Distinguishing regional from within-codon rate heterogeneity in DNA sequence alignments

Alexander V. Mantzaris and Dirk Husmeier[**]

Biomathematics and Statistics Scotland, JCMB, KB, Edinburgh EH9 3JZ, UK
Email: alexm@bioss.ac.uk and dirk@bioss.ac.uk

Abstract. We present an improved phylogenetic factorial hidden Markov model (FHMM) for detecting two types of mosaic structures in DNA sequence alignments, related to (1) recombination and (2) rate heterogeneity. The focus of the present work is on improving the modelling of the latter aspect. Earlier papers have modelled different degrees of rate heterogeneity with separate hidden states of the FHMM. This approach fails to appreciate the intrinsic difference between two types of rate heterogeneity: long-range regional effects, which are potentially related to differences in the selective pressure, and the short-term periodic patterns within the codons, which merely capture the signature of the genetic code. We propose an improved model that explicitly distinguishes between these two effects, and we assess its performance on a set of simulated DNA sequence alignments.

1 Introduction

DNA sequence alignments are usually not homogeneous. Mosaic structures may result as a consequence of recombination or rate heterogeneity. Interspecific recombination, in which DNA subsequences are transferred between different (typically viral or bacterial) species may result in a change of the topology of the underlying phylogenetic tree. Rate heterogeneity corresponds to a change of the nucleotide substitution rate. Two Bayesian methods for simultaneously detecting recombination and rate heterogeneity in DNA sequence alignments are the dual multiple change-point model (DMCP) of [13], and the phylogenetic factorial hidden Markov model (PFHMM) of [9] and [12]. The idea underlying the DMCP is to segment the DNA sequence alignment by the insertion of change-points, and to infer different phylogenetic trees and nucleotide substitution rates for the separate segments thus obtained. Two separate change-point processes associated with the tree topology and the nucleotide substitution rate are employed. Inference is carried out in a Bayesian way with reversible jump (RJ) Markov chain Monte Carlo (MCMC). Of particular interest are the number and locations of the change-points, which mark putative recombination break-points and regions putatively under different selective pressures. A related modelling paradigm is provided by the PFHMM, where two *a priori* independent hidden Markov chains are introduced, whose states represent the tree topology and nucleotide substitution rate, respectively. While the earlier work of [9] kept the number of hidden

[**] This work was funded by RERAD of the Scottish Government.

states fixed, [12] generalised the inference procedure with RJMCMC and showed that this framework subsumes the DMCP as a special case. This model has recently been extended to larger numbers of species [16].

Common to all these models are two simplifications. First, the no-common mechanism model of [15] is introduced, which assumes separate branch lengths for each site in the DNA sequence alignment. Second, there is no distinction between regional and within-codon rate heterogeneity. Following [14], the first assumption was introduced with the objective to reduce the computational complexity of the inference scheme. The no-common-mechanism model allows the branch lengths to be integrated out analytically. This is convenient, as the marginal likelihood of the tree topology, the nucleotide substitution rate, and further parameters of the nucleotide substitution model (like the transition-transversion ratio) can be computed in closed from. In this way, the computational complexity of sampling break-points (DMCP) or hidden state sequences (PFHMM) from the posterior distribution with MCMC is substantially reduced. However, in the no-common-mechanism model the branch lengths are incidental rather than structural parameters. As we discussed in [10], this implies that maximum likelihood no longer provides a consistent estimator, and that the method systematically infers the wrong tree topology in the Felsenstein zone defined in [3]. The second simplification does not distinguish between two different types of rate heterogeneity: (1) a regional effect, where larger consecutive segments of the DNA sequence alignment might be differently evolved, e.g. as a consequence of changes of the selective pressure; (2) and a codon effect, where the third codon position shows more variation than the first or the second. Not allowing for this difference and treating both sources of rate heterogeneity on an equal footing implies the risk that subtle regional effects might be obscured by the short-range codon effect, as discussed in [12]. The latter effect is of no biological interest, though, as it only represents the signature of the genetic code.

In the present work, we address this issue and develop a model that properly distinguishes between these two effects. Our work is based on the model we introduced in [10]. We modify this approach so as to explicitly take the signature of the genetic code into account. In this way, the within-codon effect of rate heterogeneity is imposed on the model *a priori*, which makes it easier to learn the biologically more interesting effect of regional rate heterogeneity *a posteriori*.

