THE PERFORMANCE OF A REVERSIBLE JUMP MARKOV CHAIN MONTE CARLO ALGORITHM FOR DNA SEQUENCES ALIGNMENT

LUIS J. ÁLVAREZ
Instituto de Matemáticas – UNAM, Unidad Cuernavaca,
Av. Universidad, s/n
Lomas de Chamilpa
62210 Cuernavaca, Morelos, México.
E-mail: lja@matcuer.unam.mx

NANCY L. GARCIA
Departamento de Estatística, IMECC - UNICAMP
Caixa Postal 6065
13081-970 - Campinas, SP - Brasil
E-mail: nancy@ime.unicamp.br

ELIANE R. RODRIGUES
Instituto de Matemáticas – UNAM
Area de la Investigación Científica
Circuito Exterior, Ciudad Universitaria
México, D.F. 04510, México
E-mail: eliane@math.unam.mx

Assume that K independent copies are made from a common prototype DNA sequence whose length is considered to be a random variable. In this paper the problem of aligning these copies and therefore the problem of estimating the prototype sequence that produced the copies is addressed. A hidden Markov chain is used to model the copying procedure and a reversible jump Markov chain Monte Carlo algorithm is used to sample the parameters of the model from their posterior distribution. Using the sample obtained, the Bayesian model selection may be made and the prototype sequence may be selected using the *maximum a posteriori* estimate. A prior distribution for the prototype DNA sequence that incorporates a correlation among neighbouring bases is also considered. Additionally, an analysis of the performance of the algorithm is presented when different scenarios are taken into account.

KEY WORDS: Bayesian inference; Sequences alignment; Reversible jump Markov Chain Monte Carlo method; Hidden Markov model; Potts model.

1 INTRODUCTION

One of the aims of GENOME Projects is to decode the genetic code of living creatures. The decoding procedure used to read the genetic code in a region of interest (for example, a portion of a chromosome) is such that some discrepancies may occur from one decoded sequence to another (sequences coding the same region). Thus, it is necessary to obtain an alignment of the copies produced from a common prototype sequence so this prototype sequence may be inferred. (For more information about DNA decoding see, for example, Apostolico and Giancarlo (1999), Blackwell (1993), Drasdo *et al.* (1998), Liu and Lawrence (1995), Liu *et al.* (1995, 1999), Meidanis and Setubal (1995), Milanesi *et al.* (1999), Schleif (1993), Waterman (1989a, 1989b), Weir (1985) and references therein.)

Some methods in the literature use hidden Markov chains to model the base composition of each fragment and with that construct an alignment by maximum likelihood (see, for example, Bishop and Thompson (1986), Churchill (1989, 1992, 1995) and Krogh *et al.* (1994), Thorne and Churchill (1995), Thorne *et al.* (1991)). However, the methods and the initial distribution for the data, used up to now, do not take into account the correlation among neighbouring bases present in a DNA sequence. The present work intends to take into account that type of correlation. Another problem with most of the existing methods is that they are static in the sense that they reconstruct a DNA sequence after fixing the maximum alignment length.

In this work a two-step method for producing an alignment is proposed. Even though in the beginning a hidden Markov chain is used to obtain a likelihood function for the sequences copied from the prototype, the present work differs from previous ones when a reversible jump Markov chain Monte Carlo method is used to select the Bayesian model and to estimate the parameters of the model. This is made in the following way. The prototype sequence, its length and the parameters of the hidden Markov chain are considered parameters of the Bayesian model. After obtaining the likelihood function, the reversible jump Markov chain Monte Carlo method is used to obtain samples from the joint posterior distribution of the parameters. These sampled values are then used to select the Bayesian model and within this model the sequence that has the largest marginal posterior distribution is the one chosen to represent the prototype DNA word. Garcia and Rodrigues (1999) use reversible jump Markov chain Monte Carlo in a similar problem. However, the DNA decoding procedure is not modelled using hidden Markov chains. In Garcia and Rodrigues (2001) hidden Markov chains are used to model the decoding procedure. Nevertheless, in neither work an implementation of the algorithm is presented.

Unless otherwise stated, the following assumptions are considered throughout this paper. The order in which the bases appear in the decoded sequences are final. The possible mutations are the usual ones: deletion, insertion and replacement. These mutations may be considered as produced during the decoding procedure (i.e., if there are not enough pieces of DNA ending in a specific position, the letter appearing at that position may not be read and consequently the base is deleted - bases may also be misread or inserted). Transposition between two consecutive bases is considered an occurrence of two substitutions. Bases are read one by one by the scanner. Homogeneity within a sequence is also assumed. (For heterogeneous sequences one may use the approach presented by Boys and Henderson (2003) to identify the homogeneous segments in heterogeneous sequences and then apply the procedure described here to each homogeneous segment.)

The outline of the paper is as follows. In Section 2 the basic assumptions for the hidden Markov model are given. The Bayesian model is described in Section 3 and the spatial correlation between sites in the prototype sequence is described as a four colour Potts model (commonly used in image restoration problems - Wu (1982)). Section 4 presents a reversible jump Markov chain Monte Carlo algorithm used to obtain a sample from the *joint posterior distribution* of the parameters of the Bayesian model. Some simulated results obtained by the implementation of the algorithm proposed in Section 4 is presented in Section 5. Finally, in the last section some remarks about the method proposed in this paper are made.

2 A HIDDEN MARKOV MODEL

Let M be a random variable assuming values on $\{1, 2, \ldots\}$; $K \geq 1$ be a known and fixed natural number; and $\mathbf{X}^{(M)} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$, where $\mathbf{X}_i \in \{\text{A,C,G,T}\} = \mathcal{A}$, be a DNA word called prototype sequence. Assume that K independent copies of $\mathbf{X}^{(M)}$ are produced by a given decoding mechanism (see, Blackwell (1993), Casella and Robert (1995) and Churchill (1995), for example). Denote these copies by $\chi_{(i)} = \chi_{i,1}, \ldots, \chi_{i,q_i}$, $i = 1, 2, \ldots, K$, where q_i indicates the length of the ith copy. Note that due to errors that may occur during the decoding procedure the decoded sequences and the prototype sequence may not have the same length. Different copies may also have different lengths.

The prototype DNA word $\mathbf{X}^{(M)}$ is an unknown vector of bases belonging to $\{\text{A,C,G,T}\}^M$. The observed data is $\mathbf{Y} = (\chi_{(1)}, \ldots, \chi_{(K)})$, i.e., the outcome produced by K independent realisations of the mechanism used to decode the prototype sequence $\mathbf{X}^{(M)}$. Assume that the decoded sequences $\mathbf{Y}$ are results of a hidden Markov chain denoted by $\mathbf{s} = \{s_k,\ k = 0, 1, 2, \ldots\}$ where each s_k is of one of the following type of states: R-states, representing mutations (this can be either a replacement of a base by a different one or a replacement of a base by itself, and therefore a correct copy); D-states, meaning that a deletion occurred; and I-states indicating that an insertion has occurred. Besides the I, R, and D states, two spurious states are added to the state space of $\mathbf{s}$. These states, denoted by B and E, are used to indicate the beginning and the end of the decoding process, respectively, (see Churchill (1995) and Churchill and Lazareva (1999)). The spurious states are mute states, i.e., they do not produce an output. The initial state of the sequence $\mathbf{s}$ is set to be B with probability one. Unless otherwise stated, from now on the procedure will be described for the case $K = 1$.

Each base $\mathbf{X}_t$, $t = 1, 2, \ldots, M$ will be associated to one of the R, I, or D states as follows. $\mathbf{X}_t$ is associated to the states I_t, R_t, D_t in a way that if the value of the hidden Markov chain is R_t, that means that the base $\mathbf{X}_t$ was either correctly copied or was replaced by another base; if its value is I_t, then that means that a base has been inserted before the base $\mathbf{X}_t$ was processed; if its value is D_t then that means that the base $\mathbf{X}_t$ was deleted during the decoding process. An additional state is also considered. The state I_{M+1} will indicate the possible insertions that may occur after the last base of $\mathbf{X}$ is processed. Hence, given M, the state space of the hidden Markov chain is $\mathcal{S} = \{B, R_1, I_1, D_1, \ldots, R_M, I_M, D_M, I_{M+1}, E\}$ where l consecutive visits to state I_t means that l insertions were made before base $\mathbf{X}_t$ was examined, $t = 1, 2, \ldots, M$; and l consecutive visits to state I_{M+1} implies that l insertions were made after the base $\mathbf{X}_M$ was processed. Let $n = \min\{k : s_k = E,\, k \geq 2\} - 1$. Therefore, the hidden Markov chain is given by $\mathbf{s} = s_1, \ldots, s_n$. The decoded sequence $\chi = \chi_1, \ldots, \chi_n$ produced as a realisation of the chain $\mathbf{s}$ has state space $\mathcal{R} = \{A, C, G, T, *, -\}$, where $*$ appears in a specific position to indicate that at that position it was not possible to decide what base was present, and "$-$" indicates that the base in the prototype sequence that would occupy that position was deleted during the decoding process. (The states B and E are suppressed in $\mathbf{s}$ since they produce no output and are used only to indicate the beginning and the end of the decoding process.) Hence, for a given DNA word $\mathbf{X}^{(M)} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ the output data is $\{\chi_{(i)} = \chi_{i,1}, \ldots, \chi_{i,n_i},\, i = 1, 2, \ldots, K\}$ and the corresponding hidden states are $\{s_{(i)} = s_{i,1}, \ldots, s_{i,n_i},\, i = 1, 2, \ldots, K\}$.

