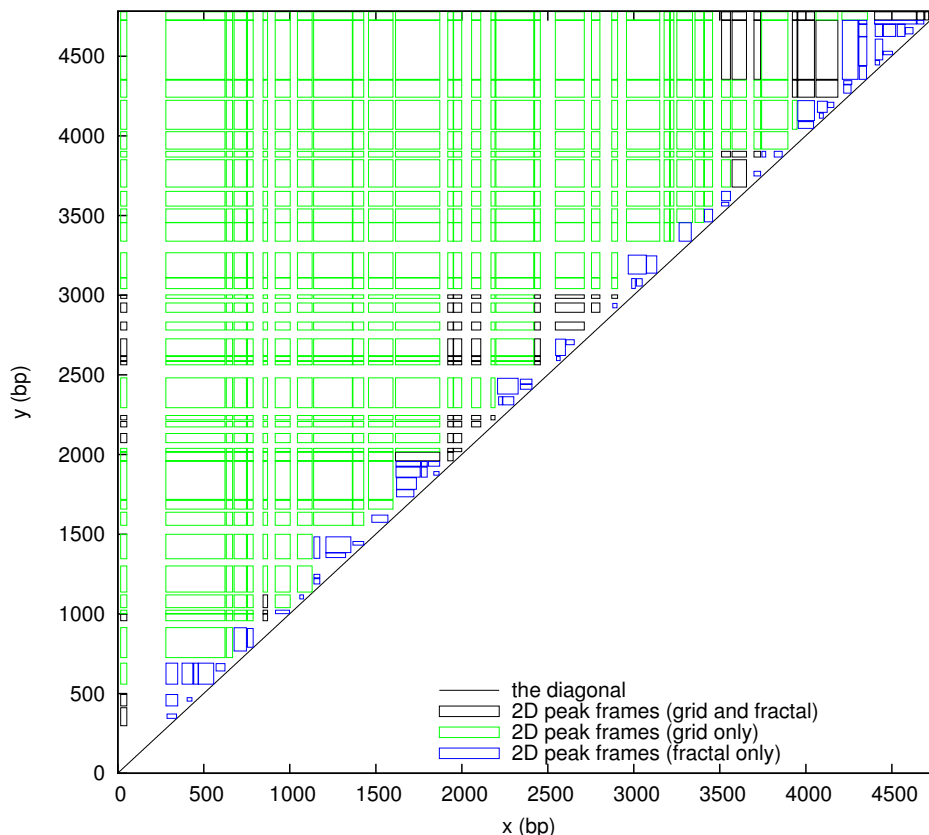


Figure 5
The set P of all 2D peaks plotted in the xy -plane. The 2D peak frames are colored to distinguish those that are fractal frames (blue), fractal frames and grid frames (black), or grid frames (green), thus illustrating the subsets P_F , P_{FG} and P_G , respectively. Frames with side lengths below 20 bp are not shown to unclutter the figure.

(ii) (a, b) is a 2D peak and $D(\sigma(a, b)) < D_{max}$ iff (a, b) is a fractal frame and there is a grid frame (a', b') that is crossing the diagonal, such that $(a', b') \in \Sigma(a, b)$.

Proof. (i): Immediate by Def. 8 and Lemma 1.

(ii): If a 2D peak (a, b) has $D(\sigma(a, b)) < D_{max}$ then (a, b) is a fractal frame by Def. 8. Applying Lemma 2 to $\sigma(a, b)$, there is a grid frame $(a', b') \in \Sigma(\sigma(a, b)) \subset \Sigma(a, b)$. Frame (a', b') is crossing the diagonal because it contains $\sigma(a, b)$, which is crossing the diagonal because (a, b) is a fractal frame.

Conversely, $(a, b) \neq (a', b')$ because (a, b) is above the diagonal (a fractal frame) and (a', b') is crossing the diagonal, and hence $(a', b') \in \Sigma(\sigma(a, b))$. Since (a', b') is a grid frame, Lemma 1 gives $D(a', b') < D_{max}$, which by Prop. 1 implies $D(a, b) \leq D(\sigma(a, b)) < D_{max}$ and we conclude that (a, b) is a 2D peak. \square

Note: Theorem 2 does not consider the root frame. However, if the root frame is a 2D peak, then it is of the first type: a grid frame that is above the diagonal.

It follows from Theorems 1 and 2 that a 2D peak is either a grid frame, a fractal frame, or both. The set of 2D peaks P can therefore be divided into three disjoint sets defined as follows. P_F are the 2D peaks that are fractal frames only, not grid frames. P_{FG} are the 2D peaks that are both fractal frames and grid frames. P_G are the 2D peaks that are grid frames only, not fractal frames. Let G_a , G_b and G_c be the sets of grid frames that are above, below and crossing the diagonal, respectively. Let F_d and F_s be the sets of fractal frames that are deep ($D(a, b) \geq D_{max}$) and shallow ($D(a, b) < D_{max}$), respectively. In Figs. 3, 4, 5, all these subsets are illustrated with different colors. The following corollary summarizes the relationships between grid frames, fractal frames and 2D peaks:

Corollary 1. *The set of 2D peaks is $P = F_s \cup G_a$. The intersection between the grid and the fractal is $P_{FG} = F_s \cap G_a = F \cap G$.*

