Segmentation with an isochore distribution*

Miklós Csűrös[1], Ming-Te Cheng[1], Andreas Grimm[2], Amine Halawani[1], and Perrine Landreau[3]

[1] Department of Computer Science and Operations Research, Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal, Québec, Canada, H3C 3J7
csuros@iro.umontreal.ca
[2] Lehr- und Forschungseinheit für Bioinformatik
Ludwig-Maximilians-Universität München, 80333 München, Germany
[3] Insitut Scientifique Polytechnique Galilée — Université Paris XIII
93430 Villetaneuse, France

Abstract. We introduce a novel generative probabilistic model for segmentation problems in molecular sequence analysis. All segmentations that satisfy given minimum segment length requirements are equally likely in the model. We show how segmentation-related problems can be solved with similar efficacy as in hidden Markov models. In particular, we show how the best segmentation, as well as posterior segment class probabilities in individual sequence positions can be computed in $O(nC)$ time in case of C segment classes and a sequence of length n.

1 Introduction

Let $\mathbf{x} = x_1 x_2 \cdots x_n$ be a sequence of characters over a finite alphabet $\mathcal{A}$. A *segmentation* of $\mathbf{x}$ is described as a sequence $\mathbf{z} = z_1 z_2 \ldots z_n$ that assigns a *segment class* to each sequence position. The segmentation is thus a sequence over an alphabet $\mathcal{C}$, where $\mathcal{C}$ is the set of segment classes. A *segment* is a maximal contiguous region of positions that belong to the same class. Many molecular sequence analysis problems can be formulated as segmentation problems [1]. Obvious examples include the identification of isochores [2] in genomic DNA, and identification of charge clusters and hydrophobic profiles for proteins. In principle, all sequence annotation tasks (with non-overlapping segments) fit this general segmentation framework. For example, even such a complex task as eukaryotic gene prediction [3], entails the segmentation of a genomic sequence into classes such as "intergenic" and "exonic." In this work we are interested in generative probabilistic models, when the sequence $\mathbf{x}$ is the observed value of a random variable that depends solely on $\mathbf{z}$, which is also a random instance. Furthermore, we assume independence in the sense that each x_i depends on z_i only. Such probabilistic models include hidden Markov models [4, 5], and other notable examples [6, 7]. Hidden Markov models (HMMs) have the computational advantage that various segmentation-related problems, including that of finding

* This work is supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

the most likely segmentation, can be solved with linear-time algorithms in the sequence length n.

This paper's main goal is to introduce a new class of prior segmentation distributions; namely, a uniform distribution over segmentations in which all segments are longer than some specified threshold. Such a distribution captures usual expectations from segmentation results. We show that it is possible to compute the most likely segmentation in linear time in n, while the minimum segment length does not affect the running time. We show the same asymptotic running times for computing the posterior probabilities for segment class memberships and segment boundaries. In other words, we describe the analogues of the Viterbi and forward-backward algorithms.

An important motivation for our segmentation model comes from the *isochore theory* [8]. It postulates that the genome of warm-blooded vertebrates is composed of *isochores* in a mosaic structure. An isochore is a long contiguous segment of genomic DNA with a "fairly homogeneous" guanine+cytosine (GC) content [9]. The old debate about the theory's utility reemerged at the completion of the human genome draft sequence and persists to this day [9–13]. Eyre-Walker and Hurst [14] review biologically relevant issues in conjunction with isochores. We do not want to settle the question of biological relevance, but rather treat isochores as a technically useful concept describing the "fairly homogeneous" GC content of a region within an environment of at least 50–300 thousand base pairs. Usual isochore computations involve sliding windows of fixed length [10, 11, 13, 14]. Window-less methods usually correspond to the minimization of some segment homogeneity measure [2, 15]. To our knowledge, no generative model exists until now that explicitly captures the notions of minimum length and homogeneity at the same time. Here we put forward such a minimalist model, along with relevant computations.

1.1 Model and model selection

First we describe a generative framework for defining segmentation problems. A sequence of random variables $\mathbf{X} = (X_i \colon i = 1, \ldots, n)$ is dependent on a sequence of (unknown) segment class memberships $\mathbf{Z} = (Z_i \colon i = 1, \ldots, n)$. Here $X_i \in \mathcal{A}$ are letters from a finite alphabet and $Z_i \in \mathcal{C}$ are segment classes. The possible segment classes $\mathcal{C}$ are known. From an observed sequence $\mathbf{x} = x_1 \cdots x_n$, we want to deduce a segmentation $\mathbf{z} = z_1 \cdots z_n$. The human genome is often analyzed in terms of isochores named L1, L2, H1, H2, H3 with typical GC level cutoffs of 0.37, 0.41, 0.46, 0.53. In our probabilistic framework, a human chromosome sequence forms $\mathbf{x}$, and $\mathcal{C}$ comprises isochore classes.