For $M = m$ given, let $\Lambda_{(m)} = \left(\lambda_{(m)}(i,j)\right)_{i,j \in \mathcal{S}}$ be the transition matrix for the hidden Markov chain, i.e., $\lambda_{(m)}(i,j) = P(s_k = j \mid s_{k-1} = i)$, $1 \leq k \leq n+1$, $i, j \in \mathcal{S}$; $\lambda_{(m)}(i, B) = \lambda_{(m)}(E, i) = 0$, for all $i \in \mathcal{S}$ and $\lambda_{(m)}(E, E) = 1$. Denote by $\Pi_{(m)} = \left(\pi_{(m)}(i,j)\right)_{i \in \mathcal{S}, j \in \mathcal{R}}$ the distribution of the observed states given the hidden states, i.e., $\pi_{(m)}(i,j) = P(\chi_k = j \mid s_k = i)$, $1 \leq k \leq n$, $i \in \mathcal{S}, j \in \mathcal{R}$. The matrices $\Lambda_{(m)}$ and $\Pi_{(m)}$ are stochastic and are $(3m+2) \times (3m+2)$ and $(3m) \times 6$ matrices, respectively.

Remarks. 1. Since homogeneity of the DNA sequence is assumed, then the reduced state space $\mathcal{S}' = \{B, I, R, D, E\}$ will be used for the hidden Markov chain. Therefore, the reduced transition matrix is considered, i.e., given that $M = m$,

$$\Lambda_{(m)} = \begin{pmatrix} \lambda_{(m)}(R,R) & \lambda_{(m)}(R,D) & \lambda_{(m)}(R,I) \\ \lambda_{(m)}(D,R) & \lambda_{(m)}(D,D) & \lambda_{(m)}(D,I) \\ \lambda_{(m)}(I,R) & \lambda_{(m)}(I,D) & \lambda_{(m)}(I,I) \end{pmatrix}.$$

That implies that the transition from any state I_t is the same. Likewise for the states R_t and D_t.

2. Also due to homogeneity we may work with the reduced observation matrix Π, i. e., given that $M = m$ we have

$$\Pi_{(m)} = \begin{pmatrix} \pi_{(m)}(I,A) & \pi_{(m)}(I,C) & \pi_{(m)}(I,G) & \pi_{(m)}(I,T) & \pi_{(m)}(I,*) & 0 \\ \pi^A_{(m)}(R,A) & \pi^A_{(m)}(R,C) & \pi^A_{(m)}(R,G) & \pi^A_{(m)}(R,T) & \pi^A_{(m)}(R,*) & 0 \\ \pi^C_{(m)}(R,A) & \pi^C_{(m)}(R,C) & \pi^C_{(m)}(R,G) & \pi^C_{(m)}(R,T) & \pi^C_{(m)}(R,*) & 0 \\ \pi^G_{(m)}(R,A) & \pi^G_{(m)}(R,C) & \pi^G_{(m)}(R,G) & \pi^G_{(m)}(R,T) & \pi^G_{(m)}(R,*) & 0 \\ \pi^T_{(m)}(R,A) & \pi^T_{(m)}(R,C) & \pi^T_{(m)}(R,G) & \pi^T_{(m)}(R,T) & \pi^T_{(m)}(R,*) & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

where $\pi^A_{(m)}(R,\cdot)$, $\pi^C_{(m)}(R,\cdot)$, $\pi^G_{(m)}(R,\cdot)$, $\pi^T_{(m)}(R,\cdot)$ indicate the observation probabilities when the R state has occurred in the hidden Markov chain and this state corresponds to the letter A, C, G, and T in the prototype sequence, respectively. The last line of $\Pi_{(m)}$, corresponds to the possible observations from a D state and the last column of $\Pi_{(m)}$ indicates the observation produced by the output "$-$" (which corresponds to a deletion). Therefore, if either I or R is associated to X_t, then the observation of a state "$-$" is not possible and if D is associated to X_t, then the only output allowed is the state "$-$".

3. Note that, when R is associated with the last base of $\mathbf{X}$ and the copied sequence still has some bases to be accounted for, that means that the remaining bases of χ are results of insertions and we will have only transitions from I to I and from I to E. If R is associated with a base in $\mathbf{X}$ (which is not the last one) and in the next step there is no base to be processed in the observed sequence, then that means that the remaining bases of $\mathbf{X}$ have been deleted during the decoding procedure.

4. Note if $\mathbf{X}_t = C$ and if R is associated to $\mathbf{X}_t$, then the most likely output is C. Similar situation occurs if we have $\mathbf{X}_t = A, G, T$ instead of $\mathbf{X}_t = C$.

Next the general Bayesian model is presented.

3 A BAYESIAN MODEL

There is a natural hierarchical structure expressed by modelling the joint distribution of the parameter $\theta = (M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$ and the output $\mathbf{Y}$ as

$$\begin{aligned} P(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)}, \mathbf{Y}) \;\; &\propto \;\; L(\mathbf{Y}|M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})\, P(\Lambda_{(M)}|M)\, P(\mathbf{s}|\Lambda_{(M)}, M) \\ & \qquad P(\Pi_{(M)}|X^{(M)}, M)\, P(\mathbf{X}^{(M)}|M)\, P(M), \end{aligned} \qquad (1)$$

(see, for example, Richardson and Green (1997) and Robert *et al.* (2000)). There are several choices for each component of the above model (1). The ones considered here are:

1. **The prior probability of M.** The length M of the prototype sequence will have a truncated Poisson distribution with parameter μ, i.e.,

   $$P(M) \;\propto\; \frac{\mu^M}{M!}\, I_{\{k_0+1,\ldots,K_0\}}(M),$$

   where $I_A(x) = 1$, if $x \in A$ and it is zero otherwise. Other choices for $P(M)$ are possible. Liu *et al.* (1999) suggested it to be uniform in a suitable range of possible lengths, say $l_0 + 1$ and L_0, and Robert *et al.* (2000) consider $P(M)$ as the uniform distribution on $\{1, 2, \ldots, M_{\max}\}$, where $M_{\max}$ is some given number. The length M may also be considered as a random variable with a geometric distribution with some parameter $0 < p < 1$.

2. **The prior probability of $\mathbf{X}^{(M)}$ given its length M.** Two cases will be considered:

   (a) **Independent case.** Churchill and Lazareva (1999) assume that, given M, the sequence $\mathbf{X}^{(M)}$ has independent and identically distributed components with known letter frequencies α_i, $i \in \{A, C, G, T\}$, that is

   $$P(\mathbf{X}^{(M)}|M) = \prod_{t=1}^{M} \alpha_{X_t}, \quad X_t \in \{A, C, G, T\}. \qquad (2)$$

   (b) **Non-independent case.** In this work the independence assumption is dropped and a spatial dependence among the sites is introduced. The dependence assumed here is that given the sequence length M the configuration of the sequence $\mathbf{X}^{(M)}$ follows a four colour Potts model, that is,

$$P(\mathbf{X}^{(M)}|M) \;=\; \frac{1}{Z_{\beta,M}} \exp\left(\sum_{\{t,r\}} \beta_{tr} \mathbf{1}_{\{\mathbf{X}_t=\mathbf{X}_r\}}(\mathbf{X}^{(M)})\right), \tag{3}$$

where the sum is over all unordered neighbours $\{t, r\}$ (neighbourhood to be defined), with $t, r \in \{1, 2, \ldots, M\}$, and $Z_{\beta,M}$ is a normalising constant. Boundary conditions are free. The k-colour Potts model ($k \geq 1$) has been extensively used in image recovery and recognition. (For more information on Potts model and its applications see, for example, Ferrari *et al.* (1995), Greig *et al.* (1989), Hebert and Leahy (1992), Hurn and Jennison (1993), and Wu (1982), among others.) In the present case the neighbourhood considered is the nearest neighbour and we also assume that $\beta_{tr} = \beta_{rt} = \beta$ for all $t, r \in \{1, 2, \ldots, M\}$. Hence, the expression (3) may be written as

$$P(\mathbf{X}^{(M)}|M) = \frac{1}{Z_{\beta,M}} \exp\left(M\,\beta + 2\,\beta \sum_{t=2}^{M} \mathbf{1}_{\{\mathbf{X}_t=\mathbf{X}_{t-1}\}}(\mathbf{X}^{(M)})\right). \tag{4}$$

It is possible to consider this particular case for β_{tr} because the sequence is considered to be homogeneous. (For heterogeneous DNA sequences one may use the approach proposed by Boys and Henderson (2003) to identify homogeneous segments and then apply the hypothesis of the same value of β_{tr} for each of the segments. Note that in this case we may have distinct values of β for distinct homogeneous segments.)

