Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 18

Pseudo-likelihood for Non-reversible Nucleotide Substitution Models with Neighbour Dependent Rates

Ole F. Christensen*

*Bioinformatics Research Center, Aarhus University, Denmark, olefc@birc.au.dk

Pseudo-likelihood for Non-reversible Nucleotide Substitution Models with Neighbour Dependent Rates*

Ole F. Christensen

Abstract

In the field of molecular evolution genome substitution models with neighbour dependent substitution rates have recently received much attention. It is well-known that substitution of nucleotides does not occur independently of neighbouring nucleotides, but there has been less focus on the phenomenon that this substitution process is also not time-reversible. In this paper I construct a pseudo-likelihood type method for inference in non-reversible substitution models with neighbour dependent substitution rates. I also construct an EM-algorithm for maximising the pseudo-likelihood. For human-mouse aligned sequence data a number of different models are investigated, where I show that strand-symmetric models are appropriate, and that overlapping di-nucleotide models do not fit the data well.

KEYWORDS: EM-algorithm, Markov processes, maximum likelihood, pseudo-likelihood

---

*The author wish to thank Gerton Lunter for generously sharing his data and programs, and also for some helpful discussions. Asger Hobolth, Ojvind Skare and two anonymous reviewers are thanked for providing helpful comments on the manuscript. Financial support from SJVF grant 2052-01-0032 is also acknowledged.

1 Introduction

In molecular evolution stochastic process models are used to elucidate the biological processes generating variation within and between species at the molecular level. When DNA sequences from different species are sampled the focus is on nucleotide substitutions, the prevalent process in the long term evolution of sequences. The substitution of DNA sequences is modelled using time-homogeneous continuous time Markov processes; an introduction to the topic is Galtier, Gascuel and Jean-Marie (2005). In practice the most commonly used models have been simple ones, i.e. Jukes-Cantor, Kimura and HKY, which all share the following properties : 1) the evolution occurs independently at each position, 2) the process is time-reversible, and 3) the process is stationary.

It is well-known that the substitution of nucleotides does not occur independently of neighbouring nucleotides, i.e. the CpG effect where an excess of substitutions is observed at positions with a CpG di-nucleotide. Substitution models with neighbour dependence have been considered by Jensen and Pedersen (2000), Duret and Galtier (2000), Pedersen and Jensen (2001), Arndt, Burge and Hwa (2003a), Arndt, Petrov and Hwa (2003b), Siepel and Haussler (2004), Jojic, Jojic, Geiger, Siepel, Haussler and Heckerman (2004), Lunter and Hein (2004), Hwang and Green (2004), Christensen, Hobolth and Jensen (2005), Jensen (2005), Arndt and Hwa (2005) and Hobolth (2006). For mammalian genomes, an explanation of the CpG effect is the CpG-methylation-deamination mutational process, where a CpG is substituted by TpG (which is seen as a CpG to CpA substitution when it happens on the other strand). This chemical process is a non-reversible phenomenon - further details, see Figure 2 in Lunter and Hein (2004). Likelihood inference for non-reversible Markov process models with neighbour dependent substitution rates is only considered by few of the papers listed above - Arndt *et al.* (2003a) and Arndt and Hwa (2005) consider the case where the ancestral sequence is known, whereas Lunter and Hein (2004) and Hwang and Green (2004) consider an unknown ancestral sequence. Here the case with an unknown ancestral sequence is considered using the nucleotide substitution model introduced by Hwang and Green (2004), which is more general than the one in Lunter and Hein (2004).

Hwang and Green (2004) assume that the model is non-stationary and they use a second order Markov chain model for the distribution of the common ancestor sequence of the observed sequences, i.e. the sequence at the root of the tree relating the observed sequences, whereas Lunter and Hein (2004) consider the model to be stationary using a second order Markov chain approximation

for the equilibrium distribution of the sequence. Yap and Speed (2005) investigate the stationarity assumption for a non-reversible model, although a very simple site independent model, and conclude that a non-stationary non-reversible model provides a much better fit to data and also seems to provide more robust conclusions. In addition, the assumption of stationarity offers no computational advantages for non-reversible models, contrary to reversible models where reversibility+stationarity implies that the root can be placed anywhere on the tree. Therefore, the presentation in this paper follows Hwang and Green (2004) and does not assume stationarity of the process.