Furthermore, the 2D peaks can be obtained by the following two expressions, in which all set unions are between disjoint sets:

$$P = F_s \bigcup_{(a',b') \in F_d} G \cap \Delta(a',b'), \quad (14a)$$

$$= G_a \bigcup_{(a',b') \in G_c} F \cap \Delta(a',b'). \quad (14b)$$

Proof. $P = P_F \cup P_{FG} \cup P_G$. Theorem 1 states that $P_F \cup P_{FG} = F_s$ and that

$$P_G = \bigcup_{(a',b') \in F_d} G \cap \Delta(a',b').$$

Here, $\Delta(a', b')$ is brought into play by Prop. 6. Theorem 2 states that $P_{FG} \cup P_G = G_a$ (the root frame would go here) and that

$$P_F = \bigcup_{(a',b') \in G_c} F \cap \Delta(a',b'). \square$$

Eqs. (14a) and (14b) outline how the set of 2D peaks is built up computationally by Algorithm 1 and 2, respectively. Writing the expressions side by side shows the parallels: Algorithm 1 takes some fractal frames and then it adds some grid frames that are contained inside fractal frames. Algorithm 2 takes some grid frames and then it adds some fractal frames that are contained inside grid frames. In both cases, the additional part is the more complicated part, as it requires searching some forests. The two Algorithms are algorithmically equivalent in terms of output, but the transformation in Eq. (14) from F -based to G -based facilitates a reduction in execution time, as described in the next section.

The fast and exact algorithm

Algorithm 2 owes its speed to two important ingredients: One is the grid frame matrix G associated to the parameter D_{\max} . The other is an upper bound associated to the parameter p_c .

To compute all bubble stitches of the stitch profile, the algorithm must find those 2D peaks (a, b) in the bubble context that have a peak volume

$$p_v(a, b) = \sum_{x \in L(a)} \sum_{\gamma \in L(b)} p_{\text{bubble}}(x, \gamma) \quad (15)$$

that is greater or equal to the probability cutoff p_c . According to Eq. (14b), one can write an algorithm for obtaining all 2D peaks using two nested loops that goes through all matrix elements (a_i, b_j) of the grid frame matrix G : If $(a_i,$

$b_j)$ is above the diagonal, it is a 2D peak. If (a_i, b_j) is crossing the diagonal, a subroutine computes the set $F \cap \Delta(a_i, b_j)$. If (a_i, b_j) is below the diagonal, it is skipped. By piping the resulting frames through a probability cutoff filter, we obtain the bubble stitches.

The matrix G is not stored in memory, only the two arrays P_x and P_y that provide each a_i and b_j . Matrix elements (a_i, b_j) being above, crossing or below the diagonal refers to the diagonal line in the xy -plane, never the diagonal of the matrix. For each row and column of the matrix there may be zero, one, or more matrix elements that are crossing the diagonal, as can be seen in Figure 3.

More specifically, let G be of order $m \times n$ and let the outer loop be over $j = n$ to 1 and the inner loop over $i = m$ to 1. The iteration thus begins at the upper right corner of Figure 3 and steps along the y -axis in the outer loop and the x -axis in the inner loop. However, we do not have to start at $i = m$ for each j . If (a_i, b_j) is below the diagonal, then (a_i, b_k) is below the diagonal for all $k < j$. Therefore, we can jump directly to the i that corresponds to the first grid frame that was not below the diagonal at the previous j . In this way, most of the grid frames that are below the diagonal are ignored by the algorithm. While this is a trivial programming trick, we shall now see a less trivial trick, that ignores most of the grid frames that are above the diagonal.

Recall [30] that the bubble probability is

$$p_{\text{bubble}}(a, b) = \frac{Z_{X10}(x)\Omega(y-x)Z_{01X}(y)}{Z}. \quad (16)$$

The loop entropy factor $\Omega(y - x)$ is a monotonically decreasing function. Its largest value in a frame (a, b) is therefore in the lower right corner, i.e. $\Omega_{\max} = \Omega(y_{\text{start}}(b) - x_{\text{end}}(a))$. Then

$$p_v(a, b) \leq \frac{\Omega_{\max}}{Z} \left(\sum_{x \in L(a)} Z_{X10}(x) \right) \left(\sum_{\gamma \in L(b)} Z_{01X}(\gamma) \right)$$

and the bubble peak volume has an upper bound that factorizes. Using the 1D peak volumes

$$p_v(a) = \sum_{x \in L(a)} Z_{X10}(x) / Z, \quad (17a)$$

$$p_v(b) = \sum_{\gamma \in L(b)} Z_{01X}(\gamma) / Z, \quad (17b)$$

we can write the upper bound as

$$\tilde{p}_v(a, b) = \Omega(\gamma_{\text{start}}(b) - x_{\text{end}}(a))Zp_v(a)p_v(b). \quad (18)$$

If a grid frame (a_i, b_j) has an upper bound below the cutoff, $\tilde{p}_v(a_i, b_j) < p_c$, then also $\tilde{p}_v(a_k, b_j) < p_c$ for all $k < i$ for which $p_v(a_k) \leq p_v(a_i)$, because the loop entropy factor is decreasing. In that case, their peak volumes are also below the cutoff, of course, and the algorithm can reject all these frames. We implement this observation by calculating in advance the *next bigger goat* $\text{nbG}(i)$ defined by

$$(i) \ p_v(a_k) \leq p_v(a_i) \text{ for } \text{nbG}(i) < k < i$$

$$(ii) \ p_v(a_{\text{nbG}(i)}) > p_v(a_i)$$

The $\text{nbG}(i)$ is calculated as follows: A loop over $i = 1$ to m compares each $p_v(a_i)$ successively to $p_v(a_{i-1})$, $p_v(a_{\text{nbG}(i-1)})$, $p_v(a_{\text{nbG}(\text{nbG}(i-1))})$, ... until a bigger one is found or the list ends.

For grid frames (a_i, b_j) that are above the diagonal, the algorithm first checks if $\tilde{p}_v(a_i, b_j) < p_c$, in which case it jumps directly to $(a_{\text{nbG}(i)}, b_j)$. The $\text{nbG}(i)$ may be undefined, if there are no bigger $p_v(a_k)$, in which case the inner loop is done and the outer loop proceeds to the next j . On the other hand, if $\tilde{p}_v(a_i, b_j) \geq p_c$, then the peak volume has to be calculated and checked. Although grid frames may be skipped without having calculated neither their peak volumes nor their upper bounds, the criterion for rejection is exact. There are no false negatives (or positives).