More or less general versions of this framework were considered in the statistical literature [6, 16]. They usually involve $\Omega(n^2)$-time computations for determining optimal segmentations [16, 17]. Optimality is measured by some fitness or homogeneity measure. We focus on cases when the optimal segmentation can be found efficiently by some reasonable principle. First of all, we assume *indepen-*

dence: the distribution of each X_i is completely determined by the probabilities

$$p_z(x) = \mathbb{P}\Big\{ X_i = x \;\Big|\; Z_i = z \Big\}.$$

A direct likelihood maximization approach cannot be used to choose a hypothesis $\mathbf{z}$, since the likelihood is maximized when each $z_i = \max_z p_z(x_i)$, which is rarely a consistent estimation. (For example, in GC content analysis, the best segmentation is a binary sequence of two classes for 100% and 0% GC.) We discuss two main principles that lead to better estimates without overfitting. The first principle is a Bayesian one: by imposing a prior distribution on $\mathbf{Z}$, one can select $\mathbf{z}$ that maximizes the posterior probability $\mathbb{P}\Big\{ \mathbf{Z} = \mathbf{z} \;\Big|\; \mathbf{X} = \mathbf{x} \Big\}$. This principle is employed in hidden Markov models. If $\mathbf{Z}$ is a Markov chain with a finite state set $\mathcal{C}$, then the best segmentation can be found in $O(n|\mathcal{C}|)$ time using the Viterbi algorithm [4, 5]. An alternative principle is to incorporate a notion of complexity in the optimization. For instance, the likelihood can be combined with description length [18], which penalizes complicated segmentations. When $\mathcal{C}$ is finite, and the segmentation's complexity is measured by the number of its segments, the best segmentation can be found efficiently in $O(n|\mathcal{C}|)$ time [7]. When $\mathcal{C}$ is the set of all possible distributions over $\mathcal{A}$, then the best segmentation minimizes the entropy with an adequate complexity penalization [2, 15].

The Bayesian approach of imposing a prior distribution on $\mathbf{Z}$ has the methodological advantage that it enables one to define probabilities of the type $\mathbb{P}\Big\{ \chi(\mathbf{Z}) \;\Big|\; \mathbf{X} = \mathbf{x} \Big\}$, where $\chi(\cdot)$ is some "interesting" property. Interesting properties include segment boundaries ($\chi(\mathbf{z}) = \{z_{i-1} = z'; z_i = z\}$) and the class of a position ($\chi(\mathbf{z}) = \{z_i = z\}$). Concerning the notation $\chi(\cdot)$, we use events and their indicators interchangeably, and, thus, $\{z_i = z\}$ denotes both the event that position i belongs to class z and the indicator variable which takes the value of 1 or 0, when the event occurs or not, respectively.

2 Isochore distribution

In what follows, we focus on the case when $\mathbf{Z}$ is uniformly distributed over all segmentations satisfying certain minimum segment length requirements. We call such a distribution an *isochore distribution*. When the segmentation prior is uniform over a set $\mathcal{Z}$, the posterior probabilities can be computed as

$$\mathbb{P}\Big\{ \chi(\mathbf{Z}) \;\Big|\; \mathbf{X} = \mathbf{x} \Big\} \propto \sum_{\mathbf{z} \in \mathcal{Z} \cap \chi(\mathbf{z})} \mathbb{P}\Big\{ \mathbf{X} = \mathbf{x} \;\Big|\; \mathbf{Z} = \mathbf{z} \Big\}, \tag{1}$$

since $\mathbb{P}\{\mathbf{X} = \mathbf{x}\}$ does not depend on $\mathbf{z}$ and $\mathbb{P}\{\mathbf{Z} = \mathbf{z}\}$ is the same for every choice of $\mathbf{z} \in \mathcal{Z}$. In our case, the main difficulty is the efficient enumeration of segmentations that satisfy the minimum length requirements when the segmentation value is fixed in a position.