3. **The prior distribution of the transition matrix $\Lambda_{(M)}$ given M.** It seems reasonable to assume that given M, the error process is conditionally independent of the prototype sequence $\mathbf{X}^{(M)}$. A natural choice for the prior distribution of $\Lambda_{(M)}$ given M, is to assume that every row $\Lambda_{(M)}(k)$, $k = 1, 2, 3$ of the reduced matrix $\Lambda_{(M)}$ is independently distributed according to a suitable Dirichlet distribution on the three dimensional simplex Δ_3. The choice of the parameter of the Dirichlet distribution is arbitrary. Robert *et al.* (2000) suggest using all parameters equal to one, while Churchill and Lazareva

(1999) give some other values for the parameters. Hence, we use the Dirichlet distributions with parameters: (a_{iR}, a_{iD}, a_{iI}), $i = R, D, I$ to sample the first, the second and the third rows of $\Lambda_{(M)}$, respectively. The particular values for the parameters (a_{iR}, a_{iD}, a_{iI}), $i = R, D, I$ of will be given in Section 5.

4. **Error process.** The prior distribution of the error sequence $\mathbf{s}$ given $\Lambda_{(M)}$ and M, follows a Markov chain with transition matrix $\Lambda_{(M)}$ as described in Section 2.

5. **Distribution of $\Pi_{(M)}$ given $\mathbf{X}^{(M)}$ and M.** Assume that the rows of $\Pi_{(M)}$ are independent and that given $\mathbf{X}^{(M)}$ and M, its first row $\Pi_{(M)}(1)$ has the non-zero probabilities sampled from the five dimensional simplex Δ_5 using a Dirichlet distribution with parameter $(a_{IA},\ a_{IC},\ a_{IG},\ a_{IT}, a_{I*})$. Rows $\Pi_M(k)$, $k = 2, 3, 4, 5$, have non-zero probabilities sampled from the simplex Δ_5 using a Dirichlet distribution with parameter $(a_{XA},\ a_{XC},\ a_{XG},\ a_{XT}, a_{X*})$, where the $\mathbf{X}$ is to indicate the dependence on the value that appears in the sequence $\mathbf{X}$ at a given position. So, if we are sampling the values related to the fifth letter of the sequence $\mathbf{X}$ and this is an A (with R the corresponding state of the hidden Markov chain), then we use the probabilities that appear on the second row of $\Pi_{(M)}$ to obtain the respective observation probability. For the last row of $\Pi_{(M)}$ we have that $\Pi_{(M)}(6) = (0, 0, 0, 0, 0, 1)$ with probability one. Note that since the parameter space is enlarged to include the prototype sequence $X^{(M)}$, it is not necessary to deal with a mixture of Dirichlet distributions as in Churchill and Lazareva (1999).

6. **The likelihood function.** Since the output $\mathbf{Y}$ is produced through the hidden Markov chain mechanism, the likelihood of $\mathbf{Y}$ given the parameter $\theta = (M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$ depends only on M, $\Pi_{(M)}$ and $\mathbf{s}$. Therefore,

$$\begin{aligned} L(\mathbf{Y}|M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)}) &= \prod_{i=1}^{K} L(\chi_{(i)}|M, \mathbf{s}^{(i)}, \Pi_{(M)}^{(i)}) \\ &= \prod_{i=1}^{K} \pi_{(M)}^{(i)}(s_{i,1}, \chi_{i,1})\pi_{(M)}^{(i)}(s_{i,2}, \chi_{i,2}) \ldots \pi_{(M)}^{(i)}(s_{i,n_i}, \chi_{i,n_i}) \end{aligned}$$

where $\mathbf{s}^{(i)} = (s_{i,1}, \ldots, s_{i,n_i})$ is the error sequence that produced the observation $\chi_{(i)}$ (recall that $\chi_{(i)}$ has been extended to have the same length as $\mathbf{s}^{(i)}$ by assigning $\chi_{i,j} = -$ whenever $s_{i,j} = D$) and $\Pi^{(i)}_{(M)}$ is the observation matrix associated to the ith observation and the Markov chain that produced it.

The main purpose of this work is to sample values of $(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$ from their joint posterior distribution $P(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)}|\mathbf{Y})$ and use the sample for the Bayesian model selection. This will be done in Section 4, when a reversible jump Markov chain Monte Carlo method (Carlin and Chib (1995) and Green (1995)) is used to obtain a sample from the joint posterior distribution. The joint posterior distribution, in the present case, can be written as

$$\begin{aligned}
P(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)}|\mathbf{Y}) &= P(\mathbf{s} \,|\, \Lambda_{(M)}, M, \mathbf{Y})\, P(\Lambda_{(M)}|M, \mathbf{Y})\, P(\Pi_{(M)}|\mathbf{X}^{(M)}, M, \mathbf{Y}) \\
&\quad P(\mathbf{X}^{(M)}|M, \mathbf{Y})\, P(M|\mathbf{Y}) \\
&\propto L(\mathbf{Y}|M, \mathbf{s}, \Pi_{(M)})\, P(\mathbf{s}|\Lambda_{(M)}, M)\, P(\Lambda_{(M)}|M) \qquad (5) \\
&\quad P(\Pi_{(M)}|\mathbf{X}^{(M)}, M)\, P(\mathbf{X}^{(M)}|M)\, P(M),
\end{aligned}$$

and the main task is to construct a Markov chain whose stationary distribution is $P(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)}|\mathbf{Y})$. This is done in the next section.

4 *MAXIMUM A POSTERIORI* THROUGH REVERSIBLE JUMP MCMC

Maximum a posteriori methods have been widely used for image restoration when the dimension (number of pixels) of the image is known (see the survey paper by Geman (1990) and references therein, and also Ferrari *et al.* (1995), Greig *et al.* (1989), Hebert and Leahy (1992) and Hurn and Jennison (1993), among others). However, the usual methodology is not appropriate when the dimension is unknown or random, as is the case analysed here. Carlin and Chib (1995) and Green (1995) introduced the concept of Bayesian model determination when the dimension of the model is unknown. Note that given M, the vector $(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$ lies in

$$\mathcal{C}_M \;=\; \{M\} \times \{\text{A,C,G,T}\}^M \times (\Delta_3)^3 \times \{I, R, D\}^n \times (\Delta_5 \times \{0\})^5 \times (0,0,0,0,0,1),$$

where $\Delta_5 \times \{0\}$ is used to indicate the set of all six dimensional vectors whose first five coordinates form a vector in the simplex Δ_5 and the sixth coordinate is zero. In general, $(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$ lies in $\mathcal{C} \;=\; \cup_{M=1}^{\infty} \mathcal{C}_M$.

Bayesian inference about M and $\mathbf{X}^{(M)}$ will be done in two steps. A sample $\{(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})_j, j \;=\; 1, 2, \ldots, J\}$ is drawn from the joint posterior distribution using a reversible jump Markov chain Monte Carlo. The model $M = m$ chosen is the one that maximises the marginal posterior $P(M|\mathbf{Y})$ which is estimated by the proportion of time that the parameter $(M,\ \mathbf{X}^{(M)},\ \Lambda_{(M)},\ \mathbf{s}, \Pi_{(M)})$ stays in the dimension $M = m$. Among the sequences that belong to the model in dimension $M = m$, the one chosen to estimate the prototype sequence is the sequence $\mathbf{X}^{(m)}$ that maximises $P(\mathbf{X}^{(m)}|M = m, \mathbf{Y})$.

The reversible jump Markov chain Monte Carlo used to obtain a sample $\{(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})_j, j = 1, 2, \ldots, J\}$ is described as follows. (As usual, there is a certain flexibility in choosing conveniently the kernel of transition.) If the actual state of the chain is $(M, \mathbf{X}^{(M)}, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$, at the time of a transition an independent random choice is made among attempting each of the three moves:

(r) replacement of a base at a randomly chosen site by a randomly chosen base;

(b) birth of a randomly chosen base at a randomly chosen location;

(d) death of a base at a randomly chosen site;

with probabilities r_M, b_M and d_M, respectively, depending only on the dimension of $\mathbf{X}^{(M)}$ and satisfying

$$r_M + b_M + d_M \;=\; 1.$$

These probabilities are chosen so that

$$b_M = c\ \min\Big\{1, \frac{p(M+1)}{p(M)}\Big\}$$

and

$$d_M = c \min\Big\{1, \frac{p(M-1)}{p(M)}\Big\},$$

with $c > 0$ a suitable constant subject to $b_M + d_M < 1$, for all $M \geq 1$. Therefore, the reversibility condition $p(M)\, b_M = p(M+1)\, d_{M+1}$ is satisfied.

Remark. Note that **(b)** and **(d)** involve changing the dimension of the parameter space, hence standard Markov chain Monte Carlo methods do not apply.

In order to describe the steps of the Markov chain Monte Carlo, the dependence on M is dropped from some of the notation. Therefore, from now on Λ is used to represent the transition matrix of the hidden Markov chain that will possibly be updated and Λ' is the updated matrix; Π is the observation matrix that will possibly be updated and Π' is the updated matrix. The proposal distributions for the jumps of the reversible jump Markov chain are given in the following way.

