Biometrics 60, 573–588
September 2004

A Bayesian Approach to DNA Sequence Segmentation

Richard J. Boys

School of Mathematics and Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.
email: Richard.Boys@ncl.ac.uk

and

Daniel A. Henderson

Department of Statistics, The Open University, Milton Keynes, MK7 6AA, U.K.
email: D.A.Henderson@open.ac.uk

Summary. Many deoxyribonucleic acid (DNA) sequences display compositional heterogeneity in the form of segments of similar structure. This article describes a Bayesian method that identifies such segments by using a Markov chain governed by a hidden Markov model. Markov chain Monte Carlo (MCMC) techniques are employed to compute all posterior quantities of interest and, in particular, allow inferences to be made regarding the number of segment types and the order of Markov dependence in the DNA sequence. The method is applied to the segmentation of the bacteriophage *lambda* genome, a common benchmark sequence used for the comparison of statistical segmentation algorithms.

Key words: Bacteriophage *lambda*; Bioinformatics; Hidden Markov model; Model selection; Order of Markov dependence; Transdimensional MCMC.

1. Introduction

Vast amounts of DNA sequence data are currently available for analysis, primarily as a result of large-scale sequencing projects such as the Human Genome Project. Consequently, there is an increasing need to develop efficient computational and statistical tools to analyze this profusion of biological data. In this article, we focus on the fundamental problem of analyzing compositional variability in genome sequences, that is, the complete DNA sequence of an organism. Many genome sequences display heterogeneity in base composition in the form of *domains* or *segments* of similar structure; Li (2004) provides an extensive up-to-date bibliography.

Several statistical techniques have been developed in an attempt to identify these homogeneous DNA segments, many of which are reviewed in Braun and Müller (1998). Other recent work includes the Bayesian approach of Liu and Lawrence (1999) and the quasi-likelihood method of Braun, Braun, and Müller (2000), both of which use a multiple change-point framework with the change points delimiting the segments. An alternative approach, introduced by Churchill (1989), describes the DNA sequence structure by a hidden Markov model (HMM), which, in essence, is a mixture model with Markov dependent component indicators; see MacDonald and Zucchini (1997) for an introduction. HMMs have become a popular choice for the analysis of DNA sequences and have been used subsequently by Muri (1997), Boys, Henderson, and Wilkinson (2000), and Nicolas et al. (2002), among others. We follow these papers by using an HMM as our basic model. One reason for the popularity of HMMs is their flexibility in allowing noncontiguous parts of the sequence to be described by the same underlying structure. This can lead to more realistic biological interpretations when compared to the results of standard change-point analyses.

Typically, HMMs assume that the observed process—here the DNA sequence—evolves independently given the unobserved Markov chain which locates the position of the segments. In this article, however, we allow the observed process to evolve as a qth-order Markov chain, conditional on the hidden Markov chain. A choice of $q = 0$ corresponds to the usual independence assumption, whereas a choice of $q > 0$ allows us to account for the additional short-range structure that is often evident in DNA sequences; such a model has been used by Churchill (1992), Boys et al. (2000), and Nicolas et al. (2002).

In general, both the order of dependence q and the number of hidden states which define the segment structure r will be unknown and therefore it will be necessary to make inferences about them from the data. Throughout the article, we adopt a Bayesian approach to inference which allows us to take full account of the uncertainty in the locations and the composition of the various segments. It also permits the incorporation of prior knowledge about these unknowns and provides a coherent framework for model comparison/selection. The complex structure of the model precludes a fully analytic treatment and we therefore use transdimensional MCMC (Green, 2003) to explore both the parameter and model space.

The remainder of the article is organized as follows. The model is presented in Section 2 together with details of the prior-to-posterior analysis. Section 3 contains a description

of the MCMC algorithm and is followed by an analysis of the bacteriophage *lambda* genome in Section 4. The article concludes with some brief remarks in Section 5.

2. The Bayesian Model

DNA sequence data can be represented by a string of letters $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ from the four-letter alphabet $\mathcal{Y} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The letters represent the four nucleic acids, or *bases*, adenine, cytosine, guanine, and thymine, respectively. However, for reasons of generality and simplicity of notation, we denote the state space by $\mathcal{Y} = \{1, 2, \ldots, b\}$, and hence we code $\mathtt{A}, \mathtt{C}, \mathtt{G}$, and $\mathtt{T}$ as 1, 2, 3, and 4 when referring to the DNA sequence.

The observed sequence $\boldsymbol{y}$ is assumed to be a realization of a hidden Markov model with *observation equations*

$$\begin{aligned}\Pr(Y_t \,|\, Y_{1:t-1}, S_{1:t}) &= \Pr(Y_t = j \,|\, Y_{t-q} = y_{t-q}, \ldots, Y_{t-1} = y_{t-1}, S_t = k) \\ &= p_{ij}^{(k)}, \qquad i \in \mathcal{Y}_q = \{1, 2, \ldots, b^q\}, \quad j \in \mathcal{Y}, \\ &\qquad\qquad k \in \mathcal{S}_r = \{1, 2, \ldots, r\},\end{aligned}$$

where $i = \mathcal{I}(\boldsymbol{y}, t, q, b) \equiv 1 + \sum_{\ell=1}^{q} (y_{t-\ell} - 1) b^{\ell-1}$, and *state equations*

$$\Pr(S_t \,|\, S_{1:t-1}) = \Pr(S_t = j \,|\, S_{t-1} = i) = \lambda_{ij}, \qquad i, j \in \mathcal{S}_r,$$

for $t = q_{\max} + 1, q_{\max} + 2, \ldots, n$. We take this range of values for t to remove the need to specify marginal models that describe the evolution at the beginning of the sequence. Also, throughout this article we use the notation $x_{i:j}$ to denote the sequence $x_i, x_{i+1}, \ldots, x_j$.

The above specification indicates that the observed process $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ evolves as a qth-order Markov chain, conditional on the hidden process $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$. Also, the unobserved process $\boldsymbol{S}$ follows a first-order homogeneous r-state Markov chain, with transition matrix $\Lambda = (\lambda_{ij})$, where each row $\boldsymbol{\lambda}_i \in \mathscr{S}_r$, the r-dimensional simplex. For the sake of notational conciseness, we work with the *reduced* form of the qth-order transition matrix which, for each hidden state or "segment type" k, is a $b^q \times b$ matrix $P^{(k)}$ consisting of elements $p_{ij}^{(k)}$. Each reshaped matrix $P^{(k)}$ has rows $\boldsymbol{p}_i^{(k)} \in \mathscr{S}_b$ and we denote the collection of these transition matrices by $\mathcal{P} = \{P^{(1)}, P^{(2)}, \ldots, P^{(r)}\}$. Also, we take $r \in \mathcal{R} = \{1, 2, \ldots, r_{\max}\}$ and $q \in \mathcal{Q} = \{0, 1, \ldots, q_{\max}\}$. Finally, we denote the set of unknown hidden state and base transition matrices by $\boldsymbol{\theta} = \{\Lambda, \mathcal{P}\} \in \mathscr{S}_r^r \times \mathscr{S}_b^{rb^q}$, where the space $\mathscr{S}_r^x$ denotes the product of x simplices, each one r-dimensional.

2.1 *Prior Distributions*

The aim of the analysis is to make inferences about the unknown quantities in the model: the order of dependence q, the number of segment types r, the model transition parameters $\boldsymbol{\theta}$, and the sequence of segment types or "segmentation" $\boldsymbol{s}$. We begin by quantifying our uncertainty about these unknowns (before observing the data) through a prior distribution which takes the form

$$\pi(r, q, \boldsymbol{\theta}) = \pi(r)\pi(q)\pi(\boldsymbol{\theta} \,|\, r, q) = \pi(r)\pi(q)\pi(\Lambda \,|\, r)\pi(\mathcal{P} \,|\, r, q)$$

and assumes, inter alia, that r and q are independent a priori. Specifically, we adopt independent truncated Poisson prior distributions for r and q, namely

$$r \sim Po(a_r; \mathcal{R}) \qquad \text{and} \qquad q \sim Po(a_q; \mathcal{Q}), \tag{1}$$

where $a_r > 0$ and $a_q > 0$ are fixed hyperparameters, and $X \sim Po(a; \mathcal{X})$ denotes a random variable with probability function $\Pr(X = i) \propto a^i / i!, i \in \mathcal{X}$.

The components of $\boldsymbol{\theta}$ are all defined on simplices and there are therefore many choices of prior distribution which could be made; see Boys and Henderson (2002). We follow their recommendation and take independent Dirichlet distributions for the rows of each $P^{(k)}$ and Λ, that is

$$\boldsymbol{p}_i^{(k)} = \left(p_{ij}^{(k)}\right) \,\big|\, r, q \sim \mathscr{D}\left(\boldsymbol{c}_i^{(k)}\right), \quad i \in \mathcal{Y}_q, \quad j \in \mathcal{Y}, \quad k \in \mathcal{S}_r, \tag{2}$$

$$\boldsymbol{\lambda}_i = (\lambda_{ij}) \,|\, r \sim \mathscr{D}(\boldsymbol{d}_i), \quad i, j \in \mathcal{S}_r, \tag{3}$$

where the Dirichlet parameters $\boldsymbol{c}$ and $\boldsymbol{d}$ are chosen to reflect the goal of the analysis; see Section 4.1.

2.2 *Likelihood*

Information from the data on the unknown quantities is expressed through the likelihood function. For HMMs, it is computationally convenient to treat the hidden states $\boldsymbol{s}$ as "missing" data and work with the *complete-data* likelihood. For our model, this is

$$\begin{aligned}\pi(\boldsymbol{y}, \boldsymbol{s} \,|\, r, q, \boldsymbol{\theta}) &= \pi(\boldsymbol{y} \,|\, \boldsymbol{s}, r, q, \boldsymbol{\theta})\pi(\boldsymbol{s} \,|\, r, q, \boldsymbol{\theta}) \\ &\propto \prod_{i \in \mathcal{Y}_q} \prod_{j \in \mathcal{Y}} \prod_{k \in \mathcal{S}_r} \left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}} \prod_{i \in \mathcal{S}_r} \prod_{j \in \mathcal{S}_r} \lambda_{ij}^{m_{ij}}, \end{aligned} \tag{4}$$

where

$$n_{ij}^{(k)} = \sum_{t=q_{\max}+1}^{n} \mathbb{I}\left(\mathcal{I}(\boldsymbol{y}, t, q, b) = i, y_t = j, s_t = k\right) \quad \text{and}$$

$$m_{ij} = \sum_{t=q_{\max}+1}^{n} \mathbb{I}(s_{t-1} = i, s_t = j)$$

denote the relevant transition counts and $\mathbb{I}(x)$ is the usual indicator function which equals 1 if x is true and 0 otherwise. Implicit in this formulation is an assumption that the DNA sequence is *linear* rather than *circular*; modifications required for circular sequences are given in Churchill (1989).