We are interested in segmentations where segments of class $z \in \{1, \ldots, C\}$ are of minimum length $m_z > 0$. The notion of minimum segment length is captured

through the following notation. We define $\mathsf{left}(\mathbf{z}, i)$ as the number of positions to the left of i that belong to the same segment class, and $\mathsf{right}(\mathbf{z}, i)$ as the number of positions to the right that belong to the same segment class. Formally,

$$\mathsf{left}(\mathbf{z}, i) = \big(\min_{d>0}\{d\colon z_{i-d} \neq z_i\}\big) - 1; \qquad \mathsf{right}(\mathbf{z}, i) = \big(\min_{d>0}\{d\colon z_{i+d} \neq z_i\}\big) - 1.$$

We extend the notation so that $z_i = 0$ whenever $i \leq 0$ or $i > n$: if $z_j = z$ for all $j \leq i$ then $\mathsf{left}(\mathbf{z}, j) = j - 1$ for all $j \leq i$, and an analogous statement holds for $\mathsf{right}()$ in the rightmost segment. Clearly, the length of the segment that includes position i is the value

$$\mathsf{length}(\mathbf{z}, i) = \mathsf{left}(\mathbf{z}, i) + \mathsf{right}(\mathbf{z}, i) + 1.$$

Definition 1. *Let $m_1, \ldots, m_C > 0$ be the minimum segment lengths for the segment classes. A segmentation $\mathbf{z}$ is* valid *if and only if*

$$\mathsf{length}(\mathbf{z}, i) \geq m_{z_i}$$

for all $i = 1, \ldots, n$. A random variable $\mathbf{Z}$ has an isochore distribution *if it is drawn uniformly from the set of valid segmentations.*

2.1 Number of valid segmentations

It is useful to compute the number of valid segmentations, since it defines our prior. Let $N_z(n)$ be the number of valid segmentations for a sequence of length n which end with a segment of class z, and let $N(n) = \sum_z N_z(n)$ be the total number of valid segmentations. These values can be computed exactly:

$$N_z(n) = \begin{cases} 0 & \text{if } n < m_z; \\ 1 & \text{if } n = m_z; \\ N_z(n-1) + \sum_{z' \neq z} N_{z'}(n - m_z) & \text{if } n > m_z. \end{cases}$$

For the particular case of $\forall z\colon m_z = m$, i.e., identical segment length thresholds, we have the recursion $N(n) = N(n-1) + (C-1)N(n-m)$ for $n > m$, with the initial values $N(n) = 0$ for $n < m$ and $N(m) = C$. Clearly, $N(n)$ grows exponentially with n. In general, $N(n) = \Theta(\beta^{n/m})$ where β is the root of the characteristic equation $\beta - \beta^{1-1/m} - (C-1) = 0$. The value $N(n)$ provides the normalizing value in Eq. (1) and can be used for normalization in upcoming formulas.

2.2 Computing the best segmentation

Finding the best segmentation under the isochore distribution prior is not difficult. The dynamic programming method outlined in [7] for $C = 2$ can be generalized to an arbitrary number C of classes. Define

$$\xi_z(i) = p_z(x_i) \qquad \text{and} \qquad \Xi_z(i, i') = \prod_{j=i}^{i'} \xi_z(j).$$

In other words, $\Xi_z(i', i)$ is the likelihood for a segment $i..i'$ in class z. We derive a dynamic programming algorithm for the variables $V_z(i)$ for all $z \in \{1, \ldots, C\}$ and $i = 1, \ldots, n$. The variable $V_z(i)$ gives the likelihood for the best segmentation that is valid within the prefix $x_{1..i}$ and ends with class $z_i = z$.

$$V_z(i) = \begin{cases} 0 & i < m_z; \\ \Xi_z(1, m_z) & i = m_z; \\ \max\Big\{\xi_z(i)V_z(i-1), \Xi_z(i-m_z+1, i)\max_{z'} V_{z'}(i-m_z)\Big\} & i > m_z. \end{cases} \quad (2)$$

After carrying out the computations for all z and i, the best segmentation ends with $\arg\max_z V_z(n)$ and previous classes can be found by tracing back the maxima in (2). An advantageous technique is to keep track of letter counts

$$c_a(i) = \sum_{j=1}^{i} \{x_j = a\}$$

for all $a \in \mathcal{A}$ and i and then compute $\Xi_z(i, j) = \prod_{a \in \Sigma} \big(p_z(a)\big)^{c_a(j) - c_a(i-1)}$ (with $c_a(0) = 0$). In order to reduce costly floating-point calculations, $\big(p_z(a)\big)^c$ should be computed beforehand for all $z \in \{1, \ldots, C\}$, $a \in \mathcal{A}$ and $c \in \{0, 1, \ldots, m\}$. One can also work with $\log V_z(i)$ instead to avoid underflow, and to expedite the computations by performing additions instead of multiplications.