For each grid frame (a_i, b_j) that is crossing the diagonal, the algorithm calculates a set of 2D peaks, $F \cap \Delta(a_i, b_j)$, and checks the peak volume of each. This set consists of all fractal frames that are contained inside (a_i, b_j) . A mental picture is that (a_i, b_j) must be broken into fractal frames (fractured) to avoid crossing the diagonal. The algorithm searches the subtree $\Delta(a_i, b_j)$ top-down (breadth-first) with a recursive subroutine. A given input frame (a, b) is split into its father frame $\pi(a, b)$ and mother frame $\mu(a, b)$. Each in turn is then checked as follows: If it is crossing the diagonal, it is further split by giving it recursively as input to the subroutine. If instead it is above the diagonal, it is a fractal frame. With (a_i, b_j) as input, the subroutine finds $F \cap \Delta(a_i, b_j)$. (If instead the input is the root frame (ρ_x, ρ_y) , the subroutine will find all fractal frames F . This was applied in Algorithm 1.)

Figure 6 shows the resulting search process, by plotting only frames that are processed by the algorithm, while the ignored grid frames are blank. Comparing with Figure 3, we see that the blank areas correspond to the great bulk of

grid frames both above and below the diagonal, leaving just an irregular band of frames along the diagonal to be searched. This is a nice geometric illustration of the reduction from $O(N^2)$ to $O(N \log N)$ in execution time. Figure 6 also shows that some bubble stitches are fractal frames contained inside grid frames that are crossing the diagonal.

The peak volumes $p_v(a, b)$ of some frames must be calculated. Algorithm 2 spends a considerable fraction of its time on doing these summations. The summation over a bubble frame can be done faster if the frame is big enough, by exploiting the Fixman-Freire approximation à la Yeramian [32,41]. This does not improve the time complexity, but significantly reduces the total execution time by some factor.

To compute all helix stitches of the stitch profile, the algorithm follows exactly the same procedure as described above, but in the helix context. Eq. (14b) and the analysis in the previous section applies equally well to the bubble and the helix contexts. The various quantities are, of course, replaced by their helix counterparts. For example, the appropriate diagonal line is applied (Def 6). The main difference is the upper bound on helix peak volume. Since x and y decouples in the helix probability [29],

$$p_{\text{helix}}(x, y) = \frac{p_{\text{helix}}(x, N)p_{\text{helix}}(1, y)}{p_{\text{helix}}(1, N)}, \quad (19)$$

we can simply use the peak volume as its own upper bound:

$$\tilde{p}_v(a, b) = p_v(a, b) = \frac{p_v(a)p_v(b)}{p_{\text{helix}}(1, N)}. \quad (20)$$

The $\Xi(x, y)$ factor [29] is the counterpart of $\Omega(y - x)$, but an explicit consideration of its monotonicity is not necessary here, because it is absorbed in the above quantities. A next bigger goat is then calculated and applied in the same way as for bubbles.

Results and Discussion

Time complexity

By inspection of Algorithm 2, we observe that it visits at least $O(N)$ and at most $O(N^2)$ matrix elements of G . Furthermore, it performs sorting, which is known to scale as $O(N \log N)$. The time complexity is therefore between $O(N \log N)$ and $O(N^2)$. The execution time depends on the fraction of ignored grid frames above the diagonal, which depends on the specific sequence, temperature, and other input parameters. A theoretical analysis of these dependencies is complicated.

