

Details of the MCMC scheme used in “Segmenting bacterial and viral DNA sequence alignments using a FHMM”

Wolfgang Lehrach, Dirk Husmeier

15th July 2008

1 Outline of the MCMC scheme

The following moves are performed for each iteration of the MCMC sampler:

1. Sample $\mathbf{H}_A \sim P(\cdot | \nu_A, \boldsymbol{\rho}_A, k_A, \mathbf{H}_{\{S,R,T\} \setminus A}, \mathcal{D})$ for $A = \{S, R, T\}$.
2. Sample $\nu_A \sim P(\cdot | C_A^{\min}, C_A^{\max}, k_A, \mathbf{H}_A, \mathcal{D})$ for $A = \{S, R, T\}$.
3. For $A = \{R, T\}$: propose $\boldsymbol{\rho}_A^*$ and k_A^* by adapting $\boldsymbol{\rho}_A$ and k_A . Propose $\mathbf{H}_A^* \sim P(\cdot | \nu_A, \boldsymbol{\rho}_A^*, k_A^*, \mathbf{H}_{\{S,R,T\} \setminus A}, \mathcal{D})$. Accept $\boldsymbol{\rho}_A^*, k_A^*$ and \mathbf{H}_A^* if $\mathcal{U}[0, 1] <$ acceptance probability, where $\mathcal{U}[0, 1]$ is a sample from the uniform distribution over the unit interval.

Note that the conditioning part of each distribution contains the Markov blanket (Pearl, 1988) of the respective random variable to be sampled. The Markov blanket is the set of parents, co-parents and children of a node. This set shields off a given node from all the other nodes in the domain, that is, conditional on its Markov blanket, a node is independent of all the other nodes. Hence, conditioning on the Markov blanket is equivalent to conditioning on the complete set of random variables (excluding the variable to be sampled). The Markov blanket of each random variable can easily be read off from Figure 1b from the paper: B is in A’s Markov blanket if and only if there is either an edge between A and B, or both A and B are parents of another random variable (Pearl, 1988). This proves that the proposed scheme is a valid Gibbs sampling scheme.

1.1 Sampling $\mathbf{H}_A \sim P(\cdot | \nu_A, \boldsymbol{\rho}_A, k_A, \mathbf{H}_{\{S,R,T\} \setminus A}, \mathcal{D})$

Sampling the hidden state sequences \mathbf{H}_S , \mathbf{H}_R , and \mathbf{H}_T can be effected with a Gibbs-within-Gibbs procedure, as described in Husmeier and McGuire (2003). However, the stochastic forward-backward algorithm of Boys et al. (2000) has proven to lead to faster mixing and convergence of the Markov chain (Werhli et al., 2006) and was, thus, used in the simulations reported in this paper.

1.2 Sampling $\nu_A \sim P(\cdot | C_A^{\min}, C_A^{\max}, k_A, \rho_A, \mathbf{H}_A, \mathcal{D})$

The sampling steps for ν_S , ν_R , and ν_T are straightforward due to the conjugacy of the beta distribution \mathcal{B} , as defined in Equation (8) in the paper. Define:

$$\Psi_A = \sum_{t=1}^{N-1} \mathbb{I}(H_{A,t} = H_{A,t+1}), \quad \bar{\Psi}_A = N - 1 - \Psi_A. \quad (1)$$

It is then easy to show from (6) in the paper that:

$$P(\nu_A | C_A^{\min}, C_A^{\max}, \mathbf{H}_A, k_A, \mathcal{D}) \propto \mathbb{I}(C_A^{\min} \leq \nu_A \leq C_A^{\max}) \mathcal{B}(\nu_A | \Psi_A + \alpha, \bar{\Psi}_A + \beta) \quad (2)$$

See Husmeier and McGuire (2003) for a derivation for the untruncated case. If $k_A = 1$, then Equation (2) does not apply, as there is only a single possible \mathbf{H}_A . To generate a sample from the truncated beta distribution we use a simple Metropolis-Hastings method – see Lehrach (2007) for more detail.

Additionally, Ψ_A and $\bar{\Psi}_A$ allow us to generate an estimate of the posterior distribution of ν_A :

$$P(\nu_A | \mathcal{D}, C_A^{\min}, C_A^{\max}) \approx \frac{1}{M} \sum_{i=1}^M \frac{\mathcal{B}(\nu_A | \Psi_A^{(i)} + \alpha, \bar{\Psi}_A^{(i)} + \beta)}{\int_{C_A^{\min}}^{C_A^{\max}} \mathcal{B}(\nu_A | \Psi_A^{(i)} + \alpha, \bar{\Psi}_A^{(i)} + \beta) d\nu_A} \mathbb{I}(C_A^{\min} \leq \nu_A \leq C_A^{\max}), \quad (3)$$

where the superscript (i) represents the i^{th} sample and we have M samples. The integral is easily calculated using the trapezoid method.