1. **Replacement:** If a replacement move is chosen, then,

   (a) select a position t uniformly in $\{1, 2, \ldots, M\}$;

   (b) replace $\mathbf{X}_t$ by $\mathbf{X}'_t$ chosen uniformly from $\mathcal{A}$ (note that a base can be replaced by itself and that some other distribution may be used to select from $\mathcal{A}$) and let $\mathbf{X}'^{(M)} = (\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, \mathbf{X}'_t, \mathbf{X}_{t+1}, \ldots, \mathbf{X}_M)$ be the updated sequence;

   (c) update the matrix Λ independently sampling its rows using their posterior distribution, i.e., if n^{Λ}_{ij}, $i, j = R, D, I$ count the number of transitions in the hidden Markov chain from the state i to state j, then sample sample rows 1, 2 and 3 of Λ' from a Dirichlet distribution with parameter $(a_{iR} + n^{\Lambda}_{iR}, a_{iD} + n^{\Lambda}_{iD}, a_{iI} + n^{\Lambda}_{iI})$, $i = R, D, I$, respectively.

   (d) Update the error sequence $\mathbf{s}$ using the updated matrix Λ'.

   (e) The updating of the matrix Π is made by independently sampling its rows using the following mechanism. Let R_A, R_C, R_G, R_T mean that R is an output of the hidden Markov chain and that it is associated to the letter A, C, G, T, respectively, in the sequence $\mathbf{X}^{(M)}$, and let n^{Π}_{ij}, $i = I, R_A, R_C, R_G, R_T$, $j = A, C, G, T, *$, count the number of times that the hidden Markov chain produces the output i and the character observed is j. Then, the first five elements of the first five rows of the matrix Π will be sampled from a Dirichlet distribution with parameter $(a_{iA} + n^{\Pi}_{iA}, a_{iC} + n^{\Pi}_{iC}, a_{iG} + n^{\Pi}_{iG}, a_{i*} + n^{\Pi}_{i*})$, $i = I, R_A, R_C, R_G, R_T$, respectively. The last element in each of the first five rows of Π' is zero with probability one. The row corresponding to the observation from a D state is $(0, 0, 0, 0, 0, 1)$ with probability one.

The acceptance probability of this move is

$$\min\left\{1, \frac{L(\mathbf{Y}|M, \Pi', \mathbf{s}')\, P(\mathbf{s}'|\Lambda', M)\, P(\Pi'|\mathbf{X}'^{(M)}, M)\, P(\Lambda'|M)}{L(\mathbf{Y}|M, \Pi, \mathbf{s})\, P(\mathbf{s}|\Lambda, M)\, P(\Pi|\mathbf{X}^{(M)}, M)\, P(\Lambda|M)}\right.$$
$$\left.\frac{P(\mathbf{X}'^{(M)}|M)\, Q(\theta', \theta)}{P(\mathbf{X}^{(M)}|M) Q(\theta, \theta')} J\right\}$$

where $Q(\theta, \theta')$ is the proposal transition from θ to θ' and the ratio $Q(\theta', \theta)/Q(\theta, \theta')$ is given by

$$\frac{f(\Lambda)\, g(\Pi)\, P(\mathbf{s} \mid \Lambda',\, M)}{f(\Lambda')\, g(\Pi')\, P(\mathbf{s}' \mid \Lambda',\, M)},$$

where $f(\cdot)$ and $g(\cdot)$ are the products of Dirichlet distributions used to update of the non-zero probabilities of the rows of Λ and Π, respectively, and $P(\cdot \mid \Lambda',\, M)$ is the probability of the error sequence when the updated matrix Λ' is considered. The transformation from the space where θ belongs to the space to which $\theta' = (M, \mathbf{X}'^{(M)}, \Lambda', \mathbf{s}', \Pi')$ belongs to is given by the transformations considered in items (a), (b), (c), (d) and (e) above and therefore the Jacobian, J, is equal to one. We also have, for the prior distribution (4), that

$$\begin{aligned}
\frac{p(\mathbf{X}'^{(M)}|M)}{p(\mathbf{X}^{(M)}|M)} &= \exp\left(2\beta\left[\mathbf{1}_{\{\mathbf{x}_2=\mathbf{x}'_1\}}(\mathbf{X}'^{(M)}) - \mathbf{1}_{\{\mathbf{x}_2=\mathbf{x}_1\}}(\mathbf{X}^{(M)})\right]\right),\ \text{if } t = 1,\\
\frac{p(\mathbf{X}'^{(M)}|M)}{p(\mathbf{X}^{(M)}|M)} &= \exp\left(2\beta\left[\mathbf{1}_{\{\mathbf{x}'_M=\mathbf{x}_{M-1}\}}(\mathbf{X}'^{(M)}) - \mathbf{1}_{\{\mathbf{x}_M=\mathbf{x}_{M-1}\}}(\mathbf{X}^{(M)})\right]\right),\ \text{if } t = M,\\
\frac{p(\mathbf{X}'^{(M)}|M)}{p(\mathbf{X}^{(M)}|M)} &= \exp\left(2\beta\sum_{k=t}^{t+1}\left[\mathbf{1}_{\{\mathbf{x}'_k=\mathbf{x}'_{k-1}\}}(\mathbf{X}'^{(M)}) - \mathbf{1}_{\{\mathbf{x}_k=\mathbf{x}_{k-1}\}}(\mathbf{X}^{(M)})\right]\right),\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{if } 2 \le t \le M-1.
\end{aligned}$$

2. **Birth:** If a birth move is chosen, then,

   (a) select a position t uniformly in $\{1, \ldots, M+1\}$;

   (b) rename the basis in the following way:

      i. if $t = 1$, then $\mathbf{X}'_1$ is chosen uniformly from $\mathcal{A}$ and $\mathbf{X}'_r = \mathbf{X}_{r-1}$, for $r = 2, 3, \ldots, M+1$;

      ii. if $t = M+1$, then $\mathbf{X}'_{M+1}$ is chosen uniformly from $\mathcal{A}$ and $\mathbf{X}'_r = \mathbf{X}_r$, for $r = 1, 2, \ldots, M$;

      iii. if $2 \leq t \leq M$, then $\mathbf{X}'_r = \mathbf{X}_r$, for $1 \leq r \leq t-1$, $\mathbf{X}'_t$ is chosen uniformly from $\mathcal{A}$ and $\mathbf{X}'_r = \mathbf{X}_{r-1}$ for $t+1 \leq r \leq M+1$,

   and let $\mathbf{X}'^{(M+1)} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_{M+1})$ be the updated sequence.

   (c) The updating of the matrices Λ and Π, and of the error sequence $\mathbf{s}$ is made in the same manner as proposed in the replacement move.

The acceptance probability of this move is

$$\min\left\{1, \frac{L(\mathbf{Y}|M+1, \Pi', \mathbf{s}')\, P(\mathbf{s}'|\Lambda', M+1)\, P(\Pi'|\mathbf{X}'^{(M+1)}, M+1)\, P(\Lambda'|M+1)}{L(\mathbf{Y}|M, \Pi, \mathbf{s})\, P(\mathbf{s}|\Lambda, M)\, P(\Pi|\mathbf{X}^{(M)}, M)\, P(\Lambda|M)} \right.$$
$$\left. \frac{P(\mathbf{X}'^{(M+1)}|M+1)\, Q(\theta', \theta)}{P(\mathbf{X}^{(M)}|M)\, Q(\theta, \theta')} J \right\}$$

where $Q(\theta, \theta')$ is the proposal transition from θ to θ' and the ratio $Q(\theta', \theta)/Q(\theta, \theta')$ is given by

$$\frac{f(\Lambda)\, g(\Pi)\, P(\mathbf{s} \mid \Lambda',\, M)\, M+1}{f(\Lambda')\, g(\Pi')\, P(\mathbf{s}' \mid \Lambda',\, M+1)\, M}.$$

where $f(\cdot)$, $g(\cdot)$ are the products of the Dirichlet distributions used to update the matrices Λ and Π, respectively, and $P(\cdot \mid \Lambda',\, M)$ is the probability of the error sequence when the updated matrix Λ' is considered. The Jacobian, J, of the transformation from the space where θ belongs to the space which $\theta' = (M+1, \mathbf{X}^{(M+1)}, \Lambda', \mathbf{s}', \Pi')$ belongs to is obtained from the transformations given in items (a), (b) and (c) above and therefore, $J = 1$. Furthermore,

$$\frac{P(\mathbf{X}'^{(M+1)}|M+1)}{P(\mathbf{X}^{(M)}|M)} = \frac{Z_{\beta,M}}{Z_{\beta,M+1}}\, e^{\beta}\ \exp\left(2\beta\, \mathbf{1}_{\{\mathbf{x}_1=\mathbf{x}'_1\}}(\mathbf{X}'^{(M+1)})\right),\ \text{if } t=1,$$

$$\frac{P(\mathbf{X}'^{(M+1)}|M+1)}{P(\mathbf{X}^{(M)}|M)} = \frac{Z_{\beta,M}}{Z_{\beta,M+1}}\, e^{\beta}\ \exp\left(2\beta\, \mathbf{1}_{\{\mathbf{x}'_{M+1}=\mathbf{x}_M\}}(\mathbf{X}'^{(M+1)})\right),\ \text{if } t=M+1,$$

$$\frac{P(\mathbf{X}'^{(M+1)}|M+1)}{P(\mathbf{X}^{(M)}|M)} = \frac{Z_{\beta,M}}{Z_{\beta,M+1}}\, e^{\beta}\ \exp\left(2\,\beta \sum_{k=t}^{t+1} \mathbf{1}_{\{\mathbf{x}'_k=\mathbf{x}'_{k-1}\}}(\mathbf{X}'^{(M+1)})\right),\ \text{if } 2\le t\le M,$$

3. **Death:** If a death move is chosen, then,

   (a) select a position t uniformly in $\{1,\ldots,M\}$, delete the base $\mathbf{X}_t$ and rename the remaining basis in the following way: $\mathbf{X}'_r = \mathbf{X}_r$, for $1 \le r \le t-1$, $\mathbf{X}'_r = \mathbf{X}_{r+1}$ for $t \le r \le M-1$.