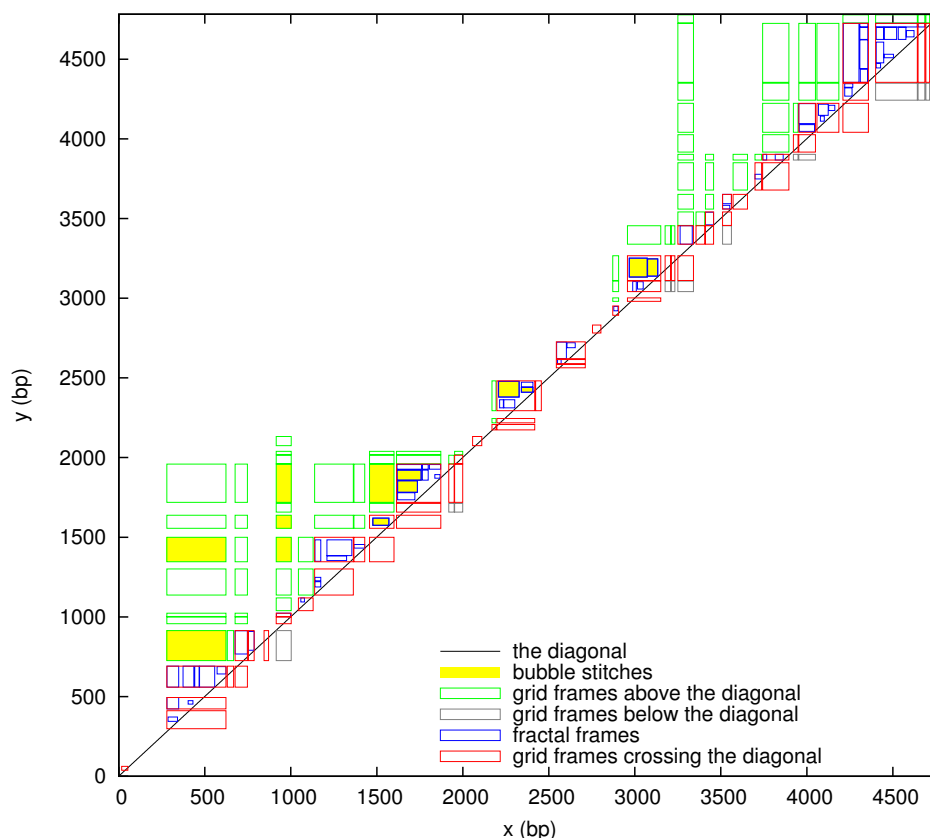


Figure 6

The footprints of Algorithm 2 plotted in the xy -plane. These are frames that are visited by Algorithm 2 during its search for the bubble stitches (filled yellow). The frames are located in a band along the diagonal, suggesting that the search space is proportional to sequence length. Grid frames below the diagonal (grey) are skipped. Grid frames crossing the diagonal (red) are broken into fractal frames (blue). The bubble stitches (filled yellow) are those grid frames above the diagonal (green) and fractal frames (blue) that have $p_v(a, b) \geq p_c$. Frames with side lengths below 20 bp are not shown to unclutter the figure.

Empirical testing of the execution times were done instead, using a test set of 14 biological sequences with lengths selected to be evenly spread on a log scale spanning three decades. A minimum length of 1000 bp was required. Most of the test sequences are genomic sequences, so as to represent the typical usage of the algorithm. The sequence lengths and accession numbers are:

- 1168 bp [GenBank:BC108918]
- 1986 bp [GenBank:BC126294]
- 4781 bp [GenBank:BC039060]
- 7904 bp [GenBank:NC_001526]
- 16571 bp [GenBank:NC_001807]
- 36001 bp [GenBank:AC_000017]
- 48502 bp [GenBank:NC_001416]
- 85779 bp [GenBank:NC_001224]
- 168903 bp [GenBank:NC_000866]
- 235645 bp [GenBank:NC_006273]
- 412348 bp [GenBank:AE001825]
- 816394 bp [GenBank:NC_000912]
- 1138011 bp [GenBank:AE000520]
- 2030921 bp [GenBank:NC_004350]

The algorithms were written in Perl and run on a Pentium 4, 2.4 GHz, 512 KB cache, 1 GB memory, PC with Linux (CentOS). In Figure 7, the speeds of Algorithms 1 and 2

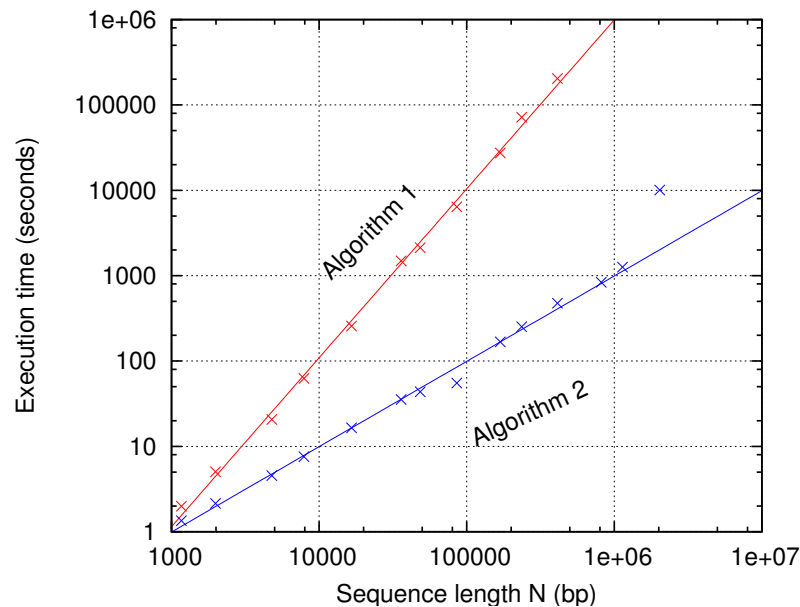


Figure 7

Algorithm 1 is quadratic and Algorithm 2 is linear. The log-log plot shows the execution time versus sequence length of Algorithm 1 (red) and Algorithm 2 (blue). The straight lines are fits to the data points with slopes 1.97955 ± 0.02923 (red) and 0.99953 ± 0.02016 (blue).

are compared. Algorithm 2 is orders of magnitude faster than Algorithm 1 for sequences longer than 100 kbp. While all the 14 sequences were computed by Algorithm 2, the three longest sequences were aborted by Algorithm 1, because of too long execution times. To ensure that the computational tasks were comparable, all sequences were computed at their melting temperatures T_m , rather than one temperature for all, such that all sequences had the same fractions of helical regions and bubbles. For both algorithms, straight lines were fitted to the data in the log-log plot. For Algorithm 2, however, the longest sequence (2 Mbp) is considered an outlier and thus excluded from the fit. This sequence's execution time was overly increased, because the required memory exceeded the available 1 gigabyte RAM. For Algorithm 1, the slope of the fit is 1.97955 ± 0.02923 , suggesting that it has time complexity $O(N^2)$. For Algorithm 2, the slope of the fit is 0.99953 ± 0.02016 . This is interpreted as the time complexity $O(N \log N)$, but with the logarithmic component being too weak to distinguish $O(N \log N)$ from $O(N)$.

The execution time of Algorithm 2 is just as much a property of the underlying energy landscape depending on the input, as it is a property of the algorithm. Could it be that other input parameters and/or sequences than was used in Figure 7 – say, away from the melting points – would exhibit the time complexity $O(N^2)$? Figure 8 shows the speed of Algorithm 2 over the whole melting range of

temperatures. Each sequence in the test set was computed at temperatures corresponding to the helicity values: 0.9995, 0.999, 0.995, 0.99, 0.95, 0.9, 0.8, 0.7, ..., 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, and 0.0005. This helicity range approximately corresponds to the temperature range $T_m \pm 10^\circ\text{C}$ and it covers most of the melting transitions. Although the curves for the individual helicity values may not be easily distinguished in Figure 8, it appears that all curves have similar slopes and that they are close to each other, i.e., the variation in execution time is below 50%. This indicates that the helicity (or temperature) value has only a small influence on the total execution time. The time complexity $O(N \log N)$ seems to be robust.