A number of specific sub-models are considered. In particular, for the strand-symmetric case an improvement of the Hwang and Green model is introduced, which has both a strand-symmetric substitution process and a strand-symmetric model for the sequence at the root.

For models with neighbour dependent substitution rates the likelihood function becomes intractable in practice and approximate methods are needed for inference. Christensen *et al.* (2005) derive a pseudo-likelihood for the evolution of one sequence to another and they also construct a corresponding EM-algorithm. The approach relies crucially on reversibility and stationarity of the model. Here I extend one of the basic ideas in that paper to models which are non-reversible and non-stationary, and derive a pseudo-likelihood for the Hwang and Green model and also a corresponding EM-algorithm. Due to the second order Markov chain used for the sequence distribution at the root, the computations for both the pseudo-likelihood and the EM-algorithm involve recursions along the sequence. In addition, the non-reversibility of the process implies that the eigenvalue decomposition of certain substitution matrices involves complex numbers.

A competing approach is to use Markov chain Monte Carlo (MCMC) methods for inference (Hwang and Green, 2004; Jensen, 2005; Hobolth, 2006). In Hwang and Green (2004) and Jensen (2005) a discrete time approximation of the substitution process is used, whereas Hobolth (2006) uses an algorithm that actually simulates continuous sample paths. The advantage of such MCMC approaches is the generality, since an implementation of a MCMC-algorithm is often easy to extend to other models, but the disadvantage is that MCMC for such models is computationally very slow to use since all the unobserved sequences in the whole evolutionary history have to be updated in each iteration of the MCMC-algorithm.

Another competing approach is the recursive algorithm for approximation of likelihood function in Lunter and Hein (2004), which avoids some of the MCMC related problems above, and in principle generalises to the more general Hwang and Green model. However, contrary to the pseudo-likelihood

considered here the Lunter and Hein approximation is not a likelihood function in itself, which implies that no corresponding EM-algorithm for finding the maximum of the approximation can be constructed. In addition, the algorithm described in the Appendix of Lunter and Hein (2004) gave negative values in the recursions, and certain modifications had be made to avoid this (personal communication with Gerton Lunter - further details can be found in his code). As regards the more general Hwang and Green model I did not succeed in actually making my attempts to construct an algorithm along the ideas in Lunter and Hein (2004) work well.

To illustrate the use of the pseudo-likelihood, in Section 5 the intergenic human-mouse data set investigated in Lunter and Hein (2004) is considered. I investigate a number of different models, with particular focus on strand-symmetric models, and demonstrate that overlapping di-nucleotide types of models used by Lunter and Hein (2004) do not fit the data well.

2 Nucleotide substitution model with neighbour dependent rates

Here I consider the substitution model in Hwang and Green (2004), which consists of two parts : the Markov process for the sequence to sequence evolution on a given branch in the species tree, and the sequence distribution at the root of the tree.

2.1 Markov process for the sequence to sequence evolution

The nucleotide substitution model describes the evolution of a sequence as a time-homogeneous continuous time Markov process, where a change in the sequence consists of a change of one nucleotide at a time only. A sequence x consisting of n nucleotides is written as $x = (x_1, \ldots, x_n)$ and for a given position k z_k denotes the new nucleotide. The rate for such a change depends upon x_k as well as the nucleotide neighbours x_{k-1} and x_{k+1} and is given by

$$\gamma(z_k; x_{k-1}, x_k, x_{k+1}). \tag{1}$$

The general model (1) has $4 \times 4 \times 4 \times 3 = 192$ parameters, $\gamma(b; l, a, r)$, and it is to be used for intronic regions where strand-asymmetric substitution rates are known to exist - see Hwang and Green (2004), whereas for intergenic

regions a strand-symmetric model is appropriate. Such a model assumes that

$$\gamma(b; l, a, r) = \gamma(\complement b; \complement r, \complement a, \complement l), \tag{2}$$

where the notation $\complement$ denotes the complementary base in the base-pairing, i.e. $\complement$A = T, $\complement$G = C, $\complement$C = G and $\complement$T = A. The number of parameters in the strand-symmetric model is 96.