   (b) The updating of the matrices Λ and Π and the error sequence $\mathbf{s}$ is made using the procedure described when the replacement move is chosen.

The acceptance probability of this move is

$$\min\left\{1,\ \frac{L(\mathbf{Y}|M-1,\Pi',\mathbf{s}')\, P(\mathbf{s}'|\Lambda',M-1)\, P(\Pi'|\mathbf{X}'^{(M-1)},M-1)\, P(\Lambda'|M-1)}{L(\mathbf{Y}|M,\Pi,\mathbf{s})\, P(\mathbf{s}|\Lambda,M)\, P(\Pi|\mathbf{X}^{(M)},M)\, P(\Lambda|M)}\right.$$
$$\left.\frac{P(\mathbf{X}'^{(M-1)}|M-1)\, Q(\theta',\theta)}{P(\mathbf{X}^{(M)}|M)\, Q(\theta,\theta')}\, J\right\}$$

where

$$\frac{P(\mathbf{X}'^{(M-1)}|M-1)}{P(\mathbf{X}^{(M)}|M)} = \frac{Z_{\beta,M}}{Z_{\beta,M-1}}\, e^{-\beta}\ \exp\left(-2\beta\mathbf{1}_{\{\mathbf{x}_2=\mathbf{x}_1\}}(\mathbf{X}^{(M)})\right)\ \text{if } t=1$$

$$\frac{P(\mathbf{X}'^{(M-1)}|M-1)}{P(\mathbf{X}^{(M)}|M)} = \frac{Z_{\beta,M}}{Z_{\beta,M-1}}\, e^{-\beta}\ \exp\left(-2\beta\mathbf{1}_{\{\mathbf{x}_M=\mathbf{x}_{M-1}\}}(\mathbf{X}^{(M)})\right)\ \text{if } t=M$$

$$\frac{P(\mathbf{X}'^{(M-1)}|M-1)}{P(\mathbf{X}^{(M)}|M)} = \frac{Z_{\beta,M}}{Z_{\beta,M-1}}\, e^{-\beta}\ \exp\left(-2\beta\sum_{k=t}^{t+1}\mathbf{1}_{\{\mathbf{x}_k=\mathbf{x}_{k-1}\}}(\mathbf{X}^{(M)})\right)$$
$$\text{if } 2\le t\le M-1,$$

and the ratio $Q(\theta',\theta)/Q(\theta,\theta')$ is given by

$$\frac{f(\Lambda)\, g(\Pi)\, P(\mathbf{s}\,|\,\Lambda',\, M)\, M-1}{f(\Lambda')\, g(\Pi')\, P(\mathbf{s}'\,|\,\Lambda',\, M-1)\, M},$$

where $f(\cdot)$, $g(\cdot)$ and $P(\cdot \mid \Lambda', M-1)$ are given in a similar way as they were given when considering the replacement and birth moves. The Jacobian J of the transformation from the model where θ belongs to the model that $\theta' = (M-1, \mathbf{X}'^{(M-1)}, \Lambda', \mathbf{s}', \Pi')$ belongs to is given by the transformations presented in (a) and (b) above and therefore is equal to one.

The problem now is to run this Markov chain for a sufficiently large number of steps and when the stationary state is attained the actual state of the chain $(\mathbf{X}^{(M)}, M, \Lambda_{(M)}, \mathbf{s}, \Pi_{(M)})$ gives us a sample of the posterior distribution (5) and the usual measures may be used to decide about the Bayesian model. Once the model is chosen, the sequence within this model that has the largest marginal posterior is the one chosen to represent the prototype sequence.

5 SIMULATION

In this section some details about the implementation of the algorithm proposed in this paper are given. Besides the values for the various variables used in the programme, some graphics and comments about the results obtained are presented.

5.1 Setting the parameters

In order to perform the simulation of the algorithm some of the parameters presented in a more general framework during Sections 2, 3 and 4 must be specified. Therefore, the following is considered.

1. The choice of the hyperparameter μ is left to the researcher to choose. One suggestion is to take μ approximately the mean size of the sequences the researcher has as data (i.e., the mean length of K decoded sequences that the researcher is using as observed data). In the present case we take $\mu = 53$. Several values of c (appearing in the reversible jump Monte Carlo) are taken in order to compare the performance of the algorithm depending on what prior distribution of the sequence $\mathbf{X}$ and the length M were considered. The

values used in the simulations were basically $c = 0.001, 0.01, 0.1, 0.33333, 0.35$. The first three values were chosen to make birth and death moves with low probability and the other two values were chosen to have c as close as possible to the maximum value of c such that $d_M + b_M < 1$.

2. The prior distribution for the sequence length M is a truncated Poisson distribution on $\{49, 50, \ldots, 57\}$ whose mean is 53. This distribution will be referred to as the Poisson(53, $\{49, 50, \ldots, 57\}$). Besides this prior two others are taken into account. They are the truncated Poisson(53) with mean 53, taking values on $\{1, 2, \ldots\}$ and the Uniform distribution on $\{49, 50, \ldots, 57\}$ which will be referred to as Poisson(53, $\{1, 2, \ldots\}$) and Uniform$\{49, 50, \ldots, 57\}$, respectively.

3. The parameters of the Dirichlet distribution used to sample values for the rows of Λ are given by $(a_{iI}, a_{iR}, a_{iD}) = (1, 6, 1)$, $i = R, D, I$. The parameters of the Dirichlet distribution used to sample values for the rows of the matrix Π are given by, $(a_{\mathbf{X}A}, a_{\mathbf{X}C}, a_{\mathbf{X}G}, a_{\mathbf{X}T}, a_{\mathbf{X}*}) = (6, 1, 1, 1, 1)$ if $\mathbf{X} = A$ and for $\mathbf{X} = C, G, T$, similar sets of parameters are considered (i.e., if $\mathbf{X} = C$ then $(a_{\mathbf{X}A},\ a_{\mathbf{X}C}, a_{\mathbf{X}G}, a_{\mathbf{X}T}, a_{\mathbf{X}*}) = (1, 6, 1, 1, 1)$). Additionally, the Dirichlet distribution used to sample the row corresponding to the insertion state has parameter $(a_{IA},\ a_{IC},\ a_{IG},\ a_{IT},\ a_{I*}) = (1, 1, 1, 1, 1)$.

4. When using the assumption of independence for the bases in the prototype sequence (i.e., the prior distribution of $\mathbf{X}$ is given by (2)), from Churchill and Lazareva (1999) we use the values $\alpha_i = 0.25$ for all $i \in \{A, C, G, T\}$ and when the Potts distribution (given by (4)) is considered, we take $\beta = 1$.

5. The sequence $\mathbf{X}$ used to initialise the algorithm was the sequence
TAGACAGGGGCCCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAACTT
which is the sequence 1 given, as observed data, by Churchill and Lazareva (1999) where we write the letter G in place of the unknown letter in position 9. Only one sequence was used as observed data. This sequence is the first sequence considered by Churchill

and Lazareva (1999) as observed data, and was obtained from Seto *et. al.* (1993) and it is TAGACAGG*GCCCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAACTT.

5.2 Computational details

The programming of the algorithm was made in FORTRAN and implemented on a IBM SP3 at the Instituto de Matemáticas, UNAM. In order to compare the performance of the algorithm, several runs were made. A typical run of two million steps takes around five minutes of CPU time which, in the tests performed, corresponds to ten minutes in real time. When a Silicon Graphics machine is used the time may increase to forty five minutes in real time.

The first step taken was to diagnose the speed of convergence when taking as a measure of reference the convergence of the sample mean of the sequence length M. Before considering the several possibilities for the prior distributions of M and $\mathbf{X}$ a preliminary test was made. Figure 1 shows the plots obtained. In that case, we assume that M and $\mathbf{X}$ have as prior distributions a Poisson$(54, \{1, 2, \ldots\})$ and the Potts distribution (4) with $\beta = 1$, respectively. The values of c considered were $0.1, 0.01, 0.001, 0.0065$ and 0.0085 and initially runs of 10^6 steps were performed. It is possible to observe that by step 4×10^5 stationarity is achieved when $c = 0.1, 0.01, 0.0065$. When using $c = 0.001$ and $c = 0.0085$ it is possible to observe that apparently, stationarity is achieved at steps 6×10^5 and 5×10^5, respectively. However, since the ergodic means where different of the one obtained when considering the other values of c, we have decided to run the algorithm for $c = 0.001$ and $c = 0.0085$ again, but now an additional 10^6 steps were performed. What is observable from Figure 1 is that for $c = 0.001$ stationarity is achieved when we reach step 14×10^5 and for $c = 0.0085$ we have that stationarity is achieved when step 5×10^5 is reached. It is worth calling attention to the fact that when $c = 0.0085$ in the first run (when only 10^6 steps were performed) the sample mean stays more or less the same for around 4×10^5 steps, giving the idea that stationarity had already been reached, and then starts increasing its value.