2.3 *Posterior Inference*

In the Bayesian paradigm, inferences are based on the posterior distribution

$$\pi(r, q, \boldsymbol{\theta}, \boldsymbol{s} \,|\, \boldsymbol{y}) \propto \pi(r, q, \boldsymbol{\theta})\pi(\boldsymbol{y}, \boldsymbol{s} \,|\, r, q, \boldsymbol{\theta}),$$

which calibrates our uncertainties about the unknown parameters after observing the data. This distribution does not permit a straightforward analysis and therefore we utilize MCMC methods to make inferences about the unknown quantities of interest.

In the case of fixed r and q, as described in Boys et al. (2000), the MCMC scheme is straightforward and proceeds by standard Gibbs sampling techniques for HMMs (Robert, Celeux, and Diebolt, 1993; Chib, 1996). Briefly, values for the

transition probability parameters $\boldsymbol{\theta}$ are generated from their full conditional distributions which are independent (conjugate) Dirichlet distributions:

$$\boldsymbol{p}_i^{(k)} \,\big|\, r, q, \boldsymbol{s}, \boldsymbol{y} \sim \mathscr{D}\big(\boldsymbol{c}_i^{(k)} + \boldsymbol{n}_i^{(k)}\big), \quad i \in \mathcal{Y}_q, \quad k \in \mathcal{S}_r, \quad (5)$$

$$\boldsymbol{\lambda}_i \,|\, r, q, \boldsymbol{s}, \boldsymbol{y} \sim \mathscr{D}(\boldsymbol{d}_i + \boldsymbol{m}_i), \quad i \in \mathcal{S}_r, \quad (6)$$

where $\boldsymbol{n}_i^{(k)} = (n_{ij}^{(k)})$ and $\boldsymbol{m}_i = (m_{ij})$ are the transition counts. Similarly, a new sequence of hidden states $\boldsymbol{s}$ is generated from its full conditional distribution $\pi(\boldsymbol{s} \,|\, r,\, q,\, \boldsymbol{\theta},\, \boldsymbol{y})$ using a standard forward–backward simulation algorithm. Details of this algorithm can be found in Boys and Henderson (2002).

The novel aspect of our model is that both r and q are unknown parameters. The MCMC algorithm allows for this by requiring the sampler to jump between parameter spaces with different dimensions corresponding to models with different values of r and q. For r, this can be accomplished by using reversible jumps (Green, 1995), an adaptation of the Metropolis–Hastings algorithm which permits transdimensional movement. In particular, the method we use is based on standard reversible jump techniques for HMMs (Robert, Rydén, and Titterington, 2000). However, as we show in Section 3.1, the structure of the model allows q to be updated by a more direct method. We now describe the MCMC scheme in more detail.

3. Outline of MCMC Scheme

At each iteration of the MCMC algorithm, a fixed scan is performed of the following *moves*:

(a) update the order of dependence and the transition probability parameters using $\pi(q,\, \boldsymbol{\theta} \,|\, r,\, \boldsymbol{s},\, \boldsymbol{y})$;
(b) update the number of segment types r (and consequently update $\boldsymbol{\theta}$ and $\boldsymbol{s}$) conditional on q;
(c) update the segmentation $\boldsymbol{s}$ using $\pi(\boldsymbol{s} \,|\, r,\, q,\, \boldsymbol{\theta},\, \boldsymbol{y})$.

Repeating this scheme for N iterations beyond convergence will give values $(r^{(i)},\, q^{(i)},\, \boldsymbol{\theta}^{(i)},\, \boldsymbol{s}^{(i)})$ for $i = 1,\, 2, \ldots, N$ on which to base posterior summaries.

Moves (a) and (b) require the sampler to jump between models in different dimensional parameter spaces and will be described further in the following sections. Move (c) is accomplished using the standard forward–backward algorithm; see Section 2.3.

3.1 *Move (a)*

Move (a) is a joint move in which the order of Markov dependence q is updated using $\pi(q \,|\, r,\, \boldsymbol{s},\, \boldsymbol{y})$ and then the transition probability parameters $\boldsymbol{\theta}$ are updated using their full conditional distribution $\pi(\boldsymbol{\theta} \,|\, r,\, q,\, \boldsymbol{s},\, \boldsymbol{y})$, which is given in (5) and (6). The choice of conjugate (Dirichlet) prior distribution for $\mathcal{P}$ allows q to be updated directly from the appropriate conditional distribution without resorting to reversible jump moves, that is, from

$$\pi(q \,|\, r, \boldsymbol{s}, \boldsymbol{y}) \propto \pi(q \,|\, r, \boldsymbol{s})\pi(\boldsymbol{y} \,|\, r, q, \boldsymbol{s}) = \pi(q)\pi(\boldsymbol{y} \,|\, r, q, \boldsymbol{s}), \quad (7)$$

where the simplification results from q being independent of $(r,\, \boldsymbol{s})$ a priori. The *marginal* likelihood $\pi(\boldsymbol{y} \,|\, r,\, q,\, \boldsymbol{s})$ is calculated by using a simple rearrangement of Bayes's Theorem as

$$\begin{aligned}\pi(\boldsymbol{y} \,|\, r, q, \boldsymbol{s}) &= \frac{\pi(\mathcal{P} \,|\, r, q, \boldsymbol{s})\pi(\boldsymbol{y} \,|\, \mathcal{P}, r, q, \boldsymbol{s})}{\pi(\mathcal{P} \,|\, r, q, \boldsymbol{s}, \boldsymbol{y})} \\ &= \prod_{k=1}^{r}\prod_{i=1}^{b^q}\left[\frac{\Gamma\left(\sum_{j=1}^{b} c_{ij}^{(k)}\right)\prod_{j=1}^{b}\Gamma\big(c_{ij}^{(k)} + n_{ij}^{(k)}\big)}{\prod_{j=1}^{b}\Gamma\big(c_{ij}^{(k)}\big)\Gamma\left\{\sum_{j=1}^{b}\big(c_{ij}^{(k)} + n_{ij}^{(k)}\big)\right\}}\right]. \quad (8)\end{aligned}$$

Boys and Henderson (2002) discuss other choices of prior distribution for $\mathcal{P}$ which inherit the simplicity of this marginal likelihood calculation.

3.2 *Move (b)*

In move (b), the number of segment types r is updated using two types of birth/death reversible jump moves, implemented successively. The birth/death moves are conceptually and computationally simpler to implement than split/merge moves and we have found them to result in adequate mixing over r.

3.2.1 *Birth and death moves.* These moves are similar in style to the birth and death moves described by Viallefont, Richardson, and Green (2002). The move begins with a random choice between creating and deleting a segment type with probabilities b_r and d_r, respectively.

In the *birth* move, a new segment type $j^\star$ is proposed, thus taking the number of segment types from r to $r+1$. A set of base transitions $\boldsymbol{u}$ for the new segment type is generated from the prior distribution (2), and then we set $\tilde{P}^{(j^\star)} = \boldsymbol{u}$ and $\tilde{P}^{(j)} = P^{(j)}$ for $j \neq j^\star$. We then simulate a row vector $\boldsymbol{v}$ from the prior distribution (2) and set row $j^\star$ of the proposed transition matrix $\tilde{\Lambda}$ to be $\tilde{\boldsymbol{\lambda}}_{j^\star} = \boldsymbol{v}$. Column $j^\star$ is then filled by taking $\tilde{\lambda}_{ij^\star} = w_i$ for $i \neq j^\star$, where $w_i \sim \text{Beta}(\tilde{d}_{ij^\star}, \sum_{j \neq j^\star} \tilde{d}_{ij})$ has the appropriate marginal distribution. The elements of $\tilde{\Lambda}$ are then scaled to ensure it is stochastic. Finally, a new segmentation $\tilde{\boldsymbol{s}}$ is simulated conditional on $r+1$ and $\tilde{\boldsymbol{\theta}}$ using the forward–backward algorithm. The move is accepted with probability $\min(1, A_{\text{B}})$, where

$$\begin{aligned}A_{\text{B}} = {} & \frac{\pi(\boldsymbol{y}, \tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}})}{\pi(\boldsymbol{y}, \boldsymbol{s} \,|\, r, q, \boldsymbol{\theta})} \times \frac{\pi(r+1)}{\pi(r)} \times (r+1) \\ & \times \frac{\prod_{i \in \mathcal{S}_{r+1}} \mathcal{D}(\tilde{\boldsymbol{\lambda}}_i \,|\, \tilde{\boldsymbol{d}}_i)}{\prod_{i \in \mathcal{S}_r} \mathcal{D}(\boldsymbol{\lambda}_i \,|\, \boldsymbol{d}_i)} \frac{\prod_{k \in \mathcal{S}_{r+1}}\prod_{i \in \mathcal{Y}_q} \mathcal{D}\big(\tilde{\boldsymbol{p}}_i^{(k)} \,\big|\, \tilde{\boldsymbol{c}}_i^{(k)}\big)}{\prod_{k \in \mathcal{S}_r}\prod_{i \in \mathcal{Y}_q} \mathcal{D}\big(\boldsymbol{p}_i^{(k)} \,\big|\, \boldsymbol{c}_i^{(k)}\big)} \\ & \times \frac{d_{r+1}}{b_r(r+1)} \frac{\pi(\boldsymbol{s} \,|\, r, q, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}}, \boldsymbol{y})} \\ & \times \left\{\mathcal{D}(\tilde{\boldsymbol{v}} \,|\, \tilde{\boldsymbol{d}}_{j^\star}) \prod_{i \in \mathcal{S}_{r+1} \setminus j^\star} \mathcal{B}\left(w_i \,\middle|\, \tilde{d}_{ij^\star}, \sum_{j \in \mathcal{S}_{r+1} \setminus j^\star} \tilde{d}_{ij}\right) \right. \\ & \qquad \left. \times \prod_{i \in \mathcal{Y}_q} \mathcal{D}\big(\boldsymbol{u}_i \,\big|\, \tilde{\boldsymbol{c}}_i^{(j^\star)}\big)\right\}^{-1} \prod_{i \in \mathcal{S}_{r+1} \setminus j^\star} (1 - w_i)^{r-1},\end{aligned}$$

where $\mathcal{D}(\cdot \,|\, \boldsymbol{\theta})$ is the $\mathscr{D}(\boldsymbol{\theta})$ density and $\mathcal{B}(\cdot \,|\, a, b)$ is the Beta(a, b) density. Briefly, the first two lines in this expression consist of the likelihood ratio and the prior ratio, the remaining lines consist of the proposal ratio and the Jacobian resulting from the change of variables $(\mathcal{P}, \boldsymbol{u}) \to \tilde{\mathcal{P}}$ and $(\Lambda, \boldsymbol{v}, \boldsymbol{w}) \to \tilde{\Lambda}$. Details of the general form of such acceptance ratios can be found in Cappé, Robert, and Rydén (2003).