The over-lapping di-nucleotide model considered in Lunter and Hein (2004), when allowing only single nucleotide substitutions, is the special case of model (1) where

$$\gamma(b; l, a, r) = \nu^{\text{left}}(b; l, a) + \nu^{\text{right}}(b; a, r). \tag{3}$$

The model is over-parameterised, and the number of free parameters is $3 \times 4^2 + 3 \times 4^2 - 3 \times 4 = 84$. In fact Lunter and Hein (2004) confine their attention to strand-symmetry and do not consider the strand-asymmetric model. In that case when allowing only single nucleotide substitutions, $\nu^{\text{left}}(b; l, a) = \nu^{\text{right}}(\complement b; \complement a, \complement l)$ and the model has $3 \times 4^2 = 48$ free parameters. Finally, Lunter and Hein (2004) also allow two substitutions to happen simultaneously in a di-nucleotide, but such an assumption is outside the framework considered here.

The most simple models considered here are models where the only neighbour dependent rate parameters are the ones corresponding to the CpG-methylation-deamination process, i.e.

$$\gamma(b; a_1, a_2, a_3) = \epsilon(b; a) \text{ when } (a_1, a_2, b) \neq (\text{C}, \text{G}, \text{A}) \text{ and } (a_2, a_3, b) \neq (\text{C}, \text{G}, \text{T}). \tag{4}$$

The most general such model has $8 + 3 \times 4 = 20$ parameters, the strand-symmetric model has $4 + 3 \times 2 = 10$ parameters, the overlapping di-nucleotide model $2 + 3 \times 4 = 14$ parameters, and the strand-symmetric overlapping di-nucleotide model $1 + 3 \times 2 = 7$ parameters.

In Section 5 different models for the evolution on two branches of the tree are studied. In particular, both models which assume individual substitution processes on each branch, and models which assume the same substitution process but with evolution on different branches happening with a different speed are considered. The latter type of model assumes that the rate on the vy branch is $\tilde{\gamma} = \tau\gamma$, where γ is the rate on the vx branch, and $\tau > 0$ is the speed of evolution on the vy branch relative to the vx branch.

2.2 The sequence distribution at the root

Since the model is not time-reversible we need to consider evolution from an unknown common ancestor sequence v. The distribution at the root is

$$p^{\text{root}}(v;\pi) = \prod_{k=3}^{n} \pi(v_k \mid v_{k-2}, v_{k-1})\pi_c(v_1, v_2), \tag{5}$$

with the $3 \times 4^2 = 48$ parameters $\pi(a_3 \mid a_1, a_2)$, and π_c determined as the stationary (along the sequence) di-nucleotide frequencies, i.e. π_c and π satisfy

$$\pi_c(a_2, a_3) = \sum_{a_1} \pi(a_3 \mid a_1, a_2)\pi_c(a_1, a_2), \tag{6}$$

for all a_2 and a_3, and π_c is found by solving these equations.

When considering a strand-symmetric substitution model (2), it seems appropriate to also consider a strand-symmetric model for the root distribution, i.e. $p^{\text{root}}((v_1, \ldots, v_n);\pi) = p^{\text{root}}((\complement v_n, \ldots, \complement v_1);\pi)$ for any $(z_1, \ldots, z_n)$. Requiring that this equation holds for all n and all nucleotides v_k, $k = 1 \ldots, n$, is equivalent to

$$\pi_c(a_1, a_2) = \pi_c(\complement a_2, \complement a_1) \tag{7}$$

and

$$\pi_c(a_1, a_2)\pi(a_3 \mid a_1, a_2) = \pi_c(\complement a_3, \complement a_2)\pi(\complement a_1 \mid \complement a_3, \complement a_2), \tag{8}$$

for all (a_3, a_1, a_2). These constraints reduce the dimension of the parameter space considerably to just 25 free parameters for π and π_c. Further details are given in Appendix A. Using a strand-symmetric model for the root distribution is an improvement of the Hwang and Green model.