1.3 Proposing and conditionally accepting ρ_A^* , k_A^* , and \mathbf{H}_A^*

We adopt a Reversible Jump Metropolis-Hastings scheme (Green, 1995) where we propose a new number of rate states k_R^* and a new set of rate states ρ_R^* from k_R and ρ_R . This is done using a birth move (with probability b_k), a death move (with probability d_k) or a relocation of one of the rate states (with probability r_k). A new \mathbf{H}_R^* is then proposed given the new ρ_R^* . The new set of rate states ρ_R^* is then accepted with a probability such that given ergodicity, the Markov chain is guaranteed to converge in distribution to the correct posterior distribution. This procedure is similar to the reversible jump move (b) from Boys and Henderson (2004).

ρ_T and k_T are adapted in the same way with identical derivations, so we only show the derivation for ρ_R and drop the R subscript on k_R . To use the Reversible Jump method, we need to specify how we propose k^* and ρ_R^* . The set of all possible proposal moves is outlined in Table 1. Note that k^* is proposed such that the Hastings factor cancels out against the prior ratio. Lastly, we propose $\mathbf{H}_R^* \sim P(\cdot | \nu_R, \rho_R^*, k^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D})$ as described in Section 1.1.

The acceptance probability a of k^* , ρ_R^* , and \mathbf{H}_R^* is $\min\{1, A_B\}$, where:

$$A_B = \text{Likelihood ratio} \times \text{Prior ratio} \times \text{Inverse proposal probability ratio} \times |\det(\text{Jacobian})|, \quad (4)$$

see Green (1995) – our formulation is closer to that of Suchard et al. (2003). We first derive the acceptance probability of a birth move. We first propose

Table 1: Possible proposal moves, the probability with which they are selected, and the corresponding proposal probability $\pi_M(\boldsymbol{\rho}_R^*|\boldsymbol{\rho}_R)$ for $\boldsymbol{\rho}_R^*$. All π distributions presume that $\boldsymbol{\rho}_R^*$ is a valid proposal given the move type, as otherwise the π distributions are not normalised. We use $c = 0.4$ – see Green (1995).

Move type	Probability of move and proposal for $\boldsymbol{\rho}_R^*$	Description of how $\boldsymbol{\rho}_R^*$ is proposed
Birth $k^* = k + 1$	$b_k = c \min \left\{ 1, \frac{P(k+1)}{P(k)} \right\}$ $\pi_b(\boldsymbol{\rho}_R^* \boldsymbol{\rho}_R) = \frac{1}{k+1} Q(\boldsymbol{\rho}_R^*)$	A new rate is sampled from Q in Equation (??), the prior distribution on $\boldsymbol{\rho}_R$ for a single rate. Where to insert the new rate state is randomly and uniformly sampled from the $k + 1$ possibilities.
Death $k^* = k - 1$	$d_k = c \min \left\{ 1, \frac{P(k-1)}{P(k)} \right\}$ $\pi_d(\boldsymbol{\rho}_R^* \boldsymbol{\rho}_R) = \frac{1}{k}$	A randomly chosen rate is deleted.
Relocation $k^* = k$	$r_k = 1 - (b_k + d_k)$ $\pi_r(\boldsymbol{\rho}_R^* \boldsymbol{\rho}_R) = \frac{1}{k} Q(\boldsymbol{\rho}_R^*)$	An existing rate factor position is randomly chosen, and its position re-sampled from Q (see birth move).

a new rate state ρ_R^* from Q_R in Equation (11) in the paper. We then map $(\boldsymbol{\rho}_R, \rho_R^*)$ to $(\boldsymbol{\rho}_R^*)$. In Equation (4), the Jacobian term refers to this mapping, and \det stands for the determinant. This mapping is a permutation, hence the Jacobian is a permutation matrix, which implies $\det(\text{Jacobian}) = \pm 1$, so $|\det(\text{Jacobian})| = 1$.

From Equations (11), (12), and (13), all from the paper, and Table 1 we see that after cancelling, the terms (by their initials) are:

$$\begin{aligned}
 \text{LR} &= \frac{P(\mathcal{D}|\mathbf{H}_R, k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T)}{P(\mathcal{D}|\mathbf{H}_R^*, k + 1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T)} \\
 \text{PR} &= \frac{P(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R)}{P(\mathbf{H}_R^*|k, \nu_R, \boldsymbol{\rho}_R)} \frac{P(k+1)}{P(k)} \frac{P(\boldsymbol{\rho}_R^*|k+1)}{P(\boldsymbol{\rho}_R|k)} \\
 &= \frac{P(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R)}{P(\mathbf{H}_R^*|k, \nu_R, \boldsymbol{\rho}_R)} \frac{P(k+1)}{P(k)} [Q(\boldsymbol{\rho}_R^*)] \\
 \text{IPPR} &= \frac{P(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D})}{P(\mathbf{H}_R^*|k+1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D})} \frac{d_{k+1}\pi_d(\boldsymbol{\rho}_R|\boldsymbol{\rho}_R^*)}{b_k\pi_b(\boldsymbol{\rho}_R^*|\boldsymbol{\rho}_R)} \\
 &= \frac{P(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D})}{P(\mathbf{H}_R^*|k+1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D})} \frac{P(k)}{P(k+1)} \frac{k+1}{k+1} \frac{1}{Q(\boldsymbol{\rho}_R^*)}.
 \end{aligned}$$