Although the expression for A_{B} can be simplified, we have included all terms so that generalizations to other prior and proposal distributions are clear. We note, however, that A_{B} does not depend on $\boldsymbol{s}$ or $\tilde{\boldsymbol{s}}$ because

$$\frac{\pi(\boldsymbol{y}, \tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}})}{\pi(\boldsymbol{y}, \boldsymbol{s} \,|\, r, q, \boldsymbol{\theta})} \times \frac{\pi(\boldsymbol{s} \,|\, r, q, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}}, \boldsymbol{y})} = \frac{\pi(\boldsymbol{y} \,|\, r+1, q, \tilde{\boldsymbol{\theta}})}{\pi(\boldsymbol{y} \,|\, r, q, \boldsymbol{\theta})},$$

and the *observed-data* likelihoods $\pi(\boldsymbol{y} \,|\, r+1, q, \tilde{\boldsymbol{\theta}})$ and $\pi(\boldsymbol{y} \,|\, r,\ q,\ \boldsymbol{\theta})$ can be computed independently of the segmentation from a forward sweep of the forward–backward algorithm; see MacDonald and Zucchini (1997). Therefore, the move can be simplified slightly by leaving the simulation of a new segmentation $\tilde{\boldsymbol{s}}$ to be performed in move (c).

The *death* move is the obvious reverse of the birth move: A randomly chosen current segment type $j^\star$ is proposed to be deleted and the other parameters are adjusted accordingly. First, $\tilde{P}^{(j^\star)}$ is deleted and the remaining base transition probabilities are taken as $P^{(j)} = \tilde{P}^{(j)}$ for $j \neq j^\star$. Then row and column $j^\star$ of $\tilde{\Lambda}$ are deleted before its remaining elements are rescaled to obtain the stochastic matrix Λ. The death of segment type $j^\star$ is accepted with probability $\min(1, A_{\mathrm{B}}^{-1})$ as the birth and death moves form a reversible pair. Again, there is no need to deduce a new segmentation $\boldsymbol{s}$ as part of this death move.

3.2.2 *Birth and death of empty segment types.* To improve the mixing behavior of the reversible jump algorithm we also use birth and death moves that act solely on dormant or *empty* segment types—that is, segment types with no data currently allocated to them—in the spirit of Richardson and Green (1997). These special birth/death moves operate in largely the same way as those described in Section 3.2.1, and hence the details have been omitted. The acceptance probability for the birth of empty segment type $j^\star$ is $\min(1, A_{\mathrm{E}})$, where

$$A_{\mathrm{E}} = A_{\mathrm{B}} \times \frac{(r+1)}{(r_0+1)} \times \frac{\pi(\tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}})}{\pi(\boldsymbol{s} \,|\, r, q, \boldsymbol{\theta})} \times \frac{\pi(\boldsymbol{y}, \boldsymbol{s} \,|\, r, q, \boldsymbol{\theta})}{\pi(\boldsymbol{y}, \tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}})} \times \frac{\pi(\tilde{\boldsymbol{s}} \,|\, r+1, q, \tilde{\boldsymbol{\theta}}, \boldsymbol{y})}{\pi(\boldsymbol{s} \,|\, r, q, \boldsymbol{\theta}, \boldsymbol{y})},$$

and r_0 denotes the number of empty segment types prior to the proposed birth. The corresponding death move is accepted with probability $\min(1, A_{\mathrm{E}}^{-1})$.

4. Application to DNA Sequence Data

We illustrate the general method described in the previous sections by analyzing the genome of the bacteriophage *lambda*, a parasite of the intestinal bacterium *Escherichia coli*. This virus has become a benchmark sequence for the comparison of segmentation algorithms since the experimental segmentation based on gradient centrifugation of its CG content by Skalka, Burgi, and Hershey (1968); see Braun and Müller (1998) for further references. Its (circular) genome is relatively small at 48,502 base pairs (bp) in length, though this is long enough to provide an adequate challenge for the methods in this article. The complete genome sequence is stored in the GenBank sequence database (Benson et al., 2004) under accession number J02459 and can be obtained from the National Center for Biotechnology Information (NCBI) web pages at `http://www.ncbi.nlm.nih.gov/`.

4.1 *Prior Specification*

Our aim is to describe the structure of the bacteriophage *lambda* as parsimoniously as possible. This preference can be expressed through the prior distribution. In choosing a prior (1) for r and q, we want to express a preference for a small number of segment types and a low order of Markov dependence without being too restrictive. Therefore, taking into account the number of transition parameters and the length of the sequence, we choose upper bounds for r and q of $r_{\max} = 14$ and $q_{\max} = 3$ and prior means around $a_r = 3$ and $a_q = 1$, respectively.

Prior knowledge about the base transition probabilities in each segment is necessarily weak, and so we make the exchangeable choice $\boldsymbol{c}_i^{(k)} = (1,\ 1,\ 1,\ 1)$ for $i \in \mathcal{Y}_q$ and $k \in \mathcal{S}_r$. The specification of parameters for the segment transition structure Λ is more complex. However, our prior preference for a relatively small number of large homogeneous regions can be expressed through uncertainty about segment lengths. This can be achieved by adopting an off-diagonal exchangeable pattern of the form $(\boldsymbol{d}_i)_j = \alpha\delta_{ij} + \beta(1 - \delta_{ij})$ for some choice of α and β, where δ_{ij} is Kronecker's delta. These hyperparameters are chosen by considering prior mean segment lengths and effective prior "sample sizes." For this analysis, we have chosen a priori expected segment lengths of 1000 bases with each row $\boldsymbol{d}_i$ having the information content of a sequence with $1000 r_{\max}/r$ transitions. This prior input is not particularly strong given that this DNA sequence is nearly 50,000 bases long. It also balances the amount of prior information, in terms of equivalent transitions, for the different values of r. The sensitivity of the results to changes in the prior are discussed in Section 4.3.

4.2 *Results*

The MCMC algorithm was run from a variety of starting points and its convergence was monitored using a range of conventional convergence diagnostics. Each run produced essentially the same results and we report here one typical run consisting of a burn-in of 500,000 iterations followed by a further 100,000 iterations in which only every 10th iterate was recorded in order to reduce computing overheads. Thus posterior inferences are based on $N = 10{,}000$ sampled values $(r^{(i)},\ q^{(i)},\ \boldsymbol{\theta}^{(i)},\ \boldsymbol{s}^{(i)})$.

4.2.1 *Marginal posterior for q.* The (marginal) posterior distribution for the order of dependence parameter q, as estimated from the MCMC output, is concentrated (almost) solely at $q = 2$, that is, within each segment type, the bases almost certainly follow a second-order Markov chain. This may seem overly precise but Fan and Tsai (1999) and Boys and Henderson (2002) have also observed that highly

Table 1
Posterior distribution of $r \mid q = 2$

r	≤ 4	5	6	7	8	9	10	11	12	13	14
Probability	0.000	0.011	0.542	0.297	0.111	0.030	0.006	0.002	0.001	0.000	0.000

concentrated posterior distributions for q are often obtained when analyzing reasonably long sequences containing either a single segment or two segment types. This finding is also supported by an examination of the q-distribution in each of the segment types identified by the posterior mode estimate of the segmentation; see Section 4.2.4.

Posterior support for second-order dependence is not particularly surprising since the bacteriophage *lambda* genome is composed predominantly of coding DNA and is therefore largely governed by the genetic code, which is read in triples (y_{t-2}, y_{t-1}, y_t). Moreover, this conclusion reinforces the need to be able to make inferences about q when analyzing DNA sequences, as conventional segmentation analyses assume $q = 0$. In the following subsections, we condition our analysis on second-order dependence.

4.2.2 *Posterior for $r \mid q = 2$.* Table 1 contains an estimate of the posterior distribution for r given $q = 2$ based on empirical averaging of the sampler output.

It shows that there still remains some (posterior) uncertainty about r. The model which receives most support is one with $r = 6$ segment types, and no support is given to models with fewer than 5 or greater than 13 segment types. The 95% highest density interval (HDI) is $\{6, 7, 8\}$. It is interesting to note that other HMM-based analyses of the bacteriophage *lambda* genome, such as in Churchill (1989, 1992) and Muri (1998), have concentrated on models with $r \leq 6$ and $q \leq 1$ which receive little or no posterior support in our analysis.

The MCMC sampler mixes adequately over different values of r. This may be due to the similarity in the prior and posterior distributions for Λ, which results in birth proposals that have (relatively) high posterior density and thus a reasonable probability of acceptance. Overall, approximately 3.5% of birth/death moves and 12.1% of birth/death of empty segment types moves are accepted. These rates are quite low but they do compare favorably with those reported in the HMM analysis of Robert, Rydén, and Titterington (2000).