3 Pseudo-likelihood

In this section the pseudo-likelihood method in Christensen *et al.* (2005) is extended. As previously mentioned, the model considered here is not time-reversible, and we need to consider the evolution from an unknown common ancestor sequence, $v = (v_1, \ldots, v_n)$. Only the case with two sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ is presented here, but the generalisation to more sequences should be obvious in principle. Since divergence times and substitution rates cannot be distinguished, the substitution rates are standardised such that evolution from time $t = 0$ to time $t = 1$ is considered.

Introducing the notation γ for the rate parameters at the vx branch and $\tilde{\gamma}$ for the rate parameters at the vy branch, the likelihood is

$$L(\gamma, \tilde{\gamma}, \pi; x, y) = \sum_v L(\gamma; v, x) L(\tilde{\gamma}; v, y) p^{\text{root}}(v; \pi), \tag{9}$$

where $L(\gamma; v, x)$ and $L(\tilde{\gamma}; v, y)$ are the sequence to sequence likelihoods for the vx branch and vy branch, respectively, and $p^{\text{root}}(v; \pi)$ is given by (5).

Below we consider the approximation for the sequence to sequence likelihood, $L(\gamma; v, x)$, introduced in Christensen *et al.* (2005), and note that the approximation for $L(\tilde{\gamma}; v, y)$ is similar. Knowing the nucleotides at the flanking positions (l is the nucleotide to the left and r is the one to the right), the rate matrix for a single nucleotide position is given by the 4×4 rate matrix

$$Q^{lr}(a, b) = \gamma(b; l, a, r), \tag{10}$$

for $a \neq b$, and with the diagonal $Q^{lr}(a, a) = -\sum_{b \neq a} Q^{lr}(a, b)$. Considering the kth nucleotide, we use the approximate evolutionary events, that if $v_{k-1} = x_{k-1}$ then no substitutions have happened at the left flanking position, and if $v_{k-1} \neq x_{k-1}$ exactly one substitution has happened at time $t = 1/2$, and similarly for the nucleotide to the right. Then the rate matrix for nucleotide k becomes $Q^{v_{k-1}, v_{k+1}}$ for $0 \leq t \leq 1/2$, and $Q^{x_{k-1}, x_{k+1}}$ for $1/2 \leq t \leq 1$. The evolutionary history at position k on the vx branch, e_k^{vx}, consist of a number of substitutions m_k, at the substitution times $t_{k,\ell}$ with new states $s_{k,\ell}, \ell = 1, \ldots, m_k$, and the complete observation likelihood at position k is

$$\begin{aligned} &L_{v_{k-1}, v_{k+1}; x_{k-1}, x_{k+1}}(e_k^{vx} | v_k) \\ &\quad = \left\{ \prod_{\ell=1}^{m_k} Q^*(s_{k,\ell-1}, s_{k,\ell}, t_{k,\ell}) e^{\int_{t_{k,\ell-1}}^{t_{k,\ell}} Q^*(s_{k,\ell-1}, s_{k,\ell-1}, t) dt} \right\} e^{\int_{t_{k,m_k}}^{1} Q^*(x_k, x_k, t) dt}, \end{aligned} \tag{11}$$

where $t_{k,0} = 0$, $s_{k,0} = v_k$, $s_{k,m_k} = x_k$ and

$$Q^*(a, b, t) = \begin{cases} Q^{v_{k-1}, v_{k+1}}(a, b) & 0 \leq t \leq 1/2 \\ Q^{x_{k-1}, x_{k+1}}(a, b) & 1/2 < t \leq 1. \end{cases}$$

Marginalising the unobserved evolutionary history of nucleotide k, e_k^{vx}, we obtain the approximation of nucleotide to nucleotide probabilities at position k when knowing the flanking nucleotides

$$L_{v_{k-1}, v_{k+1}; x_{k-1}, x_{k+1}}(x_k | v_k) = \left[\exp\left(Q^{v_{k-1} v_{k+1}}/2\right) \exp\left(Q^{x_{k-1} x_{k+1}}/2\right) \right]_{v_k, x_k}. \tag{12}$$

To compute the matrix exponentials in (12), Schadt and Lange (2002) note that in numerical practice substitution matrices Q^{lr} are complex diagonalisable, and hence these matrix exponentials can be computed using an eigenvalue decomposition; further details are found in Appendix A.