2 Methodology

2.1 Modelling recombination and rate heterogeneity with a phylogenetic FHMM

Consider an alignment $\boldsymbol{\mathcal{D}}$ of m DNA sequences, N nucleotides long. Let each column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector containing the nucleotides at the tth site of the alignment, and $\boldsymbol{\mathcal{D}} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Given a probabilistic model of nucleotide substitutions based on a homogeneous Markov chain with instantaneous rate matrix $\mathbf{Q}$, a phylogenetic tree topology S, and a vector of branch lengths $\mathbf{w}$, the probability of each column $\mathbf{y}_t$, $P(\mathbf{y}_t|S, \mathbf{w}, \boldsymbol{\theta})$, can be computed, as e.g. discussed in [4]. Here, $\boldsymbol{\theta}$ denotes a (vector) of free nucleotide substitution parameters extracted from $\mathbf{Q}$. For instance, for the HKY85 model of [7], we have $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, with $\pi_i \in [0, 1]$ and $\sum_i \pi_i = 1$, is a vector of nucleotide equilibrium frequencies, and $\alpha, \beta \geq 0$ are separate nucleotide substitution rates for transitions and transversions. For identifiability between $\mathbf{w}$ and $\mathbf{Q}$, the constraint $\sum_i Q_{ii}\pi_i = -1$ is commonly introduced, which allows the branch lengths to be interpreted as expected numbers of mutations per site (see, e.g., [13]). The normalisation constraint on $\boldsymbol{\pi}$ further reduces the number of free parameters by one, so that without loss of generality we have $\boldsymbol{\theta} = (\pi_A, \pi_C, \pi_G, \zeta)$, where $\zeta = \alpha/(2\beta) \geq 0$ is the transition-transversion ratio. In what follows, we do not make the dependence on $\boldsymbol{\theta}$ explicit in our notation.

We simultaneously model recombination and rate heterogeneity with a phylogenetic FHMM, as originally proposed in [9], with the modification discussed in [10]. A hidden variable $S_t \in \{\tau_1, \dots, \tau_K\}$ is introduced, which represents one out of K possible tree topologies τ_i at site t. To allow for correlations between nearby sites – while keeping the computational complexity limited – a Markovian dependence structure is introduced: $P(\mathbf{S}) = P(S_1, \dots, S_N) = \prod_{t=2}^N P(S_t|S_{t-1})P(S_1)$. Following [5], the transition probabilities are defined as

$$P(S_t|S_{t-1}, \nu_S) = \nu_S^{\delta(S_t,S_{t-1})} \left(\frac{1-\nu_S}{K-1}\right)^{[1-\delta(S_t,S_{t-1})]} \tag{1}$$

where $\delta(S_t, S_{t-1})$ denotes the Kronecker delta symbol, which is 1 when $S_t = S_{t-1}$, and 0 otherwise. The parameter ν_S denotes the probability of not changing the tree topology between adjacent sites. Associated with each tree topology τ_i is a vector of branch lengths, $\mathbf{w}_{\tau_i}$, which defines the probability of a column of nucleotides, $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t})$. The practical computation follows standard methodology based on the pruning algorithm [4]. For notational convenience we rewrite these *emission probabilities* as $P(\mathbf{y}_t|S_t, \mathbf{w})$, where $S_t \in \{\tau_1, \dots, \tau_k\}$ determines which of the subvectors $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$ is selected. To model rate heterogeneity, a second type of hidden states R_t is introduced. Correlations between adjacent sites are modelled again by a Markovian dependence structure: $P(\mathbf{R}) = P(R_1, \dots, R_N) = \prod_{t=2}^N P(R_t|R_{t-1})P(R_1)$. The transition probabilities are defined as in (1):

$$P(R_t|R_{t-1}, \nu_R) = \nu_R^{\delta(R_t,R_{t-1})} \left(\frac{1-\nu_R}{\tilde{K}-1}\right)^{[1-\delta(R_t,R_{t-1})]} \tag{2}$$

where $\tilde{K}$ is the total number of different rate states. Each rate state is associated with a scaling parameter $R_t \in \boldsymbol{\rho} = \{\rho_1, \dots, \rho_{K'}\}$ by which the branch lengths are rescaled: $P(\mathbf{y}_t|S_t, \mathbf{w}) \rightarrow P(\mathbf{y}_t|S_t, R_t\mathbf{w})$. To ensure that the model is identifiable, we constrain the L1-norm of the branch length vectors to be equal to one: $||\mathbf{w}_k||_1 = 1$ for $k = 1, \dots, K$. To complete the specification of the probabilistic model, we introduce prior probabilities on the transition parameters ν_S and ν_R, which are given conjugate beta distributions (which subsume the uniform distribution for the uninformative case). The initial state probabilities $P(S_1)$ and