All terms not involving \mathcal{D} and \mathbf{H}_R cancel between PR and IPPR. LR and PR together form a ratio of joint distributions over \mathcal{D} and \mathbf{H}_R which in turn simplifies against the ratio of distributions over \mathbf{H}_R conditioned on \mathcal{D} in IPPR. Hence:

$$A_B = \frac{P(\mathcal{D}|k+1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T)}{P(\mathcal{D}|k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T)}, \quad (5)$$

where due to the HMM structure, $P(\mathcal{D}|\boldsymbol{\rho}_R^*, k+1, \mathbf{H}_S, \mathbf{H}_T, \nu_R)$ can be computed from Equations (12) and (13) in the paper in linear time with a dynamical programming algorithm known as the forward algorithm (Rabiner, 1989). Note that the stated dependence on the conditioning variables becomes clear from the conditional independence graph of Figure 1b in the paper and the properties of the Markov blanket, as discussed above.

The same cancellations and simplifications occur when considering the acceptance probability of the death move as the death move is the inverse of the birth move. Hence the acceptance probability of a death move is the same (after replacing $k^* = k+1$ with $k^* = k-1$). The acceptance probability of a relocation is also the same (after replacing $k^* = k+1$ with $k^* = k$) as relocation moves are symmetrical to themselves, in the same way as the birth and death moves are symmetrical.

Note that Equation (2) does not apply when $k=1$ as there is only a single possible \mathbf{H}_R . Hence ν_R has no effect on the likelihood. When moving from two rates to a single rate state, ν_R is removed from the system. Correspondingly, when moving from a single rate state to two rate states, ν_R is proposed from the prior. To see that this leaves the acceptance ratios unchanged, first consider the death move from two rate states to a single rate. We have an extra $P(\nu_R|C_R^{\min}, C_R^{\max})$ in the denominator of PR, and a new proposal term $Q(\nu_R)$ in the numerator of the IPPR. We set $Q(\nu_R) = P(\nu_R|C_R^{\min}, C_R^{\max})$ so that these terms cancel, leaving the acceptance probability unchanged. The reverse argument applies to the birth move, so the acceptance probability is again unchanged.

1.4 Specific Markov chain settings and convergence diagnostics

To check for convergence, we used the method of Gelman and Rubin (1992) and computed the Potential Scale Reduction Factors (PSRF) of $H_{R,t}$ and $H_{T,t}$ for $t \in \{1, \dots, N\}$, and ν_A . These characteristics were chosen as they are invariant to the dimensionality of the parameter space. All results presented in the paper, for all models, were run at least in triplicate (with the exception of the initial ν explorations and the synthetic codon effect study, which were repeated 10 times). For our proposed model, the initial number of rates was picked uniformly between 1 and k_{\max} , with each rate sampled randomly from the uniform distribution. In this paper, we are mainly interested in investigating the rate along the alignment, so all runs were started with only a single transition-transversion ratio, randomly sampled from the uniform distribution.

We discarded the first 10,000 iterations of the PRJ-FHMM samples as the burn-in period. Then, for the next 200,000 iterations every 10th sample was kept so that we could form the posterior summaries. The MCP of Minin et al. (2005) was run for 200,000 burn-in iterations, followed by 4,000,000 sampling iterations where every 200th sample was kept. These lengths were chosen as they resulted in similar convergence indications, as measured by the PSRF. The

high PSRF was consistently less than 1.08 (and often much less), indicating a sufficient degree of convergence. The proposed sampling scheme was extensively tested on synthetic data and its results compared to using numerical integration on simple cases. In order to keep this article concise, these results are presented elsewhere – see Lehrach (2007).

References

- Boys, R. J. and D. A. Henderson (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics* 60, 573–588.
- Boys, R. J., D. A. Henderson, and D. J. Wilkinson (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics* 49, 269–285.
- Gelman, A. and D. B. Rubin (1992, November). Inference from iterative simulation using multiple sequences. *Statistical science* 7(4), 457–472.
- Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Husmeier, D. and G. McGuire (2003). Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution* 20(3), 315–337.
- Lehrach, W. P. (2007). *Predicting protein-protein interactions and characterizing rate heterogeneity along DNA sequence alignments*. Ph. D. thesis, University of Edinburgh, School of Informatics.
- Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* 21(13), 3034–3042.
- Pearl, J. (1988, September). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2), 257–286.
- Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer (2003). Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of the American Statistical Association* 98(462), 427–437.
- Werhli, A., M. Grzegorzcyk, M. Chiang, and D. Husmeier (2006). *Statistics in Genomics and Proteomics*, Chapter Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments, pp. 23–34. Centro Internacional de Matematica.