However, a stronger temperature dependence is revealed when considering the computations of bubble stitches and helix stitches separately. Two independent subroutines of Algorithm 2 compute the bubble stitches and the helix stitches, both following the procedure outlined in the previous section. The rest of Algorithm 2's computation, including the initial computation of at least four partition function arrays [30], is called the overhead. Correspondingly, the total execution time t_{total} is the sum of the bubble execution time t_{bubble} , the helix execution time t_{helix} and the overhead execution time t_{overhead} . By simply switching off the bubble subroutine (i.e. $t_{\text{bubble}} = 0$) and measuring the total execution time, we obtain $t_{\text{helix}} + t_{\text{overhead}}$. Likewise, by switching off the helix subroutine,

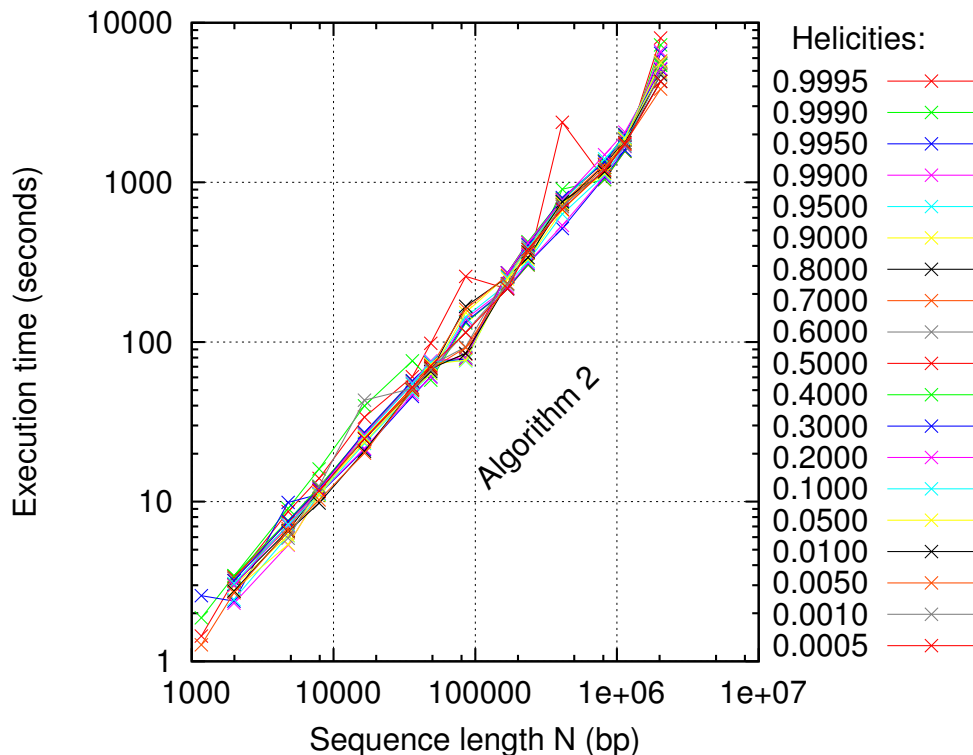


Figure 8
Algorithm 2 is fast at all temperatures. The total execution times are plotted versus sequence length for each of the listed helicity values.

we measure $t_{\text{bubble}} + t_{\text{overhead}}$. In the following, we refer to $t_{\text{bubble}} + t_{\text{overhead}}$ as the bubble time and $t_{\text{helix}} + t_{\text{overhead}}$ as the helix time. As an example, Figure 9 shows the results for the 16571 bp [GenBank:NC_001807]. The bubble and helix times are divided by sequence length and plotted as a function of temperature. Both of them have clearly a strong temperature dependence. The melting curve is also plotted in Figure 9, indicating that most of the melting occurs in the temperature range 80–85°C. Plots like Figure 9 were made for each sequence in the test set, but the average behavior is more interesting. To average times of the order $O(N)$ over sequences of different lengths, one should divide them by sequence length as in Figure 9. However, to plot as a function of temperature would not be meaningful, because the sequences have different T_m 's and different melting ranges. On the horizontal axis, instead, we use a normalized temperature,

$$\tau = \log\left(\frac{1-\Theta}{\Theta}\right), \tag{21}$$

defined such that the melting curve becomes a sigmoid:

$$\Theta = \frac{1}{1+\exp(\tau)}. \tag{22}$$

For each τ -value (or equivalently for each Θ -value), the bubble times and helix times divided by sequence length averaged over all sequences are plotted in Figure 10. The curves have a similar temperature dependence as in Figure 9. The helix time decreases monotonically (except for a shoulder), while the bubble time increases monotonically (except for a shoulder). Both of them have an about four-fold difference between their maximum and minimum. Qualitatively, the curves are kind of mirror symmetric, but the helix time is generally greater than the bubble time, the two curves cross each other at $\Theta = 0.12$. It seems that adding the two curves would give a more or less horizontal curve, i.e., the total execution time has much less temperature dependence.