$P(R_1)$ are set to the uniform distribution, as in [11]. The prediction of recombination break-points and rate heterogeneity is based on the marginal posterior probabilities

$$P(S_t|\boldsymbol{\mathcal{D}}) = \sum_{S_1} \cdots \sum_{S_{t-1}} \sum_{S_{t+1}} \cdots \sum_{S_N} P(\mathbf{S}|\boldsymbol{\mathcal{D}}) \tag{3}$$

$$P(R_t|\boldsymbol{\mathcal{D}}) = \sum_{R_1} \cdots \sum_{R_{t-1}} \sum_{R_{t+1}} \cdots \sum_{R_N} P(\mathbf{R}|\boldsymbol{\mathcal{D}}) \tag{4}$$

The distributions $P(\mathbf{S}|\boldsymbol{\mathcal{D}})$ and $P(\mathbf{R}|\boldsymbol{\mathcal{D}})$ are obtained by the marginalisation

$$P(\mathbf{S}|\boldsymbol{\mathcal{D}}) = \sum_{\mathbf{R}} \int P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \mathbf{w}|\boldsymbol{\mathcal{D}}) d\nu_S d\nu_R d\mathbf{w} \tag{5}$$

$$P(\mathbf{R}|\boldsymbol{\mathcal{D}}) = \sum_{\mathbf{S}} \int P(\mathbf{R}, \mathbf{S}, \nu_S, \nu_R, \mathbf{w}|\boldsymbol{\mathcal{D}}) d\nu_S d\nu_R d\mathbf{w} \tag{6}$$

where $P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \mathbf{w}|\boldsymbol{\mathcal{D}}) \propto P(\boldsymbol{\mathcal{D}}, \mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \mathbf{w}) = P(S_1)P(R_1)P(\nu_S)P(\nu_R)$ $\prod_{t=1}^{N} P(\mathbf{y}_t|S_t, R_t\mathbf{w}) \prod_{t=2}^{N} P(S_t|S_{t-1}, \nu_S) \prod_{t=2}^{N} P(R_t|R_{t-1}, \nu_R)$. The respective integrations and summations are intractable and have to be numerically approximated with Markov chain Monte Carlo (MCMC): we sample from the joint posterior distribution $P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \mathbf{w}|\boldsymbol{\mathcal{D}})$ and then marginalise with respect to the entities of interest. Sampling from the joint posterior distribution follows a Gibbs sampling procedure [2], where each parameter group is iteratively sampled separately conditional on the others. So if the superscript (i) denotes the ith sample of the Markov chain, we obtain the $(i+1)$th sample as follows:

$$\mathbf{S}^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}) \tag{7}$$

$$\mathbf{R}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}) \tag{8}$$

$$\nu_S^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_R^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}) \tag{9}$$

$$\nu_R^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}) \tag{10}$$

$$\mathbf{w}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \nu_R^{(i+1)}, \boldsymbol{\mathcal{D}}) \tag{11}$$

The order of these sampling steps is arbitrary. Note that, in principle, the nucleotide substitution parameters $\boldsymbol{\theta}$ should be included in the Gibbs scheme, as described in [11]. In practice, a fixation of $\boldsymbol{\theta}$ at *a priori* estimated values makes little difference to the prediction of $P(S_t|\boldsymbol{\mathcal{D}})$ and $P(R_t|\boldsymbol{\mathcal{D}})$ and has the advantage of reduced computational costs. Sampling the hidden state sequences $\mathbf{S}$ and $\mathbf{R}$ in (7) and (8) is effected with the stochastic forward-backward algorithm of [1]. Sampling the transition probabilities ν_S and ν_R in (9) and (10) is straightforward due to the conjugacy of the beta distribution. Sampling the branch lengths in (11) cannot be effected from a closed-form distribution, and we have to resort to a Metropolis-Hastings-within-Gibbs scheme. Note that the branch lengths have to satisfy the constraint $||\mathbf{w}_k||_1 = 1$, $k = 1, \ldots, K$, as well as the positivity constraint $w_{ki} \geq 0$. This is automatically guaranteed when proposing new branch length vectors $\mathbf{w}_k^*$ from a Dirichlet distribution: $Q(\mathbf{w}_k^*|\mathbf{w}_k) \propto \prod_i [w_{ki}^*]^{\alpha w_{ki}-1}$, where α is a tuning parameter that can be adapted during burn-in to improve mixing. The acceptance probability for the proposed branch lengths is then given by the standard Metropolis-Hastings criterion [8].