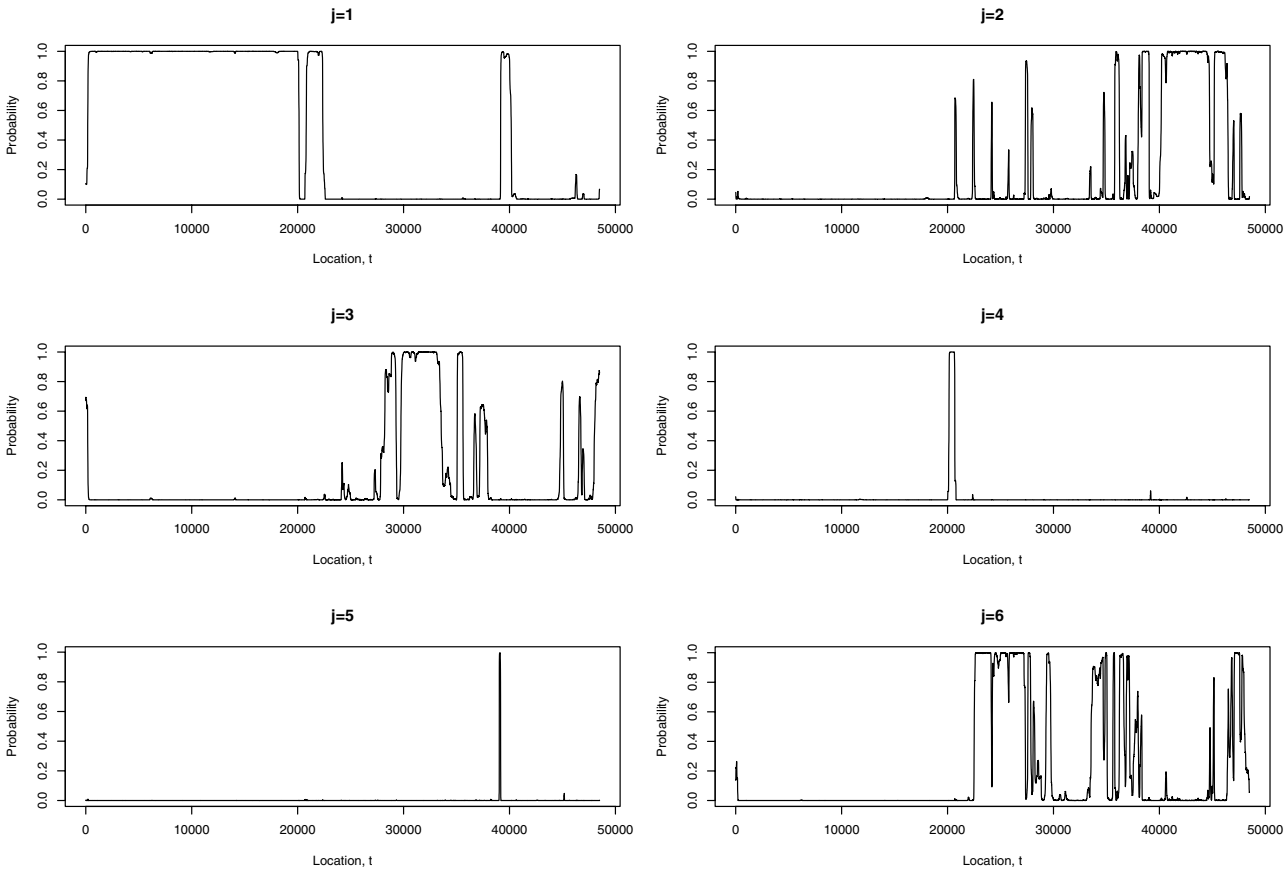

Figure 1. Posterior segment type probabilities $\widehat{\Pr}(S_t = j \mid r = 6, q = 2, \boldsymbol{y})$ for $j = 1, 2, \ldots, 6$.

4.2.3 *Problems of identifiability.* Although there are many possible summaries for this high- (and variable-) dimensional posterior distribution, we concentrate on summaries for the segmentation $\boldsymbol{s}$ and the parameters $\boldsymbol{\theta}$ conditional on r and q. This task is complicated by the nonidentifiability of the parameters in the posterior distribution, which leads to *label switching* in the MCMC sampler output; see Stephens (2000) for a detailed description of this feature. To remedy this problem we focus on one of the $r!$ symmetric posterior modes by performing a relabeling algorithm, along the lines of Stephens (2000). The algorithm permutes the MCMC output according to the permutation of the hidden states $\boldsymbol{s}$ that is most consistent with the marginal posterior mode (MPM) estimate $\hat{\boldsymbol{s}}$—the most likely hidden state at each position in the sequence. The reader is referred to Boys and Henderson (2002) for a detailed description of the algorithm.

4.2.4 *Posterior for $\boldsymbol{s} \,|\, r,\, q$.* The primary focus of all segmentation algorithms is to infer the latent segment structure of a sequence and this is done here through the posterior distribution of $\boldsymbol{s}$ given r and q. One summary of this posterior distribution is the MPM estimate $\hat{\boldsymbol{s}}$. However, it is useful to have some understanding of posterior uncertainty regarding the segmentation in addition to the MPM estimate. This can be provided by using an estimate of the segment type distribution at each location $t = q_{\max} + 1,\, q_{\max} + 2, \ldots, n$, such as, for $j \in \mathcal{S}_r$

$$\widehat{\Pr}(S_t = j \,|\, r = r^\star, q = q^\star, \boldsymbol{y}) = \sum_{i=1}^{N} \mathbb{I}\big(s_t^{(i)} = j, r^{(i)} = r^\star, q^{(i)} = q^\star\big) \Big/ \sum_{i=1}^{N} \mathbb{I}\big(r^{(i)} = r^\star, q^{(i)} = q^\star\big)$$

or its Rao–Blackwellized equivalent (Gelfand and Smith, 1990). Figure 1 displays these probabilities for the bacteriophage *lambda* genome conditional on the most likely model a posteriori, that is, six segment types and second-order dependence. It clearly shows the existence of several well-defined segments. In particular, the first half of the sequence appears to be rather homogeneous, consisting mainly of type 1 structure but also with a short fragment of different structure (type 4) shortly after base 20,000. In contrast, the second half of the sequence is comparatively heterogeneous and contains structure of types 2, 3, 5, and 6. We note that, although the segments of types 4 and 5 are quite short, they are nevertheless well defined. It is also interesting to note that, although we have not analyzed the data as a circular sequence, the segmentation has identified the same segment type (type 3) at the very beginning and end of the sequence.

4.2.5 *Posterior for $\boldsymbol{\theta} \,|\, r,\, q$.* We now focus on the transition structure of the observed sequence and concentrate again on the model with $r = 6$ and $q = 2$. Figure 2 illustrates the

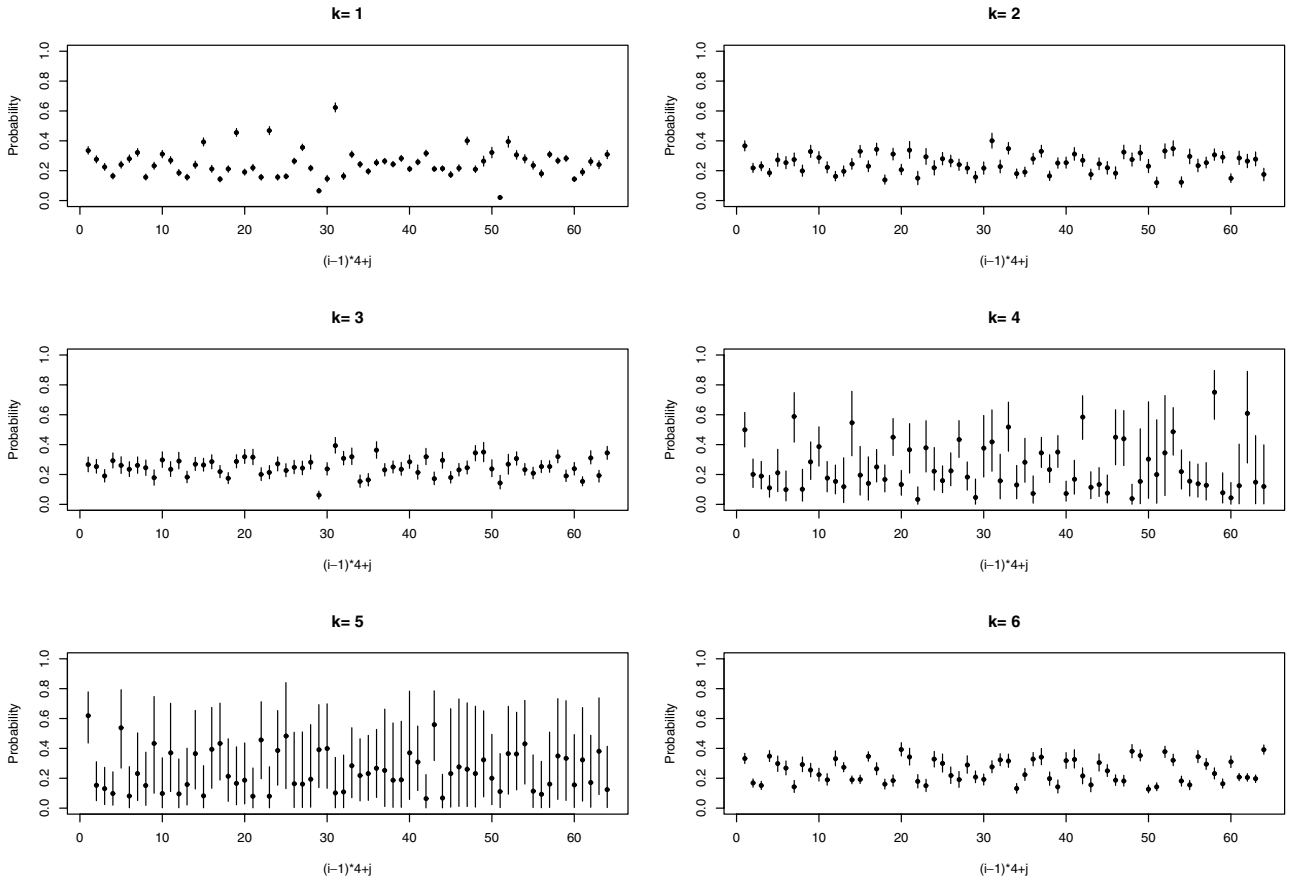

Figure 2. Marginal posterior means for the base transition probabilities $p_{ij}^{(k)}$ with 95% credible intervals.

base transition structure within each segment. It displays the marginal posterior distributions for the probabilities $p_{ij}^{(k)}$ through their estimated posterior means together with 95% equi-tailed credible intervals. The figure clearly shows that there is considerable (posterior) uncertainty about the base transition structure in segment types 4 and 5, in contrast to the other segments, especially segment type 1. This level of uncertainty is mainly due to the preponderance of each segment type in the sequence (see Figure 1). The plot also provides a useful characterization of the transition structure and shows that the patterns of transition structure differ markedly between segment types. This explains to some extent why there is so little (posterior) evidence for a lower number of segment types ($r \leq 5$), and why the locations of segment types 4 and 5 are well identified in Figure 1 despite their comparatively short length.

For reasons of brevity, we have not included summaries of the posterior distribution of the hidden state transition matrix Λ. However, information on useful quantities such as segment lengths can be gleaned from the posterior distribution of $\boldsymbol{s}$.