We may understand this interchange between bubble time and helix time in terms of the melting process. If we assume that the bubble time is proportional to the area of the footprint in Figure 6, and that this is proportional to the average length of potential bubbles at that tempera-

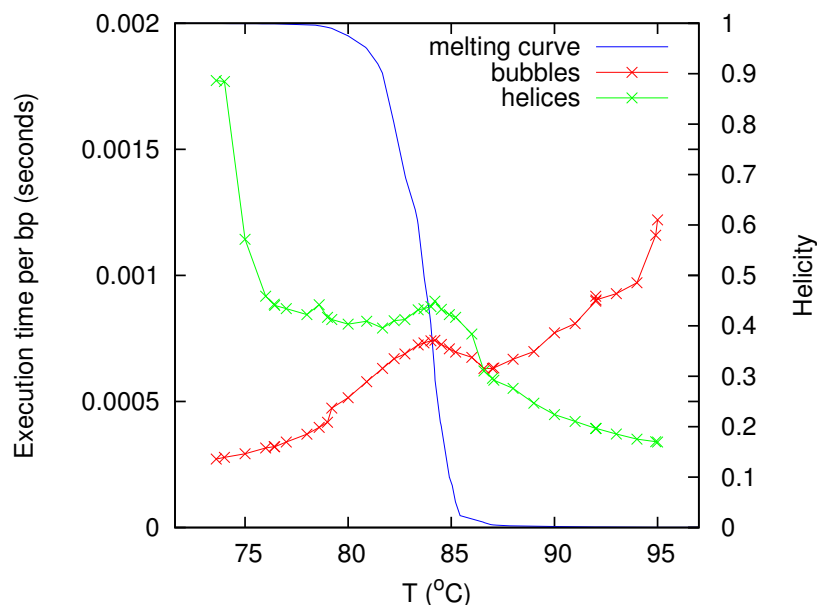


Figure 9

Bubble and helix execution times versus temperature. For the sequence [GenBank:NC_001807], the bubble time (red) and helix time (green) divided by sequence length (16571 bp) is plotted versus T . The melting curve (blue) shows the helicity Θ (on the right vertical axis) as a function of T , indicating the melting midpoint: $\Theta = 0.5$ at $T_m = 83.7^\circ\text{C}$.

ture, then we would expect the bubble time to increase with temperature, because bubbles grow as DNA melts. Likewise, we would expect the helix time to decrease with temperature, because helical regions diminish as DNA melts.

In this article, Blake & Delcourt's parameter set [42] as modified by Blossy & Carlon [43] was used with $[\text{Na}^+] = 0.075\text{ M}$. Parameters for the loop entropy approximation was obtained with our online tool [31,44]. The maximum depth and probability cutoff parameters were $D_{\max} = 5$ and $p_c = 0.01$ in Figs. 3, 4, 5, 6, $D_{\max} = 3$ and $p_c = 0.02$ in Figure 7, and $D_{\max} = 3$ and $p_c = 0.0001$ in Figs. 8, 9, 10. The sequence [GenBank:BC039060] was used for producing Figs. 2, 3, 4, 5, 6. A systematic test of how the execution time depends on D_{\max} and p_c has not been performed.

Discussion

For an algorithm to be called efficient, it should solve the task at hand with optimal time complexity. It should not introduce approximations, that would just amount to a reformulation of a simpler, but different task. In this study, the task is to compute a stitch profile based on the Poland-Scheraga model with Fixman-Freire loop entropies. With this model, the time complexity must be at least $O(N \log N)$. Indeed, this is achieved by Algorithm 2 under a wide range of conditions. Algorithm 2 does not

acquire a speedup by any *windowing* approximation, by which the sequence would first be split into smaller independent sequences. Neither does it rely on limiting the problem to a maximal bubble length. Therefore, Algorithm 2 is efficient. In computational RNA and protein studies, a maximal loop size is sometimes imposed as a heuristic for reducing time complexity by one order. Similarly, a maximal DNA bubble size of 50 bp has been reported in computations of low temperature bubble probabilities in the Peyrard-Bishop-Dauxois model [14]. In contrast, Algorithm 2 can find bubbles of whatever size at any temperature. In the 48502 bp [GenBank:NC_001416], for example, bubbles and helical regions may be up to around 20000 bp long [45]. Although Algorithm 2 has no explicit notion of a maximal bubble length, it may implicitly detect length limitations for both bubbles and helical regions by the absence of the "next bigger goat". In this way, Algorithm 2 can adapt to the input sequence. This adaptation is evident in Figure 10, where the bubble execution time grows as bubbles get bigger at higher temperatures. Conversely, the helix execution time decreases as the helical regions gradually melt away.

However, the time complexity was not proven to be $O(N \log N)$ under all conditions. It is still an open question whether there is a transition to time complexity $O(N^2)$ in

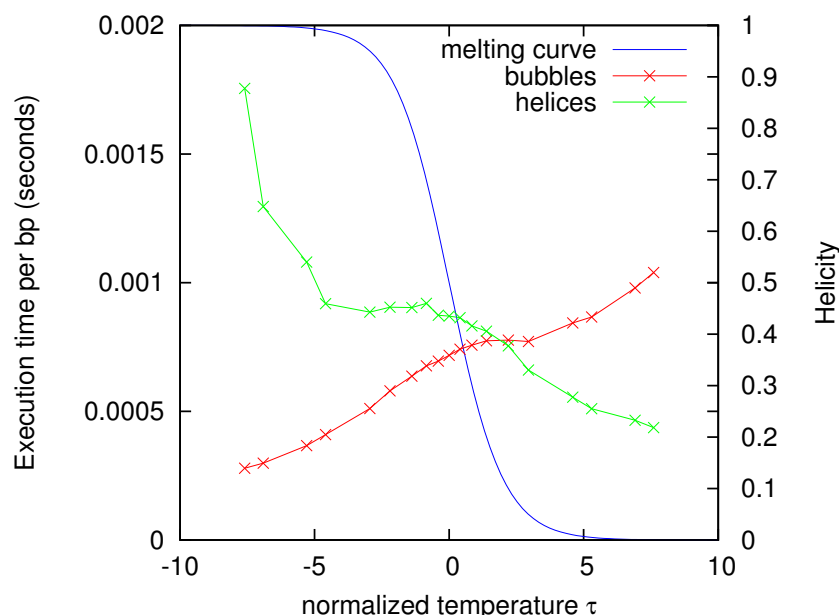


Figure 10

Sequence-averaged bubble and helix execution times. The bubble time per basepair (red) and helix time per basepair (green) averaged over all sequences are plotted versus the normalized temperature τ for each of the helicity values listed in Figure 8. The melting curve (blue) shows the helicity Θ (on the right vertical axis) as a function of τ , indicating the melting midpoint: $\Theta = 0.5$ at $\tau = 0$.

some peripheral regions of the input parameter space. But based on results so far, a fast computation would be expected in most situations.