2.2 Distinguishing regional from within-codon rate heterogeneity

We improve the model described in the previous subsection, which was proposed in [10], in two respects. First, we adapt $\boldsymbol{\rho}$ and sample it along with $\mathbf{w}$ from the posterior distribution. To make this explicit in the notation, we slightly change the definition of the rate state as $R_t \in \{1, \ldots, K'\}$ and rewrite: $P(\mathbf{y}_t|S_t, R_t\mathbf{w}) \rightarrow P(\mathbf{y}_t|S_t, \rho_{R_t}\mathbf{w})$. Second, we explicitly model codon-position-specific rate heterogeneity in a way similar to [5]. To this end, we introduce the indicator variable $I_t \in \{0, 1, 2, 3\}$, where $I_t = 0$ indicates that the tth position of the alignment does not code for protein, and $I_t = i \in \{1, 2, 3\}$ indicates that site t is the ith position of a codon. Each of the four categories is associated with a positive factor taken from $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$, by which the branch lengths are modulated. The emission probabilities are thus given by $\tilde{P}(\mathbf{y}_t|S_t, R_t, I_t, \boldsymbol{\rho}, \boldsymbol{\lambda}, \mathbf{w}) := P(\mathbf{y}_t|S_t, \rho_{R_t}\lambda_{I_t}\mathbf{w})$, where $P(.)$ was defined below equation (1), and $\tilde{P}(.)$ makes the dependence on $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ explicit. Note that as opposed to [5], we do not keep $\boldsymbol{\lambda}$ fixed, but sample it from the posterior distribution with MCMC. For identifiability we introduce the same constraint as for the branch lengths: $||\boldsymbol{\lambda}||_1 = 1$, which is automatically guaranteed when proposing $\boldsymbol{\lambda}$ from a Dirichlet distribution. Hence, to sample $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ from the posterior distribution $P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \boldsymbol{\rho}, \boldsymbol{\lambda}, \mathbf{w}|\boldsymbol{\mathcal{D}})$, we have to add two Metropolis-Hastings-within-Gibbs steps akin to equation (11) to the Gibbs sampling procedure (7-11):

$$[\boldsymbol{\rho}^{(i+1)}, \boldsymbol{\lambda}^{(i+1)}] \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \nu_R^{(i+1)}, \mathbf{w}^{(i+1)}, \boldsymbol{\mathcal{D}}) \tag{12}$$

With all other parameters and hidden states fixed, we propose new values for $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$, and accept or reject according to the Metropolis-Hastings criterion. As discussed above, we propose new values for $\boldsymbol{\lambda}$ from a Dirichlet distribution. New values for $\boldsymbol{\rho}$ are proposed from a uniform distribution (on the log scale), centred on the current values. The dispersal parameters of the proposal distributions can be adjusted during the burn-in phase using standard criteria.

3 Data

To assess the performance of the method, we tested it on synthetic DNA sequence alignments; this has the advantage that we have a known gold-standard. For a realistic simulation, we generated sequence alignments with Seq-Gen, developed by Rambaut and Grassly. This software package is widely used for Monte Carlo simulations of molecular sequence evolution along phylogenetic trees; see e.g. `http://bioweb2.pasteur.fr/docs/seq-gen/` or `http://tree.bio.ed.ac.uk/software/seqgen/` for details. We generated a DNA sequence alignment from a phylogenetic tree of four hypothetical taxa

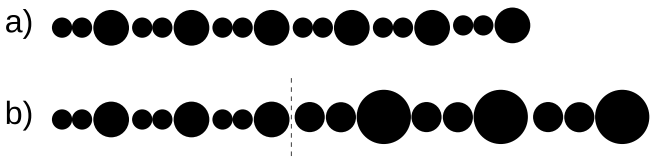

Fig. 1. Illustration of regional versus within-codon rate heterogeneity. Each circle corresponds to a nucleotide in a DNA sequence, and the circle diameter symbolises the average nucleotide substitution rate at the respective position. The top panel (a) shows a "homogeneous" DNA sequence composed of six codons, where each third position is more diverged as a consequence of the nature of the genetic code. The bottom panel (b) shows a hypothetical DNA sequence subject to regional rate heterogeneity, where the second half on the right of the dashed vertical line constitutes a region that is more evolved. The sequences used in our simulation study were similar, but longer (1.5Kbp).

with equal branch lengths, using the HKY model of nucleotide substitution [7] with a uniform nucleotide equilibrium distribution, $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$, and a transition-transversion ratio of $\zeta = 2$. We generated two types of alignments. In the first alignment, the normalised branch lengths associated with the three codon positions were set to $w_i = [0.5 - \frac{c}{2}, 0.5 - \frac{c}{2}, 0.5 + c]/1.5$, where the codon offset parameter $0 \leq c \leq 0.99$ was varied in increments of 0.1. All codons had the same structure, as illustrated in Figure 1a. We refer to these sequence alignments as "homogeneous". The second type of alignment, which we refer to as "heterogeneous" or "subject to regional rate heterogeneity", is illustrated in Figure 1b. The codons have a similar structure as before. The second half of the alignment is more evolved, though, and the branch lengths are expanded by a factor of $\varsigma = 2$. In all simulations, the total length of the alignment was 1.5 Kbp.