4.3 *Sensitivity to Prior Specification*

The sensitivity of our conclusions to the choice of prior parameters was investigated using many runs of the MCMC algorithm. The posterior distribution is fairly insensitive to changes in the prior distributions for the dimension parameters (r, q) and the base transition probabilities ($\mathcal{P}$). However, this is not the case for the hidden state transition matrix Λ. For example, the MPM segmentation is quite sensitive to the prior mean (and variance) of the segment lengths. A reanalysis of this sequence taking prior mean segment lengths to be as low as 100 bases naturally results in a segmentation with considerable switching between a much larger number of segment types.

4.4 *Biological Relevance of the Results*

We now compare our segmentation results to known biological functions of the bacteriophage *lambda* genome. Lewin (2000) provides an overview of *lambda* genome organization together with further details on the content of, and terminology used in, this section.

Focusing on the model that is most consistent with the data, that is $r = 6$ and $q = 2$, reveals segments (see Figure 1) which contain coding regions with the same direction of transcription: segments of types 1 and 2 transcribe left to right, and the others, right to left. The segmentation also identifies regions of biological significance and classes of genes with similar function. Segment type 1 comprises the structural genes for the bacteriophage particle (the "head" and "tail"). These genes are expressed when the phage is undergoing replication. It also contains the *P* protein, which is required for replication of the DNA. The type 2 segment contains the *Nin* proteins (which deal with DNA recombination), proteins *cro*, *cII*, and *Q* (which have important roles in regulation of gene expression), and proteins *R* and *Rz* (which are involved in breaking open the cells to release the new bacteriophage particles). Segment type 3 contains many coding regions; some are transcribed early in the life cycle, others have, as yet, unknown function. It also contains important regulatory proteins (*N* and *cI*) together with *int* and *xis* which are involved in integrating the phage DNA into the (host) bacterial chromosome (when it becomes a lysogen) and cutting it out again when it reactivates. The type 3 structure at the 5′ end of the sequence is clearly identified as being separate from the first segment (type 1) which contains the structural genes. Segment type 4 identifies *orf206b*, a coding region with a distinctive transition pattern. Segment type 5 contains the *O* gene, which is involved in initiating replication. Finally, segment type 6 contains the endonuclease *ea59* together with *rexA* and *rexB*, which are involved in excluding other bacteriophages that might compete with lambda.

5. Discussion

This article has described a fully Bayesian analysis of DNA sequence data assuming a hidden Markov model in which the observed process evolves as a Markov chain. The approach

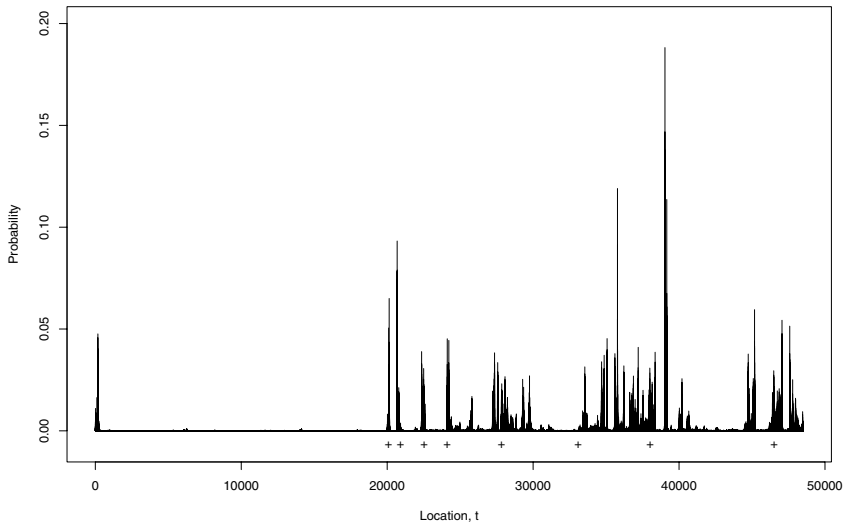

Figure 3. Posterior probability of a change point at each position in the sequence; + indicates change points identified by Braun et al. (2000).

is quite general in terms of its flexibility with respect to the number of different segment types that may be present in the sequence and the complexity of the structure within each segment type.

There are several other published HMM-based analyses of the bacteriophage *lambda* genome. However, these analyses consider a much more restricted class of models (in terms of r and q) than considered here. A more recent analysis by Braun et al. (2000) assumes an independence ($q = 0$) model for base transitions and uses quasi-likelihood to identify changes in the segmentation together with an adjusted BIC to determine the number of change points. They conclude that there are eight change points and their locations are identified on the changepoint (posterior) probability plot in Figure 3. Here the changepoint probabilities are calculated using the output from our MCMC scheme and are averaged over different values of r (and q); details of other posterior summaries that take into account the variability in r and q and that are label invariant are available from the authors.

Our analysis is in general agreement with these locations except in the latter part of the sequence. In particular, their $q = 0$ analysis fails to identify the change points corresponding to the location of the type 5 segment around base 39,000. Nevertheless, our analysis of the bacteriophage *lambda* genome reveals much of the biological structure known to be present in this sequence.

Acknowledgements

This work was carried out during DAH's tenure of a Lord Adams Fellowship at Newcastle University. We are grateful to Dr Ethan Hack (School of Biology, Newcastle University) for his comments on the biological application.

Résumé

Beaucoup de séquences d'acides désoxyribonucléiques (ADN) présentent une hétérogénéité dans la forme des segments de structure similaire. Ce papier décrit une méthode bayésienne qui identifie de tels segments en utilisant une chaîne de Markov gouvernée par un modèle de Markov caché. Les techniques de Monte Carlo par chaînes de Markov (MCMC) sont utilisées pour calculer les quantités d'intérêt a posteriori et en particulier permettre des inférences eu égard au nombre de types de segments et à l'ordre de la dépendance markovienne dans la séquence ADN. La méthode est appliquée à la segmentation du génome du bactériophage lambda, une séquence courante de référence, utilisée pour la comparaison d'algorithmes de segmentation.

References

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: Update. *Nucleic Acids Research* **32,** D23–D26.

Boys, R. J. and Henderson, D. A. (2002). On determining the order of Markov dependence of an observed process governed by a hidden Markov model. *Scientific Programming* **10,** 241–251.

Boys, R. J., Henderson, D. A., and Wilkinson, D. J. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics* **49,** 269–285.

Braun, J. V. and Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science* **13,** 142–162.

Braun, J. V., Braun, R. K., and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87,** 301–314.

Cappé, O., Robert, C. P., and Rydén, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B* **65,** 679–700.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* **75,** 79–97.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51,** 79–94.

Churchill, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Computers and Chemistry* **16,** 107–115.

Fan, T.-H. and Tsai, C.-A. (1999). A Bayesian method in determining the order of a finite state Markov chain. *Communications in Statistics—Theory and Methods* **28,** 1711–1730.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort, and S. Richardson (eds), 179–198. Oxford: Oxford University Press.

Lewin, B. (2000). *Genes VII*, 7th edition. Oxford: Oxford University Press.

Li, W. (2004). *Bibliography on Features, Patterns, Correlations in DNA and Protein Sequences*. Online bibliography. Available from `http://www.nslij-genetics.org/dnacorr/`.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15,** 38–52.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. London: Chapman and Hall.

Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMPSTAT '98 Proceedings in Computational Statistics*, R. W. Payne and P. J. Green (eds), 89–100. Heidelberg: Physica-Verlag.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B., and Bressières, P. (2002). Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Research* **30,** 1418–1426.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59,** 731–792.

Robert, C. P., Celeux, G., and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic

implementation. *Statistics and Probability Letters* **16,** 77–83.

Robert, C. P., Rydén, T., and Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B* **62,** 57–75.

Skalka, A., Burgi, E., and Hershey, A. D. (1968). Segmental distribution of nucleotides in the DNA of bacteriophage lambda. *Journal of Molecular Biology* **34,** 1–16.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B* **62,** 795–809.

Viallefont, V., Richardson, S., and Green, P. J. (2002). Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics* **14,** 181–202.

Received April 2002. *Revised January* 2003. *Accepted March* 2003.

Discussions on "A Bayesian Approach to DNA Sequence Segmentation"

By: **Charles Lawrence and Ivan Auger**
Wadsworth Center
New York State Health Department
Albany, New York 12201, U.S.A.

The use of HMMs with Markov models of segments is now one of the old statistical methods in the young science of bioinformatics. In addition to the references cited by Boys and Henderson, this approach is a fundamental component of gene-finding algorithms that have been used extensively to identify and delineate genes in the human genome and other genomes. For example, the algorithms HMMgene (Krogh, 1997) and Genescan (Burge and Karlin, 1997) take this approach. In spite of this substantial history, Boys and Henderson have addressed a largely unexplored component of this problem. Thus, through the use of reversible jumps for posterior inferences of the number of segment models appropriate for the characterization of a DNA sequence, they make a very important contribution to this field.

Unfortunately, the sensitivity analysis they report indicates that the resulting inferences maybe fairly sensitive to the assumed prior distribution on the HMMs' hidden state transition parameters. This may be the consequence of the fact that the information content of over 1000 transitions in the specification of these priors is a long way from the assumption of a uniform distribution.

Boys and Henderson suggest that HMMs are somehow uniquely appropriate in circumstances in which noncontiguous parts of the sequence are described by the same Markov model. This is not the case. Perhaps the perception that change-point algorithms were not appropriate in this circumstance stems from the fact that current change-point implementations for DNA sequence segmentation seek direct and thus exact posterior inferences on all unknowns (Liu and Lawrence, 1999) rather than resorting to an iterative MCMC approach. To achieve this end, these algorithms must simultaneously sum and integrate over all unknowns in the problem in order to marginalize over all unknowns to obtain the marginal likelihood, the key normalizing constant. This ability is lost when noncontiguous parts of the sequence are described by a common Markov model. In this case one is forced to resort to an iterative MCMC approach to produce posterior inferences from either HMMs or change-point models. Thus, the ability to identify models common to multiple noncontiguous segments is associated with the use of MCMC algorithms rather than the use of HMM instead of change-point algorithms. Another key distinction between HMMs and change-point algorithms concerns the distributional form of segment lengths. HMMs imply geometric distributions for segment lengths, while change-point algorithms used in this context assume that all segmentations of the sequences with exactly k change points are equally likely. This difference suggests that it may be productive to explore the potential of change-point algorithms to produce inferences that are less sensitive to prior specification.