From (9) and (5) the pseudo-likelihood becomes

$$\begin{aligned} & L_p(\gamma, \tilde{\gamma}, \pi) \\ & \quad = \sum_v \left(\prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(x_k|v_k) L_{v_{k-1},v_{k+1};y_{k-1},y_{k+1}}(y_k|v_k) \right) p^{\text{root}}(v;\pi) \\ & \quad = \sum_{v_1,\ldots,v_n} \left(\prod_{k=2}^{n-1} c_k(v_{k-1}, v_k, v_{k+1}) \right) \pi_c(v_1, v_2). \end{aligned} \tag{13}$$

where

$$\begin{aligned} & c_k(v_{k-1}, v_k, v_{k+1}) \\ & \quad = L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(x_k|v_k) L_{v_{k-1},v_{k+1};y_{k-1},y_{k+1}}(y_k|v_k) \pi(v_{k+1} \mid v_{k-1}, v_k), \end{aligned} \tag{14}$$

for $k = 2, \ldots, n-1$, and $v_j \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ for $j = 1, \ldots, n$.

A recursive algorithm for calculating the pseudo-likelihood is given by

$$h_2(v_1, v_2) = \pi_c(v_1, v_2),$$

$$h_j(v_{j-1}, v_j) = \sum_{v_{j-2}} c_{j-1}(v_{j-2}, v_{j-1}, v_j) h_{j-1}(v_{j-2}, v_{j-1}) / \bar{h}_{j-1}, \tag{15}$$

for $j = 3, \ldots, n$, where $\bar{h}_j$ is the average of $h_j(\cdot, \cdot, \cdot)$. The log pseudo-likelihood is then

$$\log L_p = \log \left(\sum_{v_{n-1}, v_n} h_n(v_{n-1}, v_n) \right) + \sum_{j=2}^{n-1} \log(\bar{h}_j). \tag{16}$$

The appearance of $\bar{h}_2, \ldots, \bar{h}_{n-1}$ in (15) and (16) may confuse the reader. Their entire purpose is to standardise the size of the terms in the recursion to prevent numerical underflow.

In practise, all the matrix exponentials in (12) are computed for rate parameters γ and $\tilde{\gamma}$, first. Then all the $c_k(\cdot, \cdot, \cdot)$ terms (15) are computed, and finally the recursion in (15) is carried out. The algorithm operates in linear time depending on sequence length.

Maximising the pseudo-likelihood provides the parameter estimates. However, this numerical maximisation is in practice not straight-forward since the number of parameters is large. A gradient method for the maximisation

would require both the first and second derivatives, but formulas for computing the derivatives seem not to be available for the pseudo-likelihood function (13). Alternatively, priors can be assigned to the parameters as considered in Lunter and Hein (2004), and the inference would not require maximisation of the high-dimensional approximate likelihood function, but is instead easily implemented as a MCMC simulation procedure. The prize to be paid is that for models with a high-dimensional parameter space as considered here, the specification of a high-dimensional prior may put restrictions on the parameters, which were not really intended. It may also hide undetected problems of parameter identifiability in the model, which is further discussed in Section 5 in relation to re-analysing the data analysed in Lunter and Hein (2004). An EM-algorithm for maximum pseudo-likelihood would avoid both types of problems mentioned above.

The accuracy of the pseudo-likelihood $\prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(x_k|v_k)$ for the sequence to sequence evolution was investigated in Christensen *et al.* (2005) by a simulation study, where it was seen that the parameter estimates obtained by maximising the pseudo-likelihood were almost identical to the maximum likelihood estimates. Only in extreme cases, i.e. very high neighbour-dependent rates, a noticeable difference was seen. Since (13) only involves the approximations of $L(\gamma; v, x)$ by $\prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(x_k|v_k)$ and $L(\gamma; v, y)$ by $\prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};y_{k-1},y_{k+1}}(y_k|v_k)$, the parameter estimated obtained from (13) should perform equally well.