Theorem 1. *A segmentation* $\mathbf{z}$ *that maximizes* $\mathbb{P}\Big\{\mathbf{Z} = \mathbf{z} \;\Big|\; \mathbf{X} = \mathbf{x}\Big\}$ *can be found in* $O(nC)$ *time when* $\mathbf{Z}$ *has an isochore distribution with* C *segment classes.*

2.3 Computing posteriors

For computing posterior probabilities, we need to be able to sample valid segmentations that are constrained at a position. In order to simplify the formulas, we assume from now on that the minimum segment lengths are identical, i.e., for all z, $m_z = m$, and that the minimum length m is an even number.

In order to derive recurrence relations, consider the following sets of (not necessarily valid) segmentations for $z \in \{1, \ldots, C\}$, $i \in \{1, \ldots, n\}$ and $d \in \{0, \ldots, n\}$

$$\mathcal{L}_z^{(d)}(i) = \Big\{\mathbf{z} \colon z_i = z; \mathsf{left}(\mathbf{z}, i) \geq d; \forall j < i - \mathsf{length}(\mathbf{z}, i) \colon \mathsf{length}(\mathbf{z}, j) \geq m\Big\};$$
$$\mathcal{R}_z^{(d)}(i) = \Big\{\mathbf{z} \colon z_i = z; \mathsf{right}(\mathbf{z}, i) \geq d; \forall j > i + \mathsf{length}(\mathbf{z}, i) \colon \mathsf{length}(\mathbf{z}, j) \geq m\Big\}.$$

In other words, $\mathcal{L}_z^{(d)}(i)$ is the set of segmentations that are restricted only for the prefix $z_1, \ldots, z_i$ so that (a) positions $i - d, \ldots, i$ are in class z, and (b) segments before the segment of i satisfy the minimum length requirements. The sets $\mathcal{R}_z^{(d)}(i)$

are defined analogously for suffixes of $\mathbf{z}$. Now, $\mathcal{L}_{z'}^{(m-1)}(i-1) \cap \mathcal{R}_{z}^{(m-1)}(i)$ is the set of valid segmentations that have a $z' \to z$ segment boundary at i. Hence, the posterior probability of a boundary at position $i > 1$ can be written as

$$\begin{aligned} q_{z' \to z}(i) &= \mathbb{P}\Big\{ Z_{i-1} = z'; Z_i = z \;\Big|\; \mathbf{X} = \mathbf{x} \Big\} \\ &\propto \mathbb{P}\Big\{ \mathbf{X} = \mathbf{x} \;\Big|\; \mathbf{Z} \in \mathcal{L}_{z'}^{(m-1)}(i-1) \cap \mathcal{R}_{z}^{(m-1)}(i) \Big\} \end{aligned}$$

when $z' \neq z$. It will be useful to define the posterior probabilities for position $1 < i < n$ being the left or right end of a segment in class z:

$$q_{\to z}(i) = \sum_{z' \neq z} q_{z' \to z}(i); \qquad 1 < i \leq n; \tag{3a}$$

$$q_{z \to}(i) = \sum_{z' \neq z} q_{z \to z'}(i+1); \qquad 1 \leq i < n. \tag{3b}$$

The posterior probability that position i belongs to class z is denoted by

$$q_z(i) = \mathbb{P}\Big\{ Z_i = z \;\Big|\; \mathbf{X} = \mathbf{x} \Big\}.$$

For the sake of completeness, we extend the notation of Eqs. (3) to the sequence extremities: $q_{\to z}(1) = q_z(1)$ and $q_{z \to}(n) = q_z(n)$.

Theorem 2. *Let* $\mu_z(i) = \mathbb{P}\Big\{ \mathbf{Z} \in \mathcal{L}_{z}^{(m/2)}(i) \cap \mathcal{R}_{z}^{(m/2)}(i) \;\Big|\; \mathbf{X} = \mathbf{x} \Big\}$. *For all* $i \in \{1, \ldots, n\}$ *and* $z \in \{1, \ldots, C\}$, *the probability that position* $1 < i < n$ *belongs to segment class* z *can be written as*

$$q_z(i) = \mu_z(i) + \sum_{\delta=0}^{\max\{i-1, \frac{m}{2}-1\}} q_{\to z}(i-\delta) + \sum_{\delta=0}^{\max\{n-i, \frac{m}{2}-1\}} q_{z \to}(i+\delta).$$

Proof. If $z_i = z$ and $\mathbf{z}$ is a valid segmentation, then exactly one of the following is true

1. $\mathsf{left}(\mathbf{z}, i) \geq m/2$ and $\mathsf{right}(\mathbf{z}, i) \geq m/2$ simultaneously;
2. position i's segment starts at position $i - \delta$ for some $0 \leq \delta < m/2$.
3. position i's segment ends at position $i + \delta$ for some $0 \leq \delta < m/2$.