How fast is Algorithm 2? Figures 7 and 8 show that the Perl implementation runs on an old desktop PC at the speed of roughly 1000 basepairs per second. With today's computers, assuming twice that speed and enough memory, the *E. coli* genome would take 39 minutes, the yeast genome would take 1.7 hours, and the largest human chromosome would take 35 hours. In some types of low temperature melting studies, the features of interest are the bubbles rather than the helical regions. In such applications, switching off the computation of helix stitches can speed up the algorithm several times. As Figure 10 indicates, the helix time is about twice the bubble time at helicity equal to 0.95, that is, the speedup would be about threefold. The largest human chromosome would be done in ten hours. On a computer cluster, the human genome could be computed in a day. Such bubbles could then be compared to TFBS, TSS, replication origins, viral integration sites, etc.

The required memory grows with sequence length and for sequences longer than 2 Mbp, more than 1 GB was needed. The memory usage has not been tested further

and the space complexity has not been discussed in this article. Some memory optimization of the Perl implementation must be done before such test can reflect the space complexity. While the algorithm is efficient in terms of time complexity, the code has room for optimization of both speed and memory usage. However, the space complexity is believed to be $O(N)$, which means that the algorithm would eventually become out of memory for long enough sequences. A standard solution is to introduce efficient use of disk space instead, which could reduce the memory usage to $O(1)$, without increasing the time complexity.

The fast algorithm presented in this paper relies on a factorization of certain upper bounds, which in turn relies on a factorization of partition functions in the Poland-Scheraga model, see Eq. (16). In general, it seems that a factorization of partition functions is essential for solving DNA and RNA models in polynomial time. However, some DNA melting models that explicitly consider supercoiling do not allow such a factorization [46]. For the purpose of genomic applications, supercoiling and a number of other *in vivo* interactions and constraints in the cell should ideally be accounted for in future DNA model development [1,15]. Such modelling requires quantitative knowledge yet to be obtained experimentally. When such data

becomes available, a main challenge will be to develop models and algorithms that can be solved in time $O(N \log N)$, which is necessary for many genomic applications.

Conclusion

The fast algorithm described in this article enables the computation of stitch profiles of genomic sequences. Melting features of interest, such as bubbles, helical regions, and their boundaries, are computed directly, rather than relying on visualization or educated guesses. The algorithm is exact. It does not achieve its speed by approximations, such as windowing or maximal bubble sizes. Genomewide comparisons of bubbles with TSS, replication origins, viral integration sites, etc., are proposed. The algorithm is available in Perl code from the author. Online computation of stitch profiles is available on our web server, which has recently been upgraded to run Algorithm 2 [31,44].

Competing interests

The author declares that they have no competing interests.

Acknowledgements

Discussions with Eivind Hovig, Geir Ivar Jerstad, and Torbjørn Rognes on genome browsers gave the impetus to this work. Funding to this work was provided by FUGE-The national programme for research in functional genomics in Norway.