4 EM-algorithm for maximum pseudo-likelihood estimation

In this section an EM-algorithm for maximising the pseudo-likelihood (13) is derived. The complete observation pseudo-likelihood is

$$L_c(\gamma, \tilde{\gamma}, \pi) = L_c(\gamma; e^{vx}) L_c(\tilde{\gamma}; e^{vy}) p^{\text{root}}(v; \pi), \tag{17}$$

where

$$L_c(\gamma; e^{vx}) = \prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(e_k^{vx}|v_k),$$

with $L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(e_k^{vx}|v_k)$ given by (12), and $L_c(\tilde{\gamma}; e^{vy})$ defined similarly. As noted in Christensen *et al.* (2005) the complete observation pseudo-likelihood for the sequence to sequence evolution is a complete observation likelihood for some model and data, i.e. here $L_c(\gamma; e^{vx})$ and $L_c(\tilde{\gamma}; e^{vy})$ are complete observation likelihoods. This implies that $L_c(\gamma, \tilde{\gamma}, \pi)$ in (17) is also a complete

observation likelihood function for some model and data, and therefore a corresponding EM-algorithm exists with the usual properties of EM-algorithms, i.e. improvement of the pseudo-likelihood (13) in each iteration and convergence to a local maximum.

An EM-algorithm iterates between an expectation (E) step and a maximisation (M) step. In the E-step

$$G((\gamma,\tilde{\gamma},\pi);(\gamma_0,\tilde{\gamma}_0,\pi_0)) = \mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[\log L_c(\gamma,\tilde{\gamma},\pi) \mid x,y]$$

is computed given the current parameter values, $(\gamma_0,\tilde{\gamma}_0,\pi_0)$. In the M-step, $G((\gamma,\tilde{\gamma},\pi);(\gamma_0,\tilde{\gamma}_0,\pi_0))$ is maximised as a function of $(\gamma,\tilde{\gamma},\pi)$. The E-step and the M-step for (17) are considered in Section 4.1 and Section 4.2, respectively.

4.1 The E-step

We note that (17) is the product of three terms where $L_c(\gamma;e^{vx})$ and $L_c(\tilde{\gamma};e^{vy})$ are the complete observation pseudo-likelihoods for the evolution on the v to x branch, and v to y branch, respectively. The complete observation pseudo-likelihood becomes

$$\begin{aligned} L_c(\gamma,\tilde{\gamma},\pi) = &\prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(e_k^{vx}|v_k) \\ &\times \prod_{k=2}^{n-1} L_{v_{k-1},v_{k+1};y_{k-1},y_{k+1}}(e_k^{vy}|v_k) \prod_{k=2}^{n-1} \pi(v_{k+1} \mid v_{k-1},v_k)\pi_c(v_1,v_2), \end{aligned}$$

where $L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(e_k^{vx}|v_k)$ is defined in (12), and $L_{v_{k-1},v_{k+1};y_{k-1},y_{k+1}}(e_k^{vy}|v_k)$ is defined similarly. By simplifying (12) and rearranging the terms in the formula above, we obtain an exponential family form

$$\begin{aligned} &L_c(\gamma,\tilde{\gamma},\pi) \\ &\quad= \exp\Bigg(\sum_{l,r}\sum_{a,b:a\neq b} \log\gamma(b;l,a,r)N_{l,r}(a,b) - \sum_{l,r}\sum_{a,b:a\neq b} \gamma(b;l,a,r)T_{l,r}(a) \\ &\qquad + \sum_{l,r}\sum_{a,b:a\neq b} \log\tilde{\gamma}(b;l,a,r)\tilde{N}_{l,r}(a,b) - \sum_{l,r}\sum_{a,b:a\neq b} \tilde{\gamma}(b;l,a,r)\tilde{T}_{l,r}(a) \\ &\qquad + \sum_{a_1,a_2,a_3} \log\pi(a_3 \mid a_1,a_2)N^{\mathrm{root}}(a_1,a_2,a_3) + \sum_{a_1,a_2} \log\pi_c(a_1,a_2)N^{12}(a_1,a_2) \Bigg), \end{aligned}$$

where $N_{l,r}(a,b)$ and $T_{l,r}(a)$ are the number of substitutions from a to b and the total time nucleotide a is present, respectively, with flanking nucleotides l,r