The probability for Case 1 is $\mu_z(i)$. The probability of Case 2 is $\sum_\delta q_{\to z}(i-\delta)$; the probability of Case 3 is $\sum_\delta q_{z \to}(i+\delta)$. □

3 Algorithm for posterior probabilities

Define the following likelihoods

$$L_z(i) = \sum_{\mathbf{z}\in\mathcal{L}_z^{(m-1)}(i)} \mathbb{P}\Big\{X_{1..i-1} = x_{1..i-1} \;\Big|\; \mathbf{Z} = \mathbf{z}\Big\}; \tag{4a}$$

$$\lambda_z(i) = \sum_{\mathbf{z}\in\mathcal{L}_z^{(m/2)}(i)} \mathbb{P}\Big\{X_{1..i-1} = x_{1..i-1} \;\Big|\; \mathbf{Z} = \mathbf{z}\Big\}; \tag{4b}$$

$$R_z(i) = \sum_{\mathbf{z}\in\mathcal{R}_z^{(m-1)}(i)} \mathbb{P}\Big\{X_{i+1..n} = x_{i+1..n} \;\Big|\; \mathbf{Z} = \mathbf{z}\Big\}; \tag{4c}$$

$$\varrho_z(i) = \sum_{\mathbf{z}\in\mathcal{R}_z^{(m/2)}(i)} \mathbb{P}\Big\{X_{i+1..n} = x_{i+1..n} \;\Big|\; \mathbf{Z} = \mathbf{z}\Big\}; \tag{4d}$$

$$b_{z'\to z}(i) = \sum_{\mathbf{z}\in\mathcal{L}_{z'}^{(m-1)}(i-1)\cap\mathcal{R}_z^{(m-1)}(i)} \mathbb{P}\Big\{\mathbf{X} = \mathbf{x} \;\Big|\; \mathbf{Z} = \mathbf{z}\Big\}, \qquad i > 1. \tag{4e}$$

Clearly, $b_{z'\to z}(i) = L_{z'}(i-1)\xi_{z'}(i-1)\xi_z(i)R_z(i)$. whenever $1 < i \le n$. Let

$$b_{\to z}(i) = \sum_{z'\neq z} b_{z'\to z}(i) = \xi_z(i)R_z(i)\sum_{z'\neq z}\xi_{z'}(i-1)L_{z'}(i-1), \qquad i > 1;$$

$$b_{z\to}(i) = \sum_{z'\neq z} b_{z\to z'}(i+1) = \xi_z(i)L_z(i)\sum_{z'\neq z}\xi_{z'}(i+1)R_{z'}(i+1), \qquad i < n.$$

For the sequence extremities,

$$q_z(1) \propto b_{\to z}(1) = \xi_z(1)R_z(1); \tag{5a}$$

$$q_z(n) \propto b_{z\to}(n) = \xi_z(n)L_z(n). \tag{5b}$$

By Theorem 2, the posterior probabilities for segment class memberships can be computed for all $1 < i < n$ as

$$q_z(i) \propto \lambda_z(i)\xi_z(i)\varrho_z(i) + h_z(i), \tag{6}$$

where

$$h_z(i) = \sum_{\delta=0}^{\min\{i-1,\,\frac{m}{2}-1\}} b_{\to z}(i-\delta) + \sum_{\delta=0}^{\min\{n-i,\,\frac{m}{2}-1\}} b_{z\to}(i+\delta).$$

The right-hand sides of Eqs. (5) and (6) are normalized by dividing them with $Q = \sum_z \xi_z(1)R_z(1) = \sum_z \xi_z(n)L_z(n)$. In fact, posterior probabilities for segment boundaries are computed by the same normalization:

$$q_{\to z}(i) = Q^{-1}b_{\to z}(i) \qquad \text{and} \qquad q_{z\to}(i) = Q^{-1}b_{z\to}(i).$$

Additionally, since $\mathbb{P}\{\mathbf{Z} = \mathbf{z}\} = 1/N(n)$ for all $\mathbf{z}$, Bayes' theorem gives $\mathbb{P}\{\mathbf{X} = \mathbf{x}\} = \frac{Q}{N(n)}$.