4 Simulations

Our objective is to sample topology and rate state sequences $\mathbf{S}, \mathbf{R}$, their associated transition probabilities ν_S, ν_R and rate vectors $\boldsymbol{\rho}$, the branch lengths $\mathbf{w}$ and (for the new model) the within-codon rate vector $\boldsymbol{\lambda}$ from the posterior distribution $P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \boldsymbol{\rho}, \boldsymbol{\lambda}, \mathbf{w}|\boldsymbol{\mathcal{D}})$. To this end, we apply the Gibbs sampling scheme of (7–12), which we have described in Sections 2.1 and 2.2. Our current software has not yet been optimised for speed. Hence, to improve the convergence of the Markov chain and to focus on the aspect of interest for the present study (rate heterogeneity), we have set all states in $\mathbf{S}$ to the same tree topology without allowing for recombination: $\nu_S = 1$. We also set $K' = 2$ fixed. The model was initialised with the maximum likelihood tree obtained with DNAML from Felsentein's PHYLIP package, available from `http://evolution.genetics.washington.edu/phylip/`. We tested the convergence of the MCMC simulations by computing the potential scale reduction factor of Gelman and Rubin [6] from the within and between trajectory variances of various monitoring quantities (e.g. $\mathbf{w}$, $P(R_t|\boldsymbol{\mathcal{D}})$, etc.), and took a value of 1.2 as an indication of sufficient convergence.

The main objective of our study is to evaluate the performance of the proposed model that allows for within-codon rate heterogeneity; we refer to this as the "new" model. We compare its performance with a model that does not include within-codon rate heterogeneity, that is, where $\boldsymbol{\lambda} = \mathbf{1}$ is constant. We refer to this as the "old" model. Note that the latter model is equivalent to the one proposed in [10], but with the improvement that $\boldsymbol{\rho}$ is sampled from the posterior distribution, rather than kept fixed.

In order to evaluate the performance of the methods, we want to compute the marginal posterior probability of the average effective branch length scaling for the three codon positions. The effective branch lengths are given by $\tilde{\mathbf{w}}_t = \rho_{R_t}\lambda_{I_t}\mathbf{w}_t$, where $\mathbf{w}_t$ are the normalised branch lengths. The entity of interest is

$$\Upsilon_t \ = \ \frac{||\tilde{\mathbf{w}}_t||_1}{||\mathbf{w}_t||_1} = \rho_{R_t}\lambda_{I_t} \tag{13}$$

which is the scaling factor by which the branch length vector $\tilde{\mathbf{w}}_t$ associated with position t deviates from the normalised branch lengths $\mathbf{w}_t$. Note that Υ_t is composed of two terms, associated with a region (ρ_{R_t}) and a codon (λ_{I_t}) effect. We are interested in the marginal posterior distribution of this factor, $P(\Upsilon|\boldsymbol{\mathcal{D}}, I = k)$, for the three codon positions $I \in \{1, 2, 3\}$. In practice, this distribution is estimated from the MCMC sample by the appropriate marginalisation with respect to all other quantities:

$$P(\Upsilon|\boldsymbol{\mathcal{D}}, I = k) \ \approx \ \frac{\sum_{i=1}^{M}\sum_{t=1}^{N}\delta_{I_t,k}\delta(\Upsilon - \rho^i_{R^i_t}\lambda^i_{I_t})}{M\sum_{t=1}^{N}\delta_{I_t,k}} \tag{14}$$

where the subscript t refers to positions in the alignment (of total length N), the superscript i refers to MCMC samples (sample size M), $\delta(.)$ is the delta function, the quantities on the right of its argument, $\rho^i_{R^i_t}, \lambda^i_{I_t}$, are obtained from the MCMC sample, and $\delta_{i,k}$ is the Kronecker delta. For the conventional model without explicit codon effect, we set $\lambda_{I_t} = 1/3\,\forall t$.