on the vx branch, $\tilde{N}_{l,r}(a,b)$ and $\tilde{T}_{l,r}(a)$ are defined similarly for the vy branch, $N^{\text{root}}(a_1,a_2,a_3)$ is the number of tri-nucleotides (a_1,a_2,a_3) in the root sequence v, and $N^{12}(a_1,a_2)$ is the indicator function for the first two nucleotides in the root sequence being (a_1,a_2). Since the pseudo-likelihood is derived by considering approximate evolutionary events for the flanking nucleotides, then $N_{l,r}(a,b) = N^1_{l,r}(a,b) + N^2_{l,r}(a,b)$, where $N^1_{l,r}(a,b)$ and $N^2_{l,r}(a,b)$ are the number of substitutions from a to b in $[0;1/2]$ with flanking nucleotides l,r in the v sequence, and the number of substitutions from a to b in $[1/2,1]$ with flanking nucleotides l,r in the x sequence, respectively. Similarly, $T_{l,r}(a) = T^1_{l,r}(a)+T^2_{l,r}(a)$ where $T^1_{l,r}(a)$ and $T^2_{l,r}(a)$ are the total time nucleotide a is present in $[0;1/2]$ with flanking nucleotides l,r in the v sequence, and the total time nucleotide a is present in $[1/2,1]$ with flanking nucleotides l,r in the x sequence, respectively, and $\tilde{N}_{l,r}(a,b)$ and $\tilde{T}_{l,r}(a)$ are expressed in the same manner.

From the exponential family form of the complete pseudo-likelihood above we obtain the E-step

$$\begin{aligned}
G((\gamma,\tilde{\gamma},\pi);(\gamma_0,\tilde{\gamma}_0,\pi_0)) &= \mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[\log L_c(\gamma,\tilde{\gamma},\pi) \mid x,y] \\
&= \sum_{l,r}\sum_{a,b:a\neq b} \log\gamma(b;l,a,r)\,\mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[N_{l,r}(a,b) \mid x,y] \\
&\quad - \sum_{l,r}\sum_{a,b:a\neq b} \gamma(b;l,a,r)\,\mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[T_{l,r}(a) \mid x,y] \\
&\quad + \sum_{l,r}\sum_{a,b:a\neq b} \log\tilde{\gamma}(b;l,a,r)\,\mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[\tilde{N}_{l,r}(a,b) \mid x,y] \\
&\quad - \sum_{l,r}\sum_{a,b:a\neq b} \tilde{\gamma}(b;l,a,r)\,\mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[\tilde{T}_{l,r}(a) \mid x,y] \\
&\quad + \sum_{a_1,a_2,a_3} \log\pi(a_3 \mid a_1,a_2)\,\mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[N^{\text{root}}(a_1,a_2,a_3) \mid x,y] \\
&\quad + \sum_{a_1,a_2} \log\pi_c(a_1,a_2)\,\mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[N^{12}(a_1,a_2) \mid x,y].
\end{aligned}$$

Expressions for the expectations in the formula above are given by (22)-(25) below. However, to derive (22)-(25) we use a different expression for $G((\gamma,\tilde{\gamma},\pi);(\gamma_0,\tilde{\gamma}_0,\pi_0))$ where we condition on the unobserved root sequence

v,

$$
\begin{aligned}
G((\gamma,\tilde{\gamma},\pi);(\gamma_0,\tilde{\gamma}_0,\pi_0)) &= \mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[\log L_c(\gamma,\tilde{\gamma},\pi) \mid x,y] \\
&\propto \sum_v \mathrm{E}_{(\gamma_0,\tilde{\gamma}_0,\pi_0)}[\log L_c(\gamma,\tilde{\gamma},\pi) \mid v,x,y] L(\gamma_0;v,x) L(\tilde{\gamma}_0;v,y) p^{\mathrm{root}}(v;\pi_0) \\
&= \sum_v \mathrm{E}_{\gamma_0}[\log L_c(\gamma;e^{vx}) \mid v,x] L(\gamma_0;v,x) L(\tilde{\gamma}_0;v,y) p^{\mathrm{root}}(v;\pi_0) \qquad (18) \\
&\quad + \sum_v \mathrm{E}_{\tilde{\gamma}_0}[\log L_c(\tilde{\gamma};e^{vy}) \mid v,y] L(\gamma_0;v,x) L(\tilde{\gamma}_0;v,y) p^{\mathrm{root}}(v;\pi_0) \qquad (19) \\
&\quad + \sum_v \log(p^{\mathrm{root}}(v;\pi)) L(\gamma_0;v,x) L(\tilde{\gamma}_0;v,y) p^{\mathrm{root}}(v;\pi_0). \qquad (20)
\end{aligned}
$$