In the analysis of the bacteriophage genome, an annotated diagram of the genome with the coding regions with direction of transcription would be useful. It is interesting that segment types 1 and 2 transcribe in the forward strand and all the other ones in the reverse strand. It is possible that there are fewer segment types and that models 1 and 2 are reverse complements of segment types 3–6. This seems to be a particularly important avenue to explore when DNA sequences contain coding regions. However, because there has already been much good work on gene-finding algorithms whose key focus is on distinguishing coding from noncoding sequence, the greatest potential for the applications of the procedures described here may be in the analysis of noncoding DNA sequences. Unfortunately, all too often in the biometrics community, a single available data set is repeatedly analyzed by many groups even when it may not be particularly well suited for illustrating the advantages of a new approach. If, as we believe, the future in the field of biometrics lies in far greater immersion in biology of the 21st century, then this unfortunate tendency will automatically fade as the community becomes far more knowledgeable about biochemistry, genetics, genomics, and the wonderful data resources that are emerging from the biotechnology/genomics revolution.

References

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268,** 78–94.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *Proceedings of the International Conference on Intelligence Systems on Molecular Biology* **5,** 179–186.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15,** 38–52.

By: **Mayetri Gupta**

Department of Biostatistics, University of North Carolina Chapel Hill, North Carolina 27599, U.S.A.

and

Jun S. Liu

Department of Statistics, Harvard University Cambridge, Massachusetts 02138, U.S.A.

We congratulate the authors for presenting a comprehensive Bayesian treatment to the problem of compositional variations in genomic DNA sequences and for introducing interesting extensions of the standard hidden Markov model (HMM; see Churchill, 1989; Muri, 1998). The hidden states in the authors' model represent the unknown segment types, and the observed nucleotides are assumed to be generated from a distribution completely specified by the underlying segment type. Churchill's basic HMM requires a finite number of states to be fixed in advance and the nucleotides within a segment type to follow the i.i.d. multinomial model. The form of the augmented data likelihood then is of the regular exponential family so that an ML estimation algorithm can be set up through the EM framework (Dempster, Laird, and Rubin, 1977). Churchill later extended his HMM to the case where the observed outcomes follow a first-order Markov chain, under which the estimation algorithm can be implemented through minor adjustments to the i.i.d. case. In the present article, the authors generalize Churchill's model to allow (i) an unknown number of states, and (ii) an unknown order of Markovian dependence between the sequence of observations. It is important to note that in both Churchill's model and the current authors', the underlying state boundaries are ignored when specifying the order of dependence of the observed sequence, though the transition probabilities may depend on the underlying state sequence. We discuss the implications of this later in more detail. Because the model becomes quite complicated after the authors' generalizations, the authors developed several MCMC algorithms to simultaneously estimate the total number of states and the order of dependence between observations, for an optimal sequence segmentation. However, as more uncertainty is introduced into the model, a reliable inference is more difficult to obtain and the identifiability of the parameters becomes an issue.

1. Duration Modeling

One inevitable phenomenon while modeling state transitions through a Markov chain is that the resulting distribution of the duration of the chain in a certain state is geometric. More specifically, in a two-state model with states 0 and 1, and transition probabilities p_{00}, $1 - p_{00}$, $1 - p_{11}$, p_{11}, respectively, let us denote the number of consecutive steps that the chain stays in state 1 as L_1. Then

$$P(L_1 = y) = p_{11}^{y-1}(1 - p_{11}), \quad y = 1, 2, \ldots$$

This exponentially decaying distribution of lengths is often considered inappropriate in DNA segmentation applications. More complex length distributions can be modeled by introducing an array of several hidden states (e.g., Durbin et al., 1998). For example, we can think of a transition within state 1 as an occurrence of n consecutive states (say $S_{11}, \ldots, S_{1n}$), each such state S_{1i} having probability p_{11} of transitioning to itself and $1 - p_{11}$ of moving to the next state $S_{1,i+1}$. The resulting distribution is then negative binomial (n, p_{11}),

$$P(L_1 = y) = \binom{y + n - 1}{n - 1} p_{11}^{y}(1 - p_{11})^{n}, \quad y = 0, 1, \ldots$$

Parameters n and p_{11} can be empirically estimated from the mean and variance of segment lengths from training sequences. This idea can be easily applied to a multistate model where it may provide greater flexibility and a more accurate fit to the empirical length distribution.

2. The Generalized HMM versus the Segmentation Model

In this context arise two questions—first, is it proper to consider the order q of Markovian dependence among the observation sequence the same irrespective of the underlying state? Second, what does it mean to allow Markovian dependence between observations that may not be generated from the same state? For neighboring segments, which are assumed to have different mechanisms of generation, it is a little difficult to understand what a "dependence" between, say, p residues of state 3, and the neighboring $q - p$ residues of state 1 may scientifically mean. Under the current model specifications, however, it is difficult to restrict interresidue dependence within distinct segment types without adding an amount of computational awkwardness. In this context, it may be of interest to consider a segmentation model (Liu and Lawrence, 1999) that incorporates a higher-order dependence (that may be state specific), and simultaneously avoids the problem of state overlapping of residues.

Liu and Lawrence's segmentation model can be thought of as a four-sided (for DNA) "coin-toss" model with multiple coins (a maximum of $r_{\max}$) having runs of unknown length. State transitions between different coin types are assumed i.i.d., as well as the sequence of coin tosses. The total number of states is assumed unknown, and estimated under the model (along with length of segments) through a Monte Carlo approach employing dynamic programming-like recursions. Following Boys and Henderson, we let the observed residue sequence be $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and the corresponding hidden state sequence be $\boldsymbol{S} = (S_1, \ldots, S_n)$. A higher-order dependence version of this model with q_i-order Markovian dependence under state i may be formulated as having the

complete-data likelihood (under the partition $\boldsymbol{A} = (A_0, \ldots, A_{r_{\max}})$, where $A_i = \max_j \{j : 1 \le j \le n;\, S_j = i\}$, $A_0 = 0$):

$$P(Y_1, \ldots, Y_n, A \mid \theta, r_{\max}) = \prod_{r=1}^{r_{\max}} P\big(Y_{[A_{r-1}:A_r]} \mid \theta_r\big) P(A \mid r_{\max}),$$

where for $r \in (1, r_{\max})$, and $A_{r-1} + q_r \le j \le A_r, (i \in \{1, \ldots, b^{q_r}\}; j \in \{1, \ldots, b\})$,

$$P\big(Y_j \mid Y_{[A_{r-1}:j-1]}, \theta_r, S_j = r\big) = P(Y_j \mid Y_{j-1}, \ldots, Y_{j-q_r}, \theta_r) = p_{ij}^{(r)}.$$

In this model, the order of dependence is only *conditionally* specified depending on the underlying state; and hence there is no occurrence of residue dependence overlapping across states. Generalization to more complex application-specific segmentation models can likewise be obtained using individual segment models (e.g., Schmidler, Liu, and Brutlag, 2000). Another aspect of the segmentation model is that unlike the possible repetitions of states in an HMM, here every segment is considered to be generated under a run of a state that occurs only one time in the sequence. However, this is not really a drawback of the model as any two nonadjacent segments can be tested for similarity of properties (e.g., q-order transition frequencies for some q) and may be collapsed into a single state if necessary. A between-segment Markovian model may also be incorporated, with minor modifications to the dynamic programming recursions. Conceptually, it is possible to develop a Gibbs sampling-like algorithm that samples r and the q_i's from their conditional posterior distributions and updates the partitions given r and $\boldsymbol{q}$. Alternatively, evolutionary Monte Carlo–type moves (Liang and Wong, 2000) could be used for sampling segmentations (given r and $\boldsymbol{q}$) that could lead to a faster algorithm.

3. Relation of State Identifiability to Order of Dependence

It is of both theoretical and practical interest to investigate the unidentifiability issue of the problem. Theoretically, adding layers of complexity to the structure of the model (such as a higher-order dependence) may make it more attractive for a scenario of application. However, how flexible can the model remain (viz. unknown number of states, unknown order of dependence, and resulting variability and increase in the number of model parameters) before the problem of unidentifiability makes it impracticable? To be more specific, consider a binary sequence composed solely of '0's and '1's in a repetitive pattern of three types of segments (only '0's, only '1's, alternating '0' and '1'):

$$0000000000000000000111111111111111111$$
$$0101010101010101010000000000000000000\ldots$$

Using an i.i.d. model, it is easy to determine that there are three distinct segment types. However, when we assume that the dependence structure is a first-order Markovian, the three-state model and a two-state model (with the '0'- and '1'-runs grouped into one state and the '01'-runs into another) provide virtually identical posterior likelihoods (assuming identical priors under each state). In the absence of knowledge of the true number of states, it would be a challenge for any of the above-discussed algorithms that incorporate higher-order Markov dependence to find the correct number of segments.

4. Conclusion

Boys and Henderson describe some interesting extensions to the current ideas of DNA sequence segmentation in a Bayesian framework and sound mathematical algorithms to implement these. Most importantly, their article raises fundamental questions about limitations in model specification and brings to the forefront the issue of how far one can refrain from making prior assumptions about a model while keeping it feasible in practice. The question remains an open one and any practitioner should exercise extreme caution during the actual implementation of all such models and algorithms to avoid arriving at misleading conclusions.

References

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51,** 79–94.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis.* Cambridge, U.K.: Cambridge University Press.

Liang, F. and Wong, W. H. (2000). Evolutionary Monte Carlo: Applications to C_p model sampling and change point problem. *Statistica Sinica* **10,** 317–342.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15,** 38–52.

Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMPSTAT '98 Proceedings in Computational Statistics*, R. W. Payne and P. J. Green (eds), 89–100. Heidelberg: Physica-Verlag.

Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *Journal of Computational Biology* **7,** 233–248.

By: **Terry Speed**

Division of Genetics and Bioinformatics
the Walter and Eliza Hall Institute of Medical Research
Parkville, Victoria 3052, Australia

The authors have done a very fine job illustrating the Bayesian approach to the problem of DNA sequence segmentation. Their formulation of the problem is clean, their parameterization natural, and their MCMC algorithms elegant and powerful. Their results add yet more weight to the already strong argument for the use of Bayesian MCMC methods in biological sequence analysis. What can I add to their excellent article? It seems to me, just two things. One, to make sure readers of *Biometrics* appreciate the enormity of their understatement, when they write: "HMMs have become a popular choice for the analysis of DNA sequences," and give just three references to papers like theirs. The other is to remark that we can and should ask for more biological

motivation before embarking on exercises like the one in this article, and that our assessments of the scientific value of biological sequence analysis can and should be more penetrating.