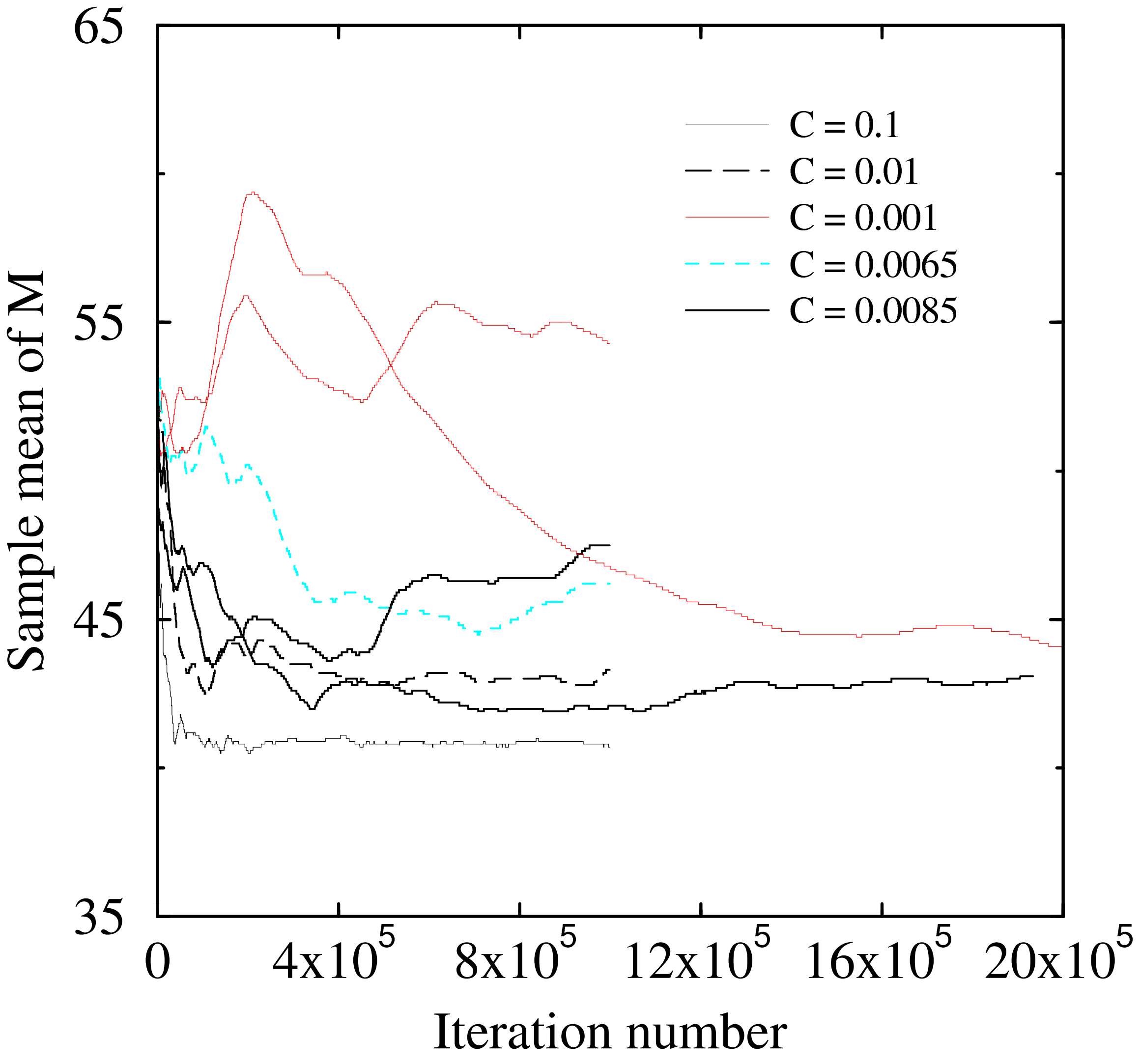

Figure 1: Initial convergence diagnose for the sample mean of M when M has as prior a Poisson(54, $\{1, 2, \ldots\}$) distribution and $\mathbf{X}$ has prior distribution the Potts distribution with $\beta = 1$.

After the preliminary convergence diagnose was made, we have decided to perform 2×10^6 iterations of the algorithm in order to verify the convergence of the empirical mean of M in each of the cases considered. Once convergence had been attained, a sample of size 6×10^4 was taken, using every fifth value generated, to perform inferences. Inferences were performed for the choice of model made by the reversible jump Markov chain Monte Carlo algorithm, for the behaviour of the probabilities b_M, r_M and d_M, as well as, for the sample of M. From now on, the possible values of c used are the ones given in Section 5.1.

Let the DNA sequence have as prior distribution the Potts distribution with parameter $\beta = 1$. When M has prior distribution the Poisson(53,$\{1, 2, \ldots\}$), then it is possible to see from Figure 2(a) that for $c = 0.01$ two million steps were not enough to make the sample mean of M to converge, whereas for all other values around 5×10^5 iterations were enough. In particular, for $c = 0.33333$ and $c = 0.35$, 1×10^5 iterations are enough to achieve convergence and for $c = 0.1$ convergence is reached around iteration 4×10^5. Having performed this test we have decided not to consider the case in which $c = 0.01$. If M has as prior distribution the Poisson(53, $\{49, 50, \ldots, 57\}$), we have by observing Figure 2(b), that for all values of c convergence is attained around iteration number 2×10^5. In the case where M has as prior distribution the Uniform$\{49, 50, \ldots, 57\}$, from Figure 2(c) we have that convergence is attained around iteration number 60×10^3 except for $c = 0.01$. Therefore, using the same procedure as for the Poisson(53, $\{1, 2, \ldots\}$) prior distribution the statistics were made using the remaining values of c. When the bases forming $\mathbf{X}$ are independent and identically distributed and M has as prior distribution the Poisson(53, $\{49, 50, \ldots, 57\}$) and the Uniform$\{49, 50, \ldots, 57\}$ the behaviour of the sample mean is similar to those shown in Figure 2(b) and 2(c) with convergence attained around iteration number 2×10^5 and 3×10^5, respectively. The case where M has as prior distribution the Poisson(53, $\{1, 2, \ldots\}$) is not considered here because of the slowness of the convergence.

Figure 3 illustrates the histograms of the values of M that are accepted by the reversible jump Markov chain Monte Carlo. Figure 3(a), 3(b) and 3(c) represent the case where $\mathbf{X}$ has prior distribution the Potts distribution and M has as prior distribution the Poisson(53, $\{1, 2, \ldots\}$), Poisson(53, $\{49, 50, \ldots, 57\}$) and the Uniform$\{49, 50, \ldots, 57\}$, respectively. Figure 3(d) is a typical example for the case where $\mathbf{X}$ is formed by independent and identically distributed bases and M has prior distribution the Poisson(53, $\{49, 50, \ldots, 57\}$). The value of c used to produce the graphics is $c = 0.35$.

The relationship among the probabilities of the occurrence of the events of birth, death and replacement in the reversible jump Markov chain Monte Carlo algorithm is shown in