The variables of Eqs. (4) are computed by the following recurrences.

$$\lambda_z(i) = \xi_z(i-1)\lambda_z(i-1) + \Xi_z(i-\frac{m}{2}, i-1) \qquad i > \frac{m}{2}+1 \quad (7a)$$
$$\times \sum_{z' \neq z} \xi_{z'}(i-\frac{m}{2}-1)L_{z'}(i-\frac{m}{2}-1);$$
$$L_z(i) = \xi_z(i-1)L_z(i-1) \qquad i > m \quad (7b)$$
$$+ \Xi_z(i-m+1, i-1) \sum_{z' \neq z} \xi_{z'}(i-m)L_{z'}(i-m);$$

Analogous formulas are used to compute $\varrho_z(i)$ and $R_z(i)$. If $\frac{m}{2} < i \leq n - \frac{m}{2} + 1$, then

$$h_z(i) = h_z(i-1) + b_{\rightarrow z}(i) - b_{\rightarrow z}(i-\frac{m}{2}) + b_{z\rightarrow}(i+\frac{m}{2}-1) - b_{z\rightarrow}(i-1). \quad (8)$$

Obviously, $h_z(1) = b_{\rightarrow z}(1)$. For $1 < i \leq \frac{m}{2}$ the recurrence of Eq. (8) does not include the subtraction of $b_{\rightarrow z}(i-\frac{m}{2})$ and for $i > n - \frac{m}{2} + 1$ the recurrence does not include the term $b_{z\rightarrow}(i+\frac{m}{2}-1)$. The variables of Eqs. (7) are initialized in an obvious manner.

A useful algorithmic technique for computing expressions of the type $A(z) = \sum_{z' \neq z} B(z')$ for all z in $O(C)$ total time is the following. First compute $B_{\rm lo}(z) = \sum_{z' < z} B(z')$ for all z. Then compute $B_{\rm hi}(z) = \sum_{z' > z} B(z')$ for all z. Clearly, this can be done in $O(C)$ time. Now, $A(z) = B_{\rm lo}(z) + B_{\rm hi}(z)$ can be set in $O(1)$ time for each z. Using this technique, all variables can be computed for every i in $O(nC)$ time. Notice that the Ξ_z can be computed in $O(1)$ time for all z, by keeping track of character counts in prefixes and suffixes as described in §2.2.

REMARK. It may seem that when the minimum lengths differ, $\sum_{z' \neq z} \xi_{z'}(i-m_z)L_{z'}(i-m_z)$ in (7b), for example, needs to be computed for each z separately, resulting in a $\Theta(C^2)$ factor in the running time. The technique, however, can be readily adapted to this case. The appropriate $B_{\rm lo}$ and $B_{\rm hi}$ values need to be kept for recent values of $j = i - m_z$, which again leads to a linear running time in C.

By the preceding discussion, we can state the following theorem.

Theorem 3. *All posterior probabilities for segment class memberships and segment boundaries can be computed in $O(nC)$ time when $\mathbf{Z}$ has an isochore distribution with C segment classes.*

The posterior probabilities can be used in an Expectation Maximization framework, as in Baum-Welch training for HMMs [4, 5]. Simply, the $p_z(x)$ are estimated as

$$\hat{p}_z(x) = \frac{\sum_{i=1}^{n} q_z(i)\{x_i = x\}}{\sum_{i=1}^{n} q_z(i)}.$$

3.1 Memory management

Since the recurrences for ϱ and R can be computed from right to left while those for λ, L and h are computed in a left to right direction, a direct implementation

would need to first compute and store the ϱ and R values and then proceed from left to right to carry out the posterior computations. The left-to-right computation proceeds in a "lookahead" fashion: for every i, $\lambda_z(i)$, $L_z(i+\frac{m}{2}-1)$, $h_z(i)$ and $q_z(i)$ are computed, in this order. Consequently, an array of size m can store the necessary values $L_z(j)$ for $i-\frac{m}{2} \leq j < i+\frac{m}{2}$ to carry out one step of the left-to-right computations. For λ_z and h_z, only the previous values are needed. It is, however, a good idea to keep track of recent values of $b_{z\rightarrow}$ and $b_{\rightarrow z}$ so that they are not computed twice.

A direct implementation, in which all $\varrho_z(i)$ and $R_z(i)$ are computed before proceeding to the left-to-right computations, may be impractical for large sequences because of large memory requirements. For longer sequences, it is possible to do the computations using a "slicing" or "checkpointing" technique, similar to those employed in pairwise sequence alignment and HMM training [19]. We do not discuss the details here due to space limitations. The technique allows for computing the probabilities on all-purpose desktop computers: our implementation was used to carry out the segmentations with five classes and $m = 50000$ for human chromosome 1 (246 Mbp), with a memory footprint below 2 Gigabytes. A recursive checkpointing technique leads to the result of the following theorem.