After reading their article, I could not help wondering how the authors managed *not* to mention the extremely effective applications of HMMs (and generalized HMMs and generalized pair HMMs) to ab initio gene finding. In my view this work is one of the great success stories involving statistics in modern biology, and it deserves to be more widely appreciated. As with the present article, gene finding is a segmentation problem: identifying intergenic and genic regions within DNA, with the genic regions having a great deal of further structure: transcribed but untranslated segments, exons, introns, polyadenylation, splice, translation start and stop signals, and so on. And as with the present article, the workhorse is the HMM, or some generalization thereof. Widely used programs include `FGENESH+`, `GENSCAN`, and `SLAM`. In principle, all of the methods of the present article should apply (with suitable modification or extension) to gene finding, and it is an interesting question to ask whether they would in practice, and if so, what the gains would be. Exact calculations for multispecies gene finding with "tree HMMs," merging the pair HMM approach of Alexandersson, Cawley, and Pachter (2003) with that of Siepel and Haussler (2003) is unlikely to be feasible, so it seems likely that MCMC methods like those in the article will be necessary to solve this problem. That would be a problem worthy of attack!

What is the biological motivation for an article like the present one, and how should we assess its results? In 1989, Churchill applied (exact, non-Bayesian) techniques similar to those of the present article to segmenting the *E. coli* genome into regions with high and low GC content, replacing the simpler method of thresholding local GC content. That was a well-defined question, and HMMs provided a natural framework for answering it. Both the number of segment types and their meaning were known a priori, and it was easy to check whether HMMs gave a good answer. By contrast, inferring compositional heterogeneity in phage λ is offered in this article simply as an illustration of the techniques, rather than a problem to be solved or a question to be answered. What did we learn from applying the techniques of the article to λ? Roughly, that each of the about $r = 6$ different segment types seems to correspond to a gene or set of genes, sharing a common composition. It is not clear that the extra flexibility obtained from being able to allow the data and the model to determine the number of segment types actually led to any biological insights. However, using techniques similar to those in this article, Nicolas et al. (2002) found DNA sequence characteristic of phages in the *B. subtilis* genome, and so it might be expected that phage λ genome harbors DNA sequence characteristic of its host, *E. coli*, which got there by some form of horizontal transfer. Indeed it does, and a comparison of Figure 1 of this article for $j = 1$ with Figure 1a of Scherer, McPeek, and Speed (1994) suggests that segment type 1 is typical *E. coli* sequence, though the correspondence between these two figures is not perfect. Perhaps there are similar interpretations for the other five segment types, but I do not know them.

Let me close with a question for the authors. Was the segmentation obtained when $q = 0$ very different from that when $q = 2$?

References

Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM—Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research* **13,** 496–502.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B., and Bressières, P. (2002). Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Research* **30,** 1418–1426.

Scherer, S., McPeek, M. S., and Speed, T. P. (1994). Atypical regions in large genomic DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 7134–7138.

Siepel, A. and Haussler, D. (2003). Combining phylogenetic and hidden Markov models in biosequence analysis. *Proceedings of the Seventh Annual International Conference on Research in Computational Biology*, 277–286.

By: **Jeffrey L. Thorne**

Bioinformatics Research Center, Box 7566
North Carolina State University
Raleigh, North Carolina 27695-7566, U.S.A.

Following the pioneering application to biological sequence data by Churchill (1989), hidden Markov models (HMMs) have assumed a central position in bioinformatics (see Durbin et al., 1998 for a good overview). To my knowledge, Boys and Henderson introduce here the most general existing application of HMMs to the analysis of DNA sequence data. With the Boys and Henderson approach, little need be prespecified. The approach does not require the number of underlying compositional categories of DNA to be known in advance nor does it require the probabilities of transitions between these categories to be predetermined. Likewise, the frequencies of the four nucleotide residues within each compositional category are free to vary.

This exploratory approach has advantages. With the availability of genomic sequence data, hypothesis-driven research has become unfashionable (e.g., Lawrence, 2001). Large data sets can be mined for unusual patterns. Once identified, the biological underpinnings of these unusual patterns can be determined. General and flexible statistical approaches such as this one by Boys and Henderson have much to recommend them.

However, these general and flexible approaches also have shortcomings which are sometimes overlooked in the statistics community. While it is statistically satisfying to estimate the parameters of a relatively general model, it may be challenging to interpret these parameters. Regardless of whether it can be well estimated, a parameter that has an unclear biological meaning prior to data analysis may continue to have an unclear meaning following the analysis.

Although Boys and Henderson have developed a general and flexible approach to characterizing compositional

heterogeneity in DNA sequences, it is not necessarily going to be straightforward to interpret its results. In the bacteriophage lambda example analysis that they present, genes can be assigned to particular compositional categories but it is less clear whether the genes within each category are connected by some biological property beyond DNA composition and the direction of transcription. When parameters can be assigned a specific interpretation prior to data analysis, then the connection of the inferences to the biological system being studied is easy to establish. For example, HMM-based analyses of single DNA sequences have proven very successful in gene finding (e.g., Burge and Karlin, 1997) and extensions to multiple aligned sequences (Pedersen and Hein, 2003) are promising. With these gene-finding techniques, hidden states have a precise biological meaning such as "phase 0 intron" or "promoter."

How can the output of their computational strategy be better exploited to advance the state of biological knowledge? This is the question that I hope Boys and Henderson address in their future work. The value of data-driven research is completely dependent upon how much it can illuminate the system being studied.

References

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268,** 78–94.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51,** 79–94.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, U.K.: Cambridge University Press.

Lawrence, C. (2001). New data driven research paradigm for the post genome era: A transcription regulation example. *Genome Informatics* **12,** 225.

Pedersen, J. S. and Hein, J. (2003). Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* **19**(2), 219–227.

The authors replied as follows:

We thank the Co-Editor for inviting discussion of our article and the discussants for their thought-provoking comments. The discussion reflects the tensions inherent in modeling the underlying biological process and ranges from points about model choice and computational simplicity to biological motivation, relevance, and interpretation.

1. Model Framework

Several important modeling issues are raised. To begin with there is the choice of overall model structure, essentially a choice between the hidden Markov model (HMM) used in our article, and the multiple change-point segmentation model (Liu and Lawrence, 1999) referred to by Gupta and Liu, and Lawrence and Auger. The change-point model has a computational advantage over the HMM in that posterior samples can be obtained without using (approximate, iterative) MCMC methods. However, if it were adapted to include the biologically more plausible scenario in which noncontiguous parts of the sequence had the same composition then the crucial intersegment independence would be lost, together with the associated computational advantages. Of course, as Gupta and Liu suggest, it is possible to assess compositional similarity in noncontiguous segments but this might just lead to a redistribution of computational effort from the model implementation stage to the follow-up analysis. The computational benefit of using HMMs in this context is the availability of a forward–backward algorithm from which to simulate segmentations efficiently (Scott, 2002) and this advantage may be substantial when analyzing long sequences.

2. Segment Lengths

Both Lawrence and Auger, and Gupta and Liu, comment on the geometric segment length distribution inherent in the first-order Markov assumption for the hidden layer in the HMM. A fundamental objection in using the geometric distribution to model segment lengths is that the whole shape of the distribution, particularly the exponential decay, is not consistent with the lengths of "known" segments. Unfortunately, similar distributional shapes for segment lengths can also occur in the change-point model: When there are m change points in total, the prior distribution for the position of the $(k+1)$st change point τ_{k+1} given the position of the kth change point τ_k for a sequence of length n is

$$\Pr(\tau_{k+1} = j | \tau_k = i) = \binom{n-j-1}{m-k-1} \Big/ \binom{n-i-1}{m-k}, \quad j = i+1, \ldots, n-m+k,$$

and this leads to a nonincreasing distribution for $\tau_{k+1} - \tau_k$, which can be lighter tailed than a geometric distribution, and frequently is for moderate to large n and m. Furthermore, in many applications, people seek to be uninformative about the number of change points m, and this can be represented by using a discrete uniform distribution on $\{0, 1, \ldots, n-1\}$, that is, where each model is equally likely. However, this uniform distribution is also the marginal distribution for m when $m \,|\, p$ has a binomial $B(n, p)$ distribution and an uninformative uniform prior distribution is used for p. For a given p, we can think of the change point locations as binary indicators occurring within a Bernoulli process with parameter p, and consequently the distribution of segment lengths is geometric. Thus, the problem of an inappropriate distributional shape can be common to both modeling frameworks.

A remedy commonly used to overcome the objections to geometric segment lengths is to generalize the HMM to allow explicit duration modeling (of segment lengths) using *hidden semi-Markov models* (HSMMs). These models have been used to great effect in gene-finding algorithms such as the `GENSCAN` algorithm (Burge and Karlin, 1997) mentioned by

the discussants (and in the original, longer, version of our article). However, they do suffer from an additional computational overhead due to the complexity of implementing the associated forward–backward algorithms (e.g., Guédon, 2003). In our fully Bayesian framework, this is somewhat undesirable given the already substantial computational burden of using MCMC techniques to implement our HMM model. However, as Gupta and Liu suggest, negative binomial (d, λ_{ii}) segment lengths (with known integer parameter d) can be incorporated easily into the HMM framework by introducing $d-1$ extra hidden states for each of the r current hidden states. The resulting HMM has a larger hidden state space with a special transition structure and so this HSMM can be fitted by using only minor modifications to our method. Also, when d is not too large, this added flexibility can be achieved at a relatively small additional computational cost. Allowing even more flexibility through unknown d or using other segment length distributions would require the use of the more computationally intensive algorithms already developed for HSMMs. Clearly, our HMM can be made more biologically plausible. However, in the absence of specific information about segment length distributions, we believe our model is a sensible compromise between biological plausibility and computational complexity.

3. Base Dependence

We now turn to comments concerning models for the observed DNA sequence within homogeneous segments. Gupta and Liu suggest a model in which the order of Markov dependence q_i is possibly different for each segment type. This might have more biological merit than our homogeneous dependence model and, when using conjugate priors for the base (residue) transition probabilities, it is straightforward to accommodate in each of the HMM, HSMM, and change-point models. For the HMM, the extension to state-specific dependence is achieved simply by adapting the MCMC algorithm so that, at each iteration, the sampling step for q is replaced by r such sampling steps for q_i, $i = 1, 2, \ldots, r$. If dependence between bases across segment boundaries is restricted then the algorithm would become considerably more cumbersome, as Gupta and Liu point out. There is little doubt that such restrictions would improve the biological credibility of the model but we are uncertain as to whether they would yield sufficient additional insight, particularly when $q_{\max}$ is small and both the sequence and segments are long. It would be interesting to investigate whether incorporating this feature really does produce a more meaningful segmentation. For their extension of the change-point model to state-specific dependence Gupta and Liu suggest a Gibbs sampling-like update for the q_i, equivalent to that for the HMM. This extension could also be implemented, perhaps more efficiently, using dynamic programming recursions which integrate out uncertainty regarding the q_i (Fearnhead, 2003).