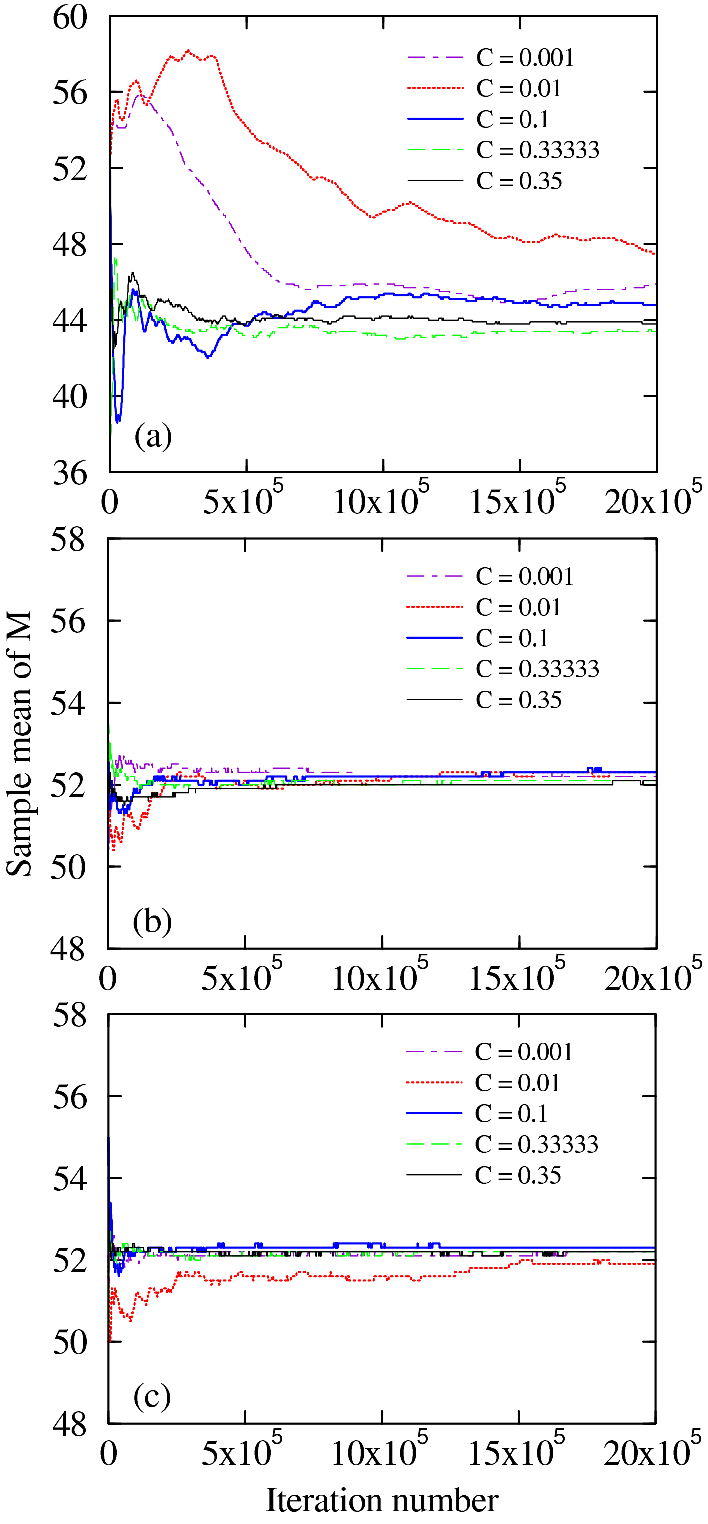

Figure 2: Convergence diagnose, for different values of c, of the sample mean of the sequence length M for the various prior distributions of $\mathbf{X}$ and M.

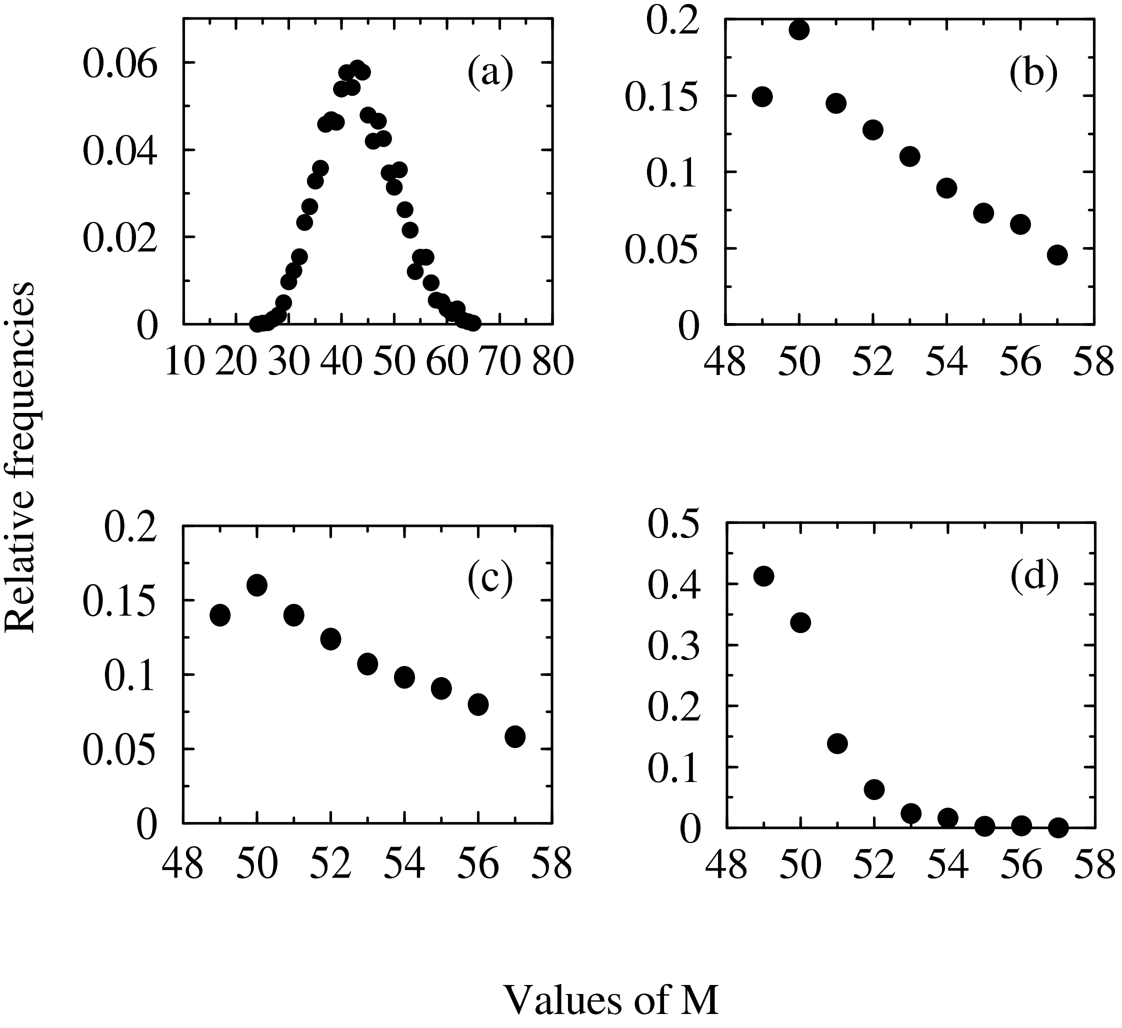

Figure 3: Histograms of values of M produced by the reversible jump Markov chain Monte Carlo algorithm using $c = 0.35$ when different prior distributions for $\mathbf{X}$ and M are considered.

the plots of Figure 4, where a portion of the values produced by the algorithm for the case where $c = 0.35$ was used. The lighter solid line on these plots is the probability of occurrence of a death event (d_M); the dashed line is the probability of a birth event (b_M); and the remaining line corresponds to the probability of a replacement (r_M). Figures 4(a) and 4(b) represent the case where $\mathbf{X}$ has prior distribution the Potts distribution and M has the Poisson(53, $\{1, 2, \ldots\}$) and the Poisson(53, $\{49, 50, \ldots, 57\}$), respectively, as prior distribution. In Figure 4(c) we have that $\mathbf{X}$ has independent and identically distributed bases and M has the Poisson(53, $\{49, 50, \ldots, 57\}$) distribution as its prior. Note that when comparing the behaviour of d_M, r_M, and b_M in the cases of Figure 4(a) and 4(b) the algorithm produces a more mixed behaviour in the case where $\mathbf{X}$ has as prior the Potts distribution and M has the Poisson(53, $\{49, 50, \ldots, 57\}$), than in the case where M has the Poisson(53, $\{1, 2, \ldots\}$) as its prior distribution. The case where M has prior distribution the Uniform$\{49, 50, \ldots, 57\}$ distribution is not presented here because the probabilities d_M, r_M, and b_M are constant.