5 Results

Figure 2 shows the posterior distribution of the (complementary) transition probability ν_R. The two models were applied to the "homogeneous" DNA sequence alignment that corresponds to the top panel in Figure 1. The left panel shows the results obtained with the old model, which does not explicitly include the codon effect. For small values of the offset parameter c, the posterior distribution of ν_R is concentrated on $\nu_R = 1$, which corresponds to a homogeneous sequence alignment. As the offset increases, the posterior distribution of ν_R gets shifted to smaller values, with a mode at $\nu_R = 0.5$. Note that ν_R is related to the average segment length $\bar{l}$ via the relation $\bar{l} = (1-\nu_R)\sum_l l\nu_R^{l-1} = (1-\nu_R)\frac{d}{d\nu_R}\sum_l \nu_R^l = (1-\nu_R)\frac{d}{d\nu_R}\frac{1}{1-\nu_R} = \frac{1}{1-\nu_R}$. For $\nu_R = 0.5$ we get $\bar{l} = 2$. The model has thus learned

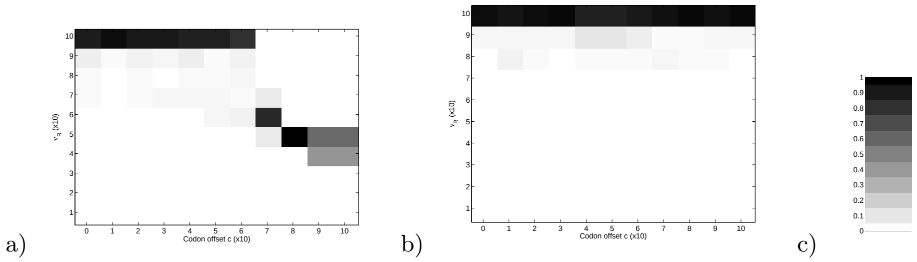

Fig. 2. Posterior distribution of ν_R (vertical axis) for different codon offsets c (horizontal axis), where the offset indicates to what extent the nucleotide substitution rate associated with the third codon position is increased over that of the first two positions. The left panel (a) shows the results obtained with the old model, the centre panel (b) shows the results obtained with the new model. The grey levels represent probabilities, as indicated by the legend in the panel on the right (c). The distributions were obtained from a "homogeneous" DNA sequence alignment, corresponding to Figure 1a.

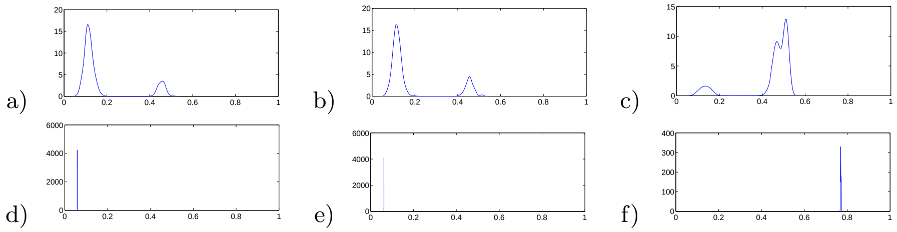

Fig. 3. Posterior distribution (vertical axes) of the combined rate Υ_t (horizontal axes), defined in equation (13), for a "homogeneous" DNA sequence alignment, corresponding to Figure 1a, with codon offset parameter $c = 0.8$. The three columns correspond to the three codon positions. The top row shows the distribution obtained with the old model. The bottom row shows the distribution obtained with the new model. The distributions were obtained from the MCMC samples with a kernel density estimator, where the delta function in (14) was replaced by a Gaussian (standard deviation: a tenth of the total range).

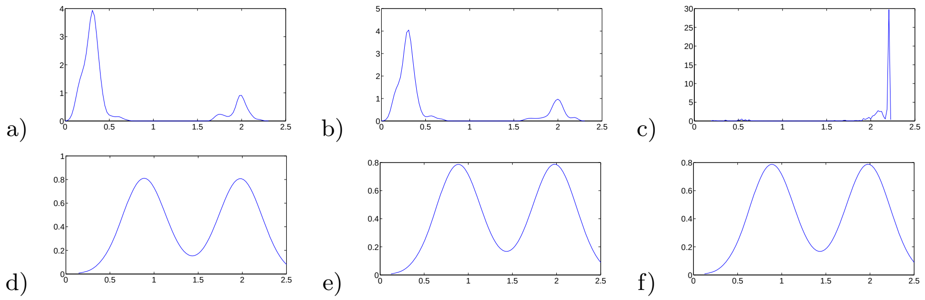

Fig. 4. Posterior distribution (vertical axes) of the rate ρ_{R_t} (horizontal axes) for a "heterogeneous" DNA sequence alignment, corresponding to Figure 1b, with codon offset parameter $c = 0.8$ and regional factor $\varsigma = 2$. The three columns correspond to the three codon positions. The top row shows the distribution obtained with the old model. The bottom row shows the distribution obtained with new model. The distributions were obtained from the MCMC samples with a kernel density estimator, where the delta function in (15) was replaced by a Gaussian (standard deviation: a tenth of the total range).