The discussants query whether the extra complexity of Markov dependence is necessary, be it state-specific q_i or simply $q > 0$. For example, Speed questions whether there are differences between segmentations when $q = 0$ and $q = 2$ for the bacteriophage *lambda* example. Although the posterior distribution for the segmentation is both high dimensional and explicitly dependent on the number of segment types r, a simple comparison can be made using the MPM estimates of the hidden states $\boldsymbol{s}$ conditional on the a posteriori most probable model; here, this is $r = 7$ when $q = 0$ and $r = 6$ when $q = 2$. Figure 1 gives these MPM estimates and also the locations of genes and their directions of transcription (indicated by the arrows) as requested by Lawrence and Auger.

The segmentations look fairly similar, with slightly more segments evident in the larger model ($q = 2$). Which model gives the "better" segmentation depends to some extent on the insight it gives to biologists. Of course, *if* the underlying base transition structure is Markovian and *if* it is related to gene function then using an appropriate choice of q (or q_i) will give a more appropriate classification of genes to segment types. In this sense the $q = 2$ model provides a better description of the sequence as it has considerably more posterior probability than the simpler model. Further details of a related $q = 0$ analysis of this sequence, with the restriction to $r_{\max} = 9$, can be found in Boys and Henderson (2001).

4. Model Identifiability

The issue of model complexity is related to Gupta and Liu's comments on model identifiability and, in particular, whether different (r, q) combinations are (roughly) equally plausible a posteriori. Clearly the problem is an important one when there is high posterior dependence between r and q. In such cases, considerations such as parsimony and biological interpretability of the segment groupings may yield a satisfactory solution. For our bacteriophage *lambda* example, there is very little uncertainty regarding q and so, fortunately, the problem does not arise. Turning to Gupta and Liu's binary data example, our analysis of their data (assuming $q = 1$) suggests that there is very strong evidence to prefer a three-state model with known segmentation to a two-state model with known segmentation, which disagrees with Gupta and Liu's conclusion. Assuming a uniform Dirichlet prior for the base transition probabilities and allowing base dependence between segments, the (log) *marginal likelihoods* (prior predictives) are -17.522 and -12.408 for the two-state and three-state models, respectively, which gives a Bayes factor of approximately 166 in favor of the three-state model. Furthermore, the three-state model is to be preferred to a two-state model regardless of whether $q = 0$, 1, or 2. In that sense the models are clearly distinguishable. However, we agree with Gupta and Liu that, in general, correctly identifying the segmentation when the sequence is only weakly informative about the number of states would be a considerable challenge for any algorithm.

5. Sensitivity to Prior Assumptions

An important issue raised by Lawrence and Auger is the sensitivity of posterior inference to prior assumptions. The prior information we used for the transition structure of the hidden states (Λ) is equivalent to that from a sequence whose length is a similar order of magnitude to the length of the observed DNA sequence, which admittedly is large. The similarity between prior and posterior means for Λ (given r) could be due to the prior dominating the likelihood function, as implied by Lawrence and Auger. However, calculation of the (univariate)

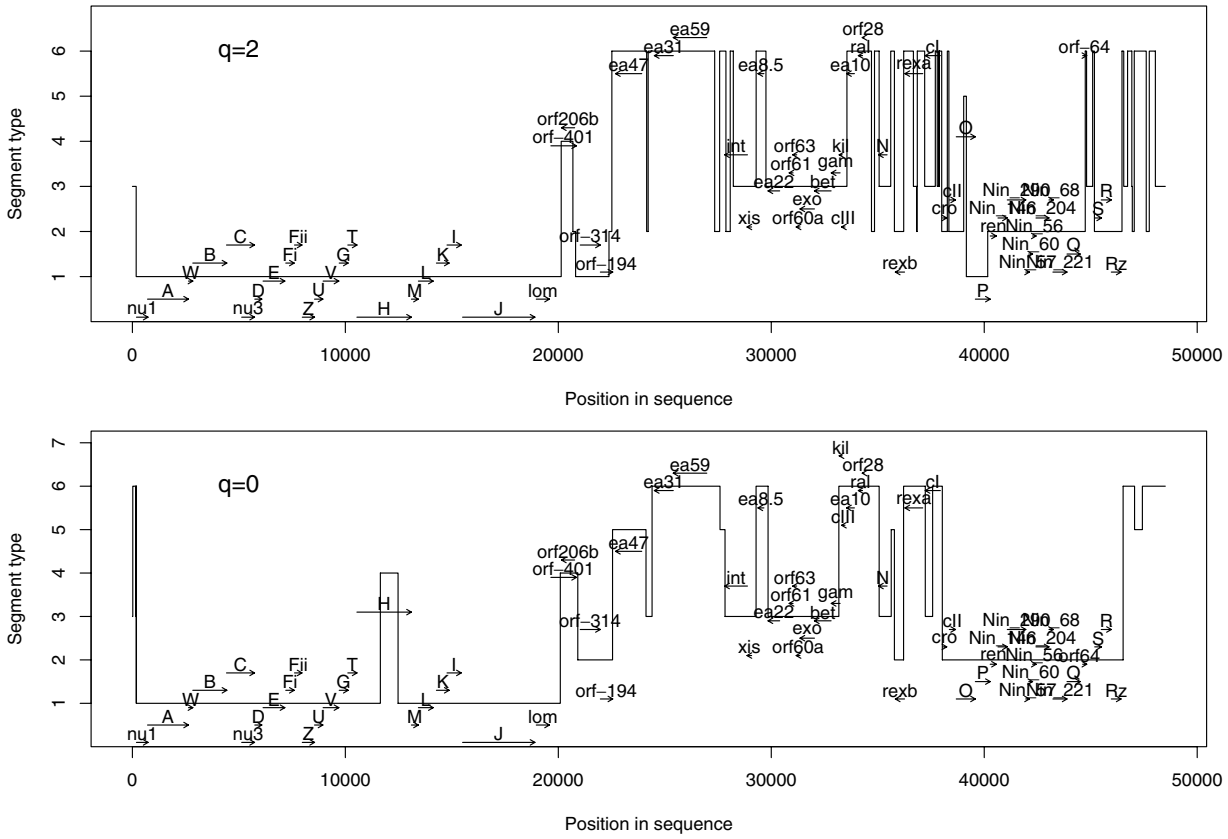

Figure 1. MPM estimates conditional on the model with $q = 2$, $r = 6$ (top), and $q = 0$, $r = 7$ (bottom).

marginal likelihood (via a ratio of a kernel density of the posterior output and the prior density) for each element reveals that the data are, in fact, fairly informative. We now recognize that the sensitivity of the posterior distribution mentioned in the article when changing the mean segment length to 100 is due to a conflict between this strong prior opinion and the data. This insight highlights the benefit of using such calculations. They also reveal that the data are reasonably informative about the other parameters and therefore that their posterior inferences should be fairly robust to modest changes in their prior specification. The allocation of genes to segment types is also reasonably robust.

6. Biological Motivation and Interpretation

We now consider the comments about biological motivation and interpretation. Speed asks what is the main biological motivation for an article such as ours. Our motivation is expressed succinctly by Thorne when he recognizes the role of segmentation algorithms as providing exploratory tools for data-driven research. Lawrence and Auger's scenario of biological scientists being "*buried in data that they don't know how to analyze*" is also a motivating factor in developing our data-driven approach. The simple HMMs and change-point models do make a serious attempt to describe general DNA sequences but inevitably fall somewhat short. We agree with Lawrence and Auger that our method may be more suited to analyzing less well-understood noncoding DNA sequences. Indeed, in an earlier paper (Boys, Henderson, and Wilkinson, 2000) we applied similar but more basic techniques (fixed r and q) to study an intron of the chimpanzee α-fetoprotein gene.

Speed and Thorne are concerned about what has been learned from our analysis of the bacteriophage *lambda* genome. Lawrence and Auger conjecture that some of the segment types are reverse complements of each other and Speed suggests that there might be similarities between the characteristics of our *lambda* segmentation and features of the host *E. coli* organism, both points worthy of investigation. It is true that our analysis did not lead to any further biological insight but our biologist colleague did feel that our gene groupings made biological sense, as did audiences at meetings where this work has been presented, and perhaps this is all that can be expected of such exploratory techniques. The motivation for using this particular DNA sequence was to facilitate comparison of our results with those made using other segmentation algorithms. Such analyses do serve a useful purpose in highlighting limitations of potential models which might otherwise be overlooked in an analysis of a carefully chosen sequence. However, we do agree with the general thrust of Lawrence and Auger's point that the presentation of results should not be restricted to those for a relatively small number of benchmark data sets.

The problem of interpreting model parameters when they are specified quite generally prior to observing the data is raised by both Thorne and Speed. Our analysis of the *lambda*

phage is a quite general exploratory investigation and has focused primarily on segmentation and classification rather than on interpretation. In such circumstances, where parameter interpretation is a secondary concern to sequence description, predictive methods which integrate out all parameter uncertainty can be instructive; see, for example, Figure 6 in Liu and Lawrence (1999) and Figure 3 in Boys and Henderson (2001). However, in many cases, biological meaning is of primary importance and this can be aided by describing the attributes of the hidden states in the prior distribution through a judicious choice of their associated Dirichlet parameters. For example, this approach can help locate segments containing genes from families with particular base transition structures. Biological information may also be included by specifying likely (prior) values for elements of the transition structure of the hidden layer, such as restricting the ordering of segment types using an asymmetric matrix.

7. The Future

The advent of simulation-based Bayesian inference has led to an explosion in the statistical analysis of complex models in many fields of application. In particular, these techniques are ideally suited to the analysis of latent process models and so can be deployed very effectively to study many of the problems of current biological interest. Of course, the analyses also gain considerably from the opportunity to input additional (prior) information and from the natural scientific interpretation that results from adopting a subjective view of probability. The discussants mention some of the statistical research into important biological questions, and progress is being made rapidly on many fronts. It is certainly an exciting time to participate at the interface between these two subjects.

References

Boys, R. J. and Henderson, D. A. (2001). A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Computing Science and Statistics* **33,** 35–49.

Boys, R. J., Henderson, D. A., and Wilkinson, D. J. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics* **49,** 269–285.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268,** 78–94.

Fearnhead, P. (2003). Exact Bayesian curve fitting and signal segmentation. Available at `http://www.maths.lancs.ac.uk/~fearnhea/PS_reg.ps`.

Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics* **12,** 604–639.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15,** 38–52.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* **97,** 337–351.