Theorem 4. *For C segment classes with minimum length m and a sequence of length n, the posterior probabilities can be computed in $O(LnC)$ time using $O\big(Cm^{1-1/L}n^{1/L}L\big)$ workspace, where L is an arbitrary positive integer. In particular, by choosing $L = \Theta\big(\log \frac{n}{m}\big)$, the probabilities are computed in $O(Cn \log \frac{n}{m})$ time using $O\big(Cm \log \frac{n}{m}\big)$ workspace.*

4 Experiments

We implemented the described procedure for posterior calculations in a Java package. Figure 1 compares in a simulated experiment the quality of HMM-based predictions and our method. The figure illustrates that HMM predictions are more easily affected by random fluctuations in the sequence composition.

For illustrative purposes, we carried out a segmentation of human chromosome 19 [20]. The results of the segmentation can be viewed as a custom annotation track in the UCSC genome browser [21]; the track can be downloaded from http://www.iro.umontreal.ca/~csuros/segmentation/hg17/chr19-segments.bed.

There are two principal questions that need to be addressed in this context: whether most of the genome can be classified into isochores, and whether there is a non-arbitrary threshold on homogeneous region lengths. Using five isochore classes, we segmented the sequence into segments within which the class membership can be established with at least 90% probability, using a minimum length of 50000 base pairs. About 85% of the sequence can be classified into one of the isochore classes with more than 90% fidelity. Almost all of the missing 15% fall into the unsequenced centromeric region, and the few percents that remain are mostly in short segment boundaries. This fact does not necessarily reflect the validity of classification, as long segments have a very small chance to fall

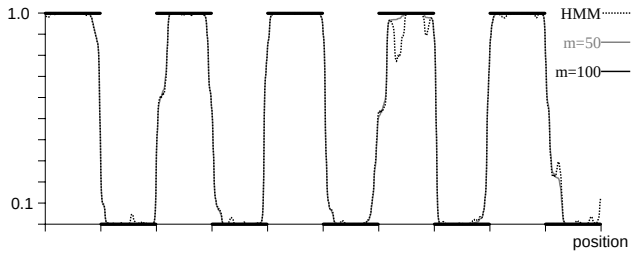

Fig. 1. Posterior segment class membership by HMM and isochore distributions. A random DNA sequence of 1000 characters was generated with alternating 30% and 70% GC level in 100bp segments. The plot compares the posterior segment class membership for the 30% GC class as computed by an HMM (two states, state switching transition probabilities are 0.01), and those computed using isochore distributions with minimum length 50 and 100. The former already gives smoother results (see especially the seventh segment), while the latter finds the true segmentation perfectly.

right between two classes in GC composition. Figure 2 plots the statistics on the segments. This chromosome is unusually GC-rich [20], 1.4%, 9.4%, 15.9%, 22.8% and 35.5% of the positions are classified into the classes L1, L2, H1, H2 and H3, respectively. It is interesting to notice that a large number of the segments have a length very close to the lower bound, which hints at heterogeneity below the minimum length cutoff. Classically, isochores are said to be hundreds of thousand base pairs in length: our segmentation does not reveal such a phenomenon.

5 Conclusion

We presented a novel probabilistic model for segmentations and showed how usual techniques associated with hidden Markov models have their equivalents, including a Viterbi-style algorithm for finding the best segmentation, a forward-backward algorithm for computing posteriors, and expectation maximization for setting class parameters. The model features an explicit minimum segment length parameter, which is not easily captured by an HMM. Our "minimalist" model assumes a uniform distribution among segmentations that obey the segment length constraints. Some additional parameters can be easily incorporated into the model. For instance, one can add conditional probabilities for changing segment classes, or have a segment length distribution that is a shifted geometric one. Using the example of Eq. (7b), write

$$L_z(i) = \tau_0 \xi_z(i-1) L_z(i-1) + \Xi_z(i-m+1, i-1) \sum_{z' \neq z} \tau_{z'} \xi_{z'}(i-m) L_{z'}(i-m).$$

The parameter τ_0 implies that segment length has a thresholded geometric distribution and the parameters $\tau_{z'}$ model different probabilities for the preceding segment class. In fact, such a parametrization is the equivalent of posterior

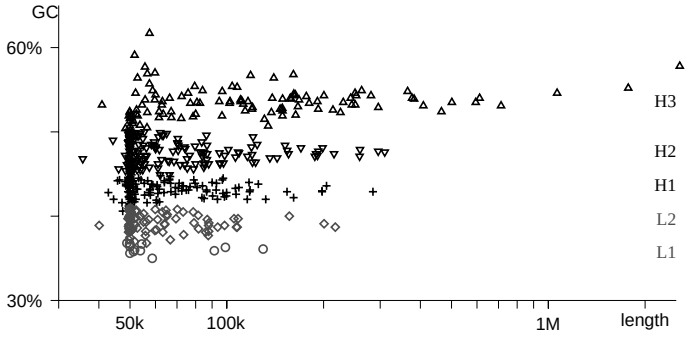

Fig. 2. Segment composition and length in the segmentation of chr19. Segment class levels are as follows: 35%, 39%, 43%, 47%, and 53% `GC` in L1–H3, respectively.

computations for HMMs when the state sequence has to obey some duration thresholds. Hidden Markov models are sometimes used along with some ad hoc thresholding on segment lengths (e.g., [22]). Our results show that such an approach can be implemented in a theoretically sound manner. There are some standard techniques [5], involving extra states or transitions, which can model minimum segment lengths at the price of increased time complexity. In contrast, our algorithms' running time is linear in the number of segment classes (using the equivalent of two states per class), and the time complexity is not affected by the minimum segment length.