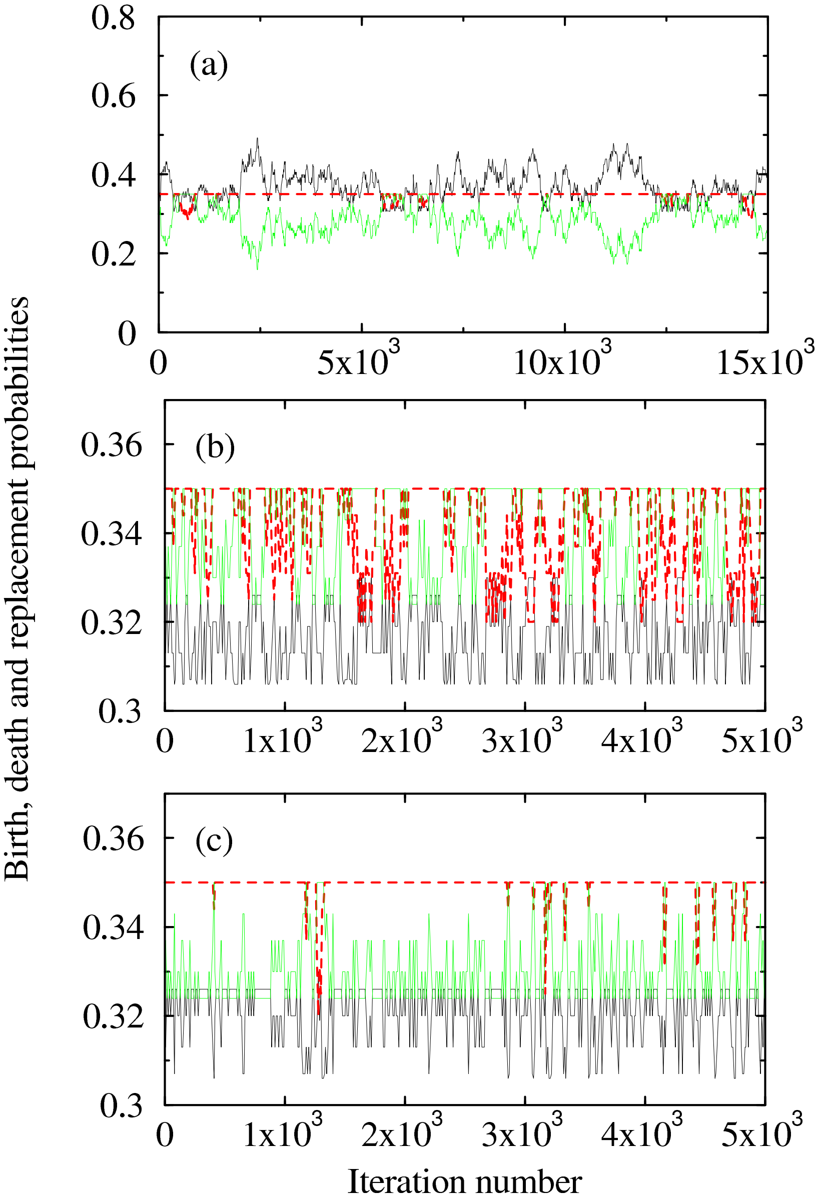

Figure 4: Probabilities of occurrence of birth, death and replacement events using $c = 0.35$ for a certain number of steps after the burn-in period for the several prior distributions of $\mathbf{X}$ and M.

Even though the convergence of the sample mean of M when $\mathbf{X}$ has independent and identically distributed bases is fast, the value of M that has the largest probability is not 52 (which corresponds to the length of the sampled DNA analysed). Instead the one with the largest probability is 49 with over than four times more weight than the one given to $M = 52$ (see Figure 3(d)). Note that when $\mathbf{X}$ has as prior distribution the Potts distribution, then the fastest convergence of the sample mean of M is when M has prior distribution the Uniform$\{49, 50, \ldots, 57\}$, second fastest is when the prior is the Poisson(53, $\{49, 50, \ldots, 57\}$) and the slowest is when M has as prior distribution the Poisson(53, $\{1, 2, \ldots\}$) (Figure 2). We also have that in the latter case the mode of the marginal posterior is around 44 which is way off the length of the sampled DNA sequence.

6 CONCLUSION

In this work our attention was focused on the presentation of an algorithm to find a Bayes estimate of a prototype DNA sequence, its length and the alignment of the copies of this prototype sequence. The novelties of our approach are: in the Bayes method we use a prior distribution for the DNA sequence that incorporates the spatial correlation among the bases given by a four colour Potts model; the length of the prototype sequence (and the length of the alignment) is considered a random variable; a sample from the joint posterior distribution is obtained from a Monte Carlo procedure based on a Markov chain with reversible jumps; during the generation of a sample from the joint distribution of the prototype sequence and its length, the length of the sequence is allowed to change.

ACKNOWLEDGEMENTS

This project started during the summer of 1998 when N.L.G. and E.R.R. were attending the Probability Intern Program at the former Center for the Mathematical Sciences (CMS), University of Wisconsin-Madison and they thank CMS for their hospitality. N.L.G. was partially funded by CNPq, Grant 301054/93-2 and FINEP (núcleo de Excelência “Fenômenos Críticos em Probabilidade e Processos Estocásticos", PRONEX-177/96). E.R.R. thanks IME-USP and IMECC-UNICAMP for financial support and hospitality.

REFERENCES

1. Apostolico, A. and Giancarlo, R. (1999). Sequence alignment in molecular biology. *In: Mathematical Support for Molecular Biology. DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences* **47**, eds., M. Farach-Colton, F. S. Roberts, M. Vingron, M. Waterman, 85–115.

2. Bishop, M. J. and Thompson, E. A. (1986). Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology*, **190**, 159–165.

3. Blackwell, T. (1993). Estimating consensus DNA sequences. Preprint.

4. Boys, R. J. and Henderson, D. A. (2003). Bayesian approach to DNA sequence segmentation. To appear in Biometrics.

5. Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*. SER. B, **57**, 473–484.

6. Casella, G. and Robert, C. (1995). Discussion: "Accurate restoration of DNA sequences" by G. A. Churchill. *In: Case studies in Bayesian Statistics*, Vol. **II**, eds., C. Gatsaris, J. S. Hodges, R. E. Kass, N. D. Sigpur-Walla, Springer-Verlag, New York, 126-138.

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

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

9. Churchill, G. A. (1995). Accurate restoration of DNA sequences (with discussion). *In: Case studies in Bayesian Statistics*, Vol. **II**, eds., C. Gatsaris, J. S. Hodges, R. E. Kass, N. D. Sigpur-Walla, Springer-Verlag, New York, 89-148.

10. Churchill, G. A. and Lazareva, B. (1999). Bayesian restoration of a hidden Markov chain with applications to DNA sequencing. *Journal of Computational Biology*, **6**, 261–277.

11. Drasdo, D., Hwa, T. and Lässig, M. (1998). A statistical theory of sequence alignment with gap. *In: Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, 52–58.

12. Ferrari, P. A., Frigessi, A. and Gonzaga de Sá, P. G. (1995). Fast approximate maximum *a posteriori* restoration of multicolour images. *Journal of the Royal Statististical Society.* SER. B, **57**, 485–500.

13. Garcia, N. L. and Rodrigues, E. R. (1999). Bayesian inference for consensus DNA sequence using reversible jump MCMC. *Publicación Preliminar* **646**. Instituto de Matemá ticas – Universidad Nacional Autónoma de México.

14. Garcia, N. L. and Rodrigues, E. R. (2001). Restoring DNA sequences using a hidden Markov model and reversible jump Markov Chain Monte Carlo. *Publicación Preliminar* **698**. Instituto de Matemáticas – Universidad Nacional Autónoma de México.

15. Geman, D. (1990). Random fields and inverse problems in imaging. *Lecture Notes in Mathematics*, **1427**, Springer-Verlag, 113–193.

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

17. Greig, D. M., Porteous, B. T. and Seheult, A. H. (1989). Exact maximum *a posteriori* estimation for binary images. *Journal of the Royal Statististical Society.* SER. B, **51**, 271–279.

18. Hebert, T. J. and Leahy, R. (1992). Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. *IEEE Transactions on Signal Processing*, **40**, 2290–2303.

19. Hurn, M. and Jennison, C. (1993). Multiple-site updates in maximum *a posteriori* (MAP) and marginal posterior modes (MPM) image estimation. *In: Statistics and Images: 1*, eds., K. V. Mardia, G. K. Kanji. Oxford, Carfax Publishing Company, 155–186.

20. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). Protein modelling using hidden Markov models. *Journal of Molecular Biology*, **235**, 1501–1531.

21. Liu, J. S. and Lawrence, C. E. (1995). Statistical models for multiple sequence alignment: unifications and generalizations. *Proceedings of the American Statistical Association: Statistical Computing Section*, 1–8.

22. Liu, J. S., Neuwald, A. F. and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and the Gibbs sampling strategies. *Journal of the American Statistical Association*, **90**, 1156–1170.

23. Liu, J. S., Neuwald, A. F. and Lawrence, C. E. (1999). Markovian structures in biological sequence alignments. *Journal of the American Statistical Association*, **94**, 1–15.

24. Meidanis, J. and Setubal, J. C. (1995). Multiple alignment of biological sequences with gap flexibility. *Lecture Notes in Computer Sciences*, **911**, Springer-Verlag, 411–426.

25. Milanesi, L., Marselli, M., Mauri, G., Rolfi, C. and Uboldi, L. (1999). Fragment assembly system for DNA sequencing projects. *In: Mathematical Support for Molecular Biology. DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences* **47**, eds., M. Farach-Colton, F. S. Roberts, M. Vingron, M. Waterman, 241–258.

26. 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.* SER. B, **59**, 731–792.

27. 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.* SER. B, **62**, 57–75.

28. Schleif, R. (1993). *Genetics and molecular biology.* Second edition. The Johns Hopkins University Press. USA.

29. Seto, D., Koop, B. F. and Hood, L. (1993). An experimentally derived data set constructed for testing large-scale DNA sequence assembly algorithms. *Genomics*, **15**, 673–676.

30. Thorne, J. L. and Churchill, G. A. (1995). Estimation and reliability of molecular sequence alignments. *Biometrics*, **51**, 100–113.

31. Thorne, J. L., Kishino, H. and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, **33**, 114–124.

32. Waterman, M. S. (1989a). Sequence alignment. *Mathematical Methods for DNA sequences*, ed., M. S. Waterman, CRC Press, USA, 53–92.

33. Waterman, M. S. (1989b). Consensus patterns in sequences. *Mathematical Methods for DNA sequences*, ed., M. S. Waterman, CRC Press, USA, 93–157.

34. Weir, B. S. (1985). Statistical Analysis of Molecular Genetic Data. *Journal of Mathematics Applied in Medicine and Biology*, **2**, 1–39.

35. Wu, F. Y. (1982). The Potts model. *Reviews of Modern Physics*, **54**, 235–268.