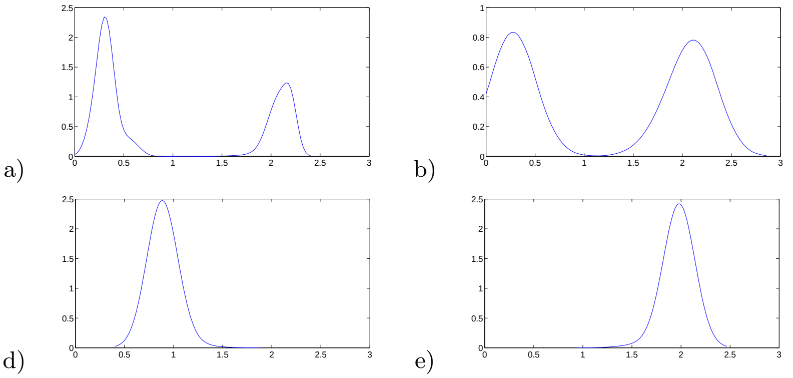

Fig. 5. Alternative representation of the posterior distribution (vertical axes) of the rate ρ_{R_t} (horizontal axes) for the "heterogeneous" DNA sequence alignment. The figure corresponds to Figure 4, but shows a separation of the distributions with respect to regions rather than codon positions. The distribution of ρ_{R_t} is defined in (16). The two columns correspond to the two differently diverged segments in the DNA sequence alignments, with the left column representing the first 750 positions, and the right column representing the last 750 positions; the latter were evolved at double the nucleotide substitution rate. The two rows correspond to the two models. The top row shows the distribution obtained with the old model. The bottom row shows the distribution obtained with new model. The distributions were obtained from the MCMC samples with a kernel density estimator, where the delta function in (16) was replaced by a Gaussian (standard deviation: a tenth of the total range).

the within-codon rate heterogeneity intrinsic to the genetic code; compare with Figure 1. The right panel of Figure 2 shows the posterior distribution of ν_R obtained with the new model. Irrespective of the codon offset c, the distribution is always concentrated on $\nu_R = 1$. This correctly indicates that there is no regional rate heterogeneity in the DNA sequence alignment. Recall that the within-codon rate heterogeneity has been explicitly incorporated into the new model and, hence, need not be learned separately via ν_R and transitions between rate states R_t.

Figure 3 shows the posterior distribution of the scaling factor Υ_t, defined in (13), for the "homogeneous" DNA sequence alignment corresponding to Figure 1a. The columns in Figure 3 correspond to the three codon positions. The posterior distribution was obtained from the MCMC samples via (14). For the new model (bottom row of Figure 3), the distributions of Υ_t are unimodal and sharply peaked. This is consistent with the fact that we have no regional rate heterogeneity, and the shift in the peak locations for the third codon position clearly indicates the within-codon rate heterogeneity. For the old model (top panel of Figure 3), the posterior distribution is always bimodal. This is a consequence of the fact that the within-codon rate heterogeneity has to be learned via the assignment of rate states R_t to the respective codon positions. The bimodality and increased width of the distribution stem from a misassignment of rate states. Note that for an alignment of $N = 1500$ sites, 500 state transitions have to be learned to model the within-codon rate heterogeneity correctly.

Figure 4 is similar to Figure 3, but was obtained for the heterogeneous DNA sequence alignment corresponding to Figure 1b. For better clarity we have shown the codon site-specific posterior distributions of the rate ρ_{R_t} rather than the scale factor Υ_t, that is, in equation (14) we have ignored the factor $\lambda^i_{I_t}$:

$$P(\rho|\boldsymbol{\mathcal{D}}, I = k) \;\approx\; \frac{\sum_{i=1}^{M}\sum_{t=1}^{N}\delta_{I_t,k}\delta(\rho - \rho^i_{R^i_t})}{M\sum_{t=1}^{N}\delta_{I_t,k}} \tag{15}$$

The bottom row shows the distributions obtained with the new model. They have a symmetric bimodal form. The bimodality reflects the regional rate heterogeneity. The symmetry reflects the nature of the DNA sequence alignment, which contains two differently diverged regions of equal size (see Figure 1b). The top panel shows the distributions obtained with the old model. The distributions are still bimodal, but the symmetry has been destroyed. This distortion results from the fact that two effects – regional and within-codon rate heterogeneity – are modelled via the same mechanism: the rate states R_t. Consequently, these two forms of rate heterogeneity are not clearly separated.