Without doubt, many genome features (such as gene density, retrotransposition and replication timing) are linked to regional `GC` composition, but there is still need for an adequate "isochore theory" that explains genome organization in terms of isochores. A main difficulty in assessing the role of isochores in mammalian genome analysis has been the lack of a widely accepted generative (as opposed to descriptive) model. In our opinion, such a falsifiable model is necessary for a useful scientific discussion, and would open up the path to meaningful hypothesis testing procedures. Refutation attemps [10, 11, 14] have been rebuked on the basis that the employed statistical models do not adequately capture the true nature of isochores [9, 12]. On the other hand, proponents of the theory largely relied on ad hoc segmentation procedures [2, 13], which result in useful genome annotations, but make it difficult to assess statistical validity. We intend to continue our work toward an adequate isochore model, by incorporating positional dependence and other essential features.

We hope that our model and the associated computational results will be useful on their own for "simple" sequence analysis tasks, such as the identification of isochores or CpG islands, or as part of more sophisticated probabilistic models for complicated analysis problems, such as *ab initio* gene prediction.

References

1. Karlin, S.: Statistical signals in bioinformatics. Proc. Natl. Acad. Sci. USA **102** (2005) 13355–13362
2. Li, W., Bernaola-Galván, P., Haghighi, F., Grosse, I.: Applications of recursive segmentation to the analysis of DNA sequences. Comput. Chem. **26** (2002) 491–510
3. Mathé, C., Sagot, M.F., Schiex, T., Rouzé, P.: Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. **30** (2002) 4103–4117
4. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE **77** (1989) 257–286
5. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis. Cambridge University Press, UK (1998)
6. Fu, Y.X., Curnow, R.N.: Maximum likelihood estimation of multiple change points. Biometrika **77** (1990) 563–573
7. Csűrös, M.: Maximum-scoring segment sets. IEEE/ACM Trans. Comput. Biol. Bioinf. **1** (2004) 139–150
8. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F.: The mosaic genome of warmblooded vertebrates. Science **228** (1985) 953–958
9. Bernardi, G.: Misunderstandings about isochores: Part I. Gene **276** (2001) 3–13
10. IHGSC: Initial sequencing and analysis of the human genome. Nature **409** (2001) 860–921
11. Cohen, N., Dagan, T., Stone, L., Graur, D.: GC composition of the human genome: in search of isochores. Mol. Biol. Evol. **22** (2005) 1260–1272
12. Clay, O., Bernardi, G.: How not to look for isochores: A reply to Cohen et al. Mol. Biol. Evol. **22** (2005) 2315–2317
13. Constantini, M., Clay, O., Auletta, F., Bernardi, G.: An isochore map of the human genome. Genome Res. **16** (2006) 536–541
14. Eyre-Walker, A., Hurst, L.D.: The evolution of isochores. Nat. Rev. Genet. **2** (2001) 549–555
15. Szpankowski, W., Ren, W., Szpankowski, L.: An optimal DNA segmentation based on the MDL principle. Int. J. Bioinformatics Research and Applications **1** (2005) 3–17
16. Barry, D., Hartigan, J.A.: Product partition models for change point problems. Ann. Statist. **20** (1992) 260–279
17. Auger, I.E., Lawrence, C.E.: Algorithms for the optimal identification of segment neighborhoods. Bull. Math. Biol. **51** (1989) 39–54
18. Rissanen, J.: A universal prior for integers and estimation by minimum description length. Ann. Statist. **11** (1983) 416–431
19. Tarnas, C., Hughey, R.: Reduced space hidden markov model training. Bioinformatics **14** (1998) 401–406
20. Grimwood, J., et al.: The DNA sequence and biology of human chromosome 19. Nature **428** (2004) 529–535
21. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J.: The UCSC genome browser database. Nucleic Acids Res. **31** (2003) 51–54
22. Klein, R.J., Misulovin, Z., Eddy, S.R.: Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc. Natl. Acad. Sci. USA **99** (2002) 7542–7547