References

- Calladine CR, Drew HR, Luisi BF, Travers AA: *Understanding DNA. The molecule and how it works* London: Elsevier academic press; 2004.
- Sumner AT: *Chromosomes. Organization and function* Oxford: Blackwell; 2003.
- Tong BY, Battersby SJ: **Melting curves, denaturation maps, and genetic map of ϕ X174: Their relationships and applications.** *Biopolymers* 1979, **18**:1917-1936.
- Wada A, Suyama A: **Stability distribution in the phage λ -DNA double helix: A correlation between physical and genetic structure.** *J Biomol Struct Dyn* 1984, **2**(3):573-591.
- King GJ: **Stability, structure and complexity of yeast chromosome III.** *Nucl Acids Res* 1993, **21**(18):4239-4245.
- Marx KA, Assil IQ, Bizzaro JW, Blake RD: **Comparison of experimental to MELTSIM calculated DNA melting of the (A+T) rich Dictyostelium discoideum genome: denaturation maps distinguish exons from introns.** *J Biomol Struct Dyn* 1998, **16**:329-39.
- Yeramian E: **Genes and the physics of the DNA double-helix.** *Gene* 2000, **255**:139-150.
- Yeramian E: **The physics of DNA and the annotation of the Plasmodium falciparum genome.** *Gene* 2000, **255**:151-168.
- Yeramian E, Jones L: **GeneFizz: a web tool to compare genetic (coding/non-coding) and physical (helix/coil) segmentations of DNA sequences.** *Nucleic Acids Res* 2003, **31**:3843-3849.
- Yeramian E, Bonnefoy S, Langsley G: **Physics-based gene identification: proof of concept for plasmodium falciparum.** *Bioinformatics* 2002, **18**:190-193.
- Carlson E, Malki ML, Blossey R: **Exons, Introns, and DNA Thermodynamics.** *Physical Review Letters* 2005, **94**(17):178101.
- Carlson E, Dkhissi A, Malki ML, Blossey R: **Stability domains of actin genes and genomic evolution.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2007, **76**(5 Pt 1):051916.
- Choi CH, Kalosakas G, Rasmussen KØ, Hiromura M, Bishop AR, Usheva A: **DNA dynamically directs its own transcription initiation.** *Nucleic Acids Res* 2004, **32**:1584-1590.
- van Erp TS, Cuesta-Lopez S, Haggmann JG, Peyrard M: **Can One Predict DNA Transcription Start Sites by Studying Bubbles?** *Physical Review Letters* 2005, **95**(21):218104.
- Benham CJ, Singh RRP: **Comment on "Can One Predict DNA Transcription Start Sites by Studying Bubbles?".** *Physical Review Letters* 2006, **97**(5):059801.
- van Erp TS, Cuesta-Lopez S, Haggmann JG, Peyrard M: **van Erp et al. Reply.** *Physical Review Letters* 2006, **97**(5):059802.
- Choi CH, Usheva A, Kalosakas G, Rasmussen KØ, Bishop AR: **Comment on "can one predict DNA transcription start sites by studying bubbles?".** *Physical Review Letters* 2006, **96**(23):239801.
- van Erp TS, Cuesta-Lopez S, Haggmann JG, Peyrard M: **van Erp et al Reply.** *Physical Review Letters* 2006, **96**(23):239802.
- Altan-Bonnet G, Libchaber A, Krichevsky O: **Bubble dynamics in double-stranded DNA.** *Phys Rev Lett* 2003, **90**(1-4):138101.
- Ambjörnsson T, Banik SK, Krichevsky O, Metzler R: **Breathing dynamics in heteropolymer DNA.** *Biophys J* 2007, **92**(8):2674-2684.
- Benham C: **Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci.** *Proc Natl Acad Sci USA* 1993, **90**(7):2999-3003.
- Wang H, Noordewier M, Benham C: **Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters.** *Genome Res* 2004, **14**(8):1575-84.
- Ak P, Benham CJ: **Susceptibility to superhelically driven DNA duplex destabilization: A highly conserved property of yeast replication origins.** *PLoS Computational Biology* 2005, **1**:e7.
- Benham CJ: **Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions.** *Journal of Molecular Biology* 1996, **255**:425-434.
- Wang H, Kaloper M, Benham C: **SIDDBASE: a database containing the stress-induced DNA duplex destabilization (SIDD) profiles of complete microbial genomes.** *Nucleic Acids Res* 2006:D373-8.
- Liu F, Tøstesen E, Sundet JK, Jenssen TK, Bock C, Jerstad GI, Thilly WG, Hovig E: **The human genomic melting map.** *PLoS Computational Biology* 2007, **3**(5):e93.
- Jerstad GI: **Merging the physical properties of DNA with genomic annotations in Ensembl.** In *Master's thesis* University of Oslo, Institute of Informatics; 2006.
- Poland D: **Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations.** *Biopolymers* 1974, **13**:1859-1871.
- Tøstesen E: **Partly melted DNA conformations obtained with a probability peak finding method.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **71**(6 Pt 1):061922.
- Tøstesen E, Liu F, Jenssen TK, Hovig E: **Speed-up of DNA melting algorithm with complete nearest neighbor properties.** *Biopolymers* 2003, **70**:364-376.
- Tøstesen E, Jerstad GI, Hovig E: **Stitchprofiles.uio.no: Analysis of partly melted DNA conformations using stitch profiles.** *Nucleic Acids Res* 2005, **33**:w573-w576.
- Fixman M, Freire JJ: **Theory of DNA melting curves.** *Biopolymers* 1977, **16**:2693-2704.
- Dill KA, Chan HS: **From Levinthal to pathways to funnels.** *Nat Struct Mol Biol* 1997, **4**:10-19.
- Wales DJ, Miller MA, Walsh TR: **Archetypal energy landscapes.** *Nature* 1998, **394**:758-760.
- Stelling J, Sauer U, Szallasi Z, Doyle FJ III, Doyle J: **Robustness of cellular functions.** *Cell* 2004, **118**:675-685.
- Hoffmann KH, Sibani P: **Diffusion in hierarchies.** *Phys Rev A* 1988, **38**(8):4261-4270.
- Stein DL, Newman CM: **Broken ergodicity and the geometry of rugged landscapes.** *Phys Rev E* 1995, **51**(6):5228-5238.
- Flamm C, Hofacker IL, Stadler PF, Wolfinger MT: **Barrier trees of degenerate landscapes.** *Z Phys Chem* 2002, **216**:155-173.
- Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF: **Exact Folding Dynamics of RNA Secondary Structures.** *J Phys A: Math Gen* 2004, **37**:4731-4741.
- Dimitrov RA, Zuker M: **Prediction of hybridization and melting for double-stranded nucleic acids.** *Biophys J* 2004, **87**:215-226.
- Yeramian E, Schaeffer F, Caudron B, Claverie P, Buc H: **An optimal formulation of the matrix method in statistical mechanics of one-dimensional interacting units: efficient iterative algorithmic procedures.** *Biopolymers* 1990, **30**:481-497.

42. Blake RD, Delcourt SG: **Thermal stability of DNA.** *Nucleic Acids Res* 1998, **26**:3323-3332.
43. Blosser R, Carlon E: **Reparametrizing the loop entropy weights: Effect on DNA melting curves.** *Phys Rev E* 2003, **68**(6):061911.
44. **DNA Melting - stitchprofiles.uio.no** [<http://stitchprofiles.uio.no>]
45. **Stitch Profiles of Bacteriophage Lambda DNA** [ftp://ftp.aip.org/epaps/phys_rev_e/E-PLFFF8-71-148506/lambda_stitch.htm]
46. Benham C, Bi C: **The analysis of stress-induced duplex destabilization in long genomic DNA sequences.** *J Comput Biol* 2004, **11**(4):519-43.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