To illustrate this effect from a different perspective, Figure 5 shows the posterior distributions of the rate ρ_{R_t} not separated according to codon positions, but according to differently diverged regions. That is, from the MCMC sample we compute the following distribution:

$$P(\rho|\boldsymbol{\mathcal{D}}, t \in r) \;\approx\; \frac{\sum_{i=1}^{M}\sum_{t=1}^{N}\mathcal{I}(t \in r)\delta(\rho - \rho^i_{R^i_t})}{M\sum_{t=1}^{N}\mathcal{I}(t \in r)} \tag{16}$$

where r represents the two regions: $r = 1$ for $1 \leq t \leq 750$, and $r = 2$ for $751 \leq t \leq 1500$, $\mathcal{I}(t \in r)$ is the indicator function, which is one if the argument is true, and zero otherwise, and the remaining symbols are as defined below equation (14). The bottom panel shows the distributions obtained with the new model, where the two columns represent the two regions. The distributions are unimodal and clearly separated, which indicates that modelling regional rate heterogeneity is properly disentangled from the within-codon rate variation. The top panel shows the distributions obtained with the old model. Here, the distributions are bimodal, which results from a lack of separation between regional and within-codon rate heterogeneity, and a tangling-up of these two effects.

6 Discussion

We have generalised the phylogenetic FHMM of [10] in two respects. First, by sampling the rate vector $\boldsymbol{\rho}$ from the posterior distribution with MCMC (rather than keeping it fixed) we have made the modelling of regional rate heterogeneity more flexible. Second, we explicitly model within-codon rate heterogeneity via a separate rate modification vector $\boldsymbol{\lambda}$. In this way, the within-codon effect of rate heterogeneity is imposed on the model *a priori*, which should facilitate the learning of the biologically more interesting effect of regional rate heterogeneity *a posteriori*. We have carried out simulations on synthetic DNA sequence alignments, which have borne out our conjecture. The old model, which does not explicitly include the within-codon rate variation, has to model both effects with the same mechanism: the rate states R_t with associated rate factors ρ_{R_t}. As expected, it was found to fail to disentangle these two effects. On the contrary, the new model was found to clearly separate within-codon from regional rate heterogeneity, resulting in a more accurate prediction.

We emphasise that our paper describes work in progress, and we have not yet applied our method to real DNA sequence alignments. This is partly a consequence of the fact that our software has not been optimised for computational efficiency yet, resulting in long MCMC simulation runs. Note that the computational complexity of our algorithm is larger than for the model described in [12]. The latter approach is based on the no-common-mechanism model of [15], which leads to a substantial model simplification, though at the price of potential inconsistency problems (as discussed in [10]). The increased computational complexity of the method proposed in the present article might require the application of more sophisticated MCMC schemes, e.g. population MCMC, which will be the objective of our future work.

As a final remark, we note that a conceptually superior approach would be the modelling of substitution processes at the codon rather than nucleotide level. However, the application of this approach to standard Bayesian analysis of single phylogenetic trees has turned out to be computationally exorbitant. A generalisation to phylogenetic FHMMs for modelling DNA mosaic structures, as described in the present article, is unlikely to be computationally feasible in the near future. We therefore believe that the method we have proposed, which is based on individual nucleotide substitution processes while taking the codon structure into account, promises a better compromise between model accuracy and practical viability.

References

1. R. J. Boys, D. A. Henderson, and D. J. Wilkinson. Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics*, 49:269–285, 2000.
2. G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
3. J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–440, 1978.
4. J. Felsenstein. Evolution trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.
5. J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104, 1996.
6. A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.
7. M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.
8. W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
9. D. Husmeier. Discriminating between rate heterogeneity and interspecific recombination in dna sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, 21:ii166–ii172, 2005.
10. D. Husmeier and A. V. Mantzaris. Addressing the shortcomings of three recent Bayesian methods for detecting interspecific recombination in DNA sequence alignments. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article 34, 2008.
11. D. Husmeier and G. McGuire. 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, 2003.
12. W. P. Lehrach and D. Husmeier. Segmenting bacterial and viral DNA sequence alignments with a trans-dimensional phylogenetic factorial hidden Markov model. *Applied Statistics*, page in print, 2009.
13. V. N. Minin, K. S. Dorman, F. Fang, and M. A. Suchard. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(13):3034–3042, 2005.
14. M. A. Suchard, R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of the American Statistical Association*, 98(462):427–437, 2003.
15. C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607, 1997.
16. A. Webb, J. Hancock, and C. Holmes. Phylogenetic inference under recombination using Bayesian stochastic topology selection. *Bioinformatics*, in press, 2009.