Considering these terms, then $\mathrm{E}_{\gamma_0}[\log L_c(\gamma;e^{vx}) \mid v,x]$ and $\mathrm{E}_{\tilde{\gamma}_0}[\log L_c(\tilde{\gamma};e^{vy}) \mid v,y]$ are E-steps for a pseudo-likelihood describing the evolution of sequence v into sequences x and y, respectively. They can be derived as in Christensen *et al.* (2005) with the complication that the eigenvalue decompositions of substitution matrices now involve complex numbers; details are found in Appendix B. Substituting (36) from Appendix B we obtain that (18) equals

$$
\begin{aligned}
&\sum_{l^v,r^v} \sum_{a,b:a\neq b} \log\gamma(b;l^v,a,r^v) \sum_{k=2}^{n-1} \sum_{v_k} w^1_{(l^v,r^v,x_{k-1},x_{k+1})}(a,b;v_k,x_k) d_k(l^v,v_k,r^v) \\
&\quad - \sum_{l^v,r^v} \sum_{a,b:a\neq b} \gamma(b;l^v,a,r^v) \sum_{k=2}^{n-1} \sum_{v_k} w^1_{(l^v,r^v,x_{k-1},x_{k+1})}(a,a;v_k,x_k) d_k(l^v,v_k,r^v) \\
&+ \sum_{l^x,r^x} \sum_{a,b:a\neq b} \log\gamma(b;l^x,a,r^x) \sum_{k\in\mathcal{K}(l^x,r^x)} \sum_{l^v,r^v} \sum_{v_k} w^2_{(l^v,r^v,l^x,r^x)}(a,b;v_k,x_k) d_k(l^v,v_k,r^v) \\
&\quad - \sum_{l^x,r^x} \sum_{a,b:a\neq b} \gamma(b;l^x,a,r^x) \sum_{k\in\mathcal{K}(l^x,r^x)} \sum_{l^v,r^v} \sum_{v_k} w^2_{(l^v,r^v,l^x,r^x)}(a,a;v_k,x_k) d_k(l^v,v_k,r^v),
\end{aligned}
$$

and (19) equals

$$
\begin{aligned}
&\sum_{l^v,r^v} \sum_{a,b:a\neq b} \log\tilde{\gamma}(b;l^v,a,r^v) \sum_{k=2}^{n-1} \sum_{v_k} \tilde{w}^1_{(l^v,r^v,y_{k-1},y_{k+1})}(a,b;v_k,y_k) d_k(l^v,v_k,r^v) \\
&\quad - \sum_{l^v,r^v} \sum_{a,b:a\neq b} \tilde{\gamma}(b;l^v,a,r^v) \sum_{k=2}^{n-1} \sum_{v_k} \tilde{w}^1_{(l^v,r^v,x_{k-1},x_{k+1})}(a,a;v_k,y_k) d_k(l^v,v_k,r^v) \\
&+ \sum_{l^y,r^y} \sum_{A,b:a\neq b} \log\tilde{\gamma}(b;l^y,a,r^y) \sum_{k\in\mathcal{K}(l^y,r^y)} \sum_{l^v,r^v} \sum_{v_k} \tilde{w}^2_{(l^v,r^v,l^y,r^y)}(a,b;v_k,y_k) d_k(l^v,v_k,r^v) \\
&\quad - \sum_{l^y,r^y} \sum_{a,b:a\neq b} \tilde{\gamma}(b;l^y,a,r^y) \sum_{k\in\mathcal{K}(l^y,r^y)} \sum_{l^v,r^v} \sum_{v_k} \tilde{w}^2_{(l^v,r^v,l^y,r^y)}(a,a;v_k,y_k) d_k(l^v,v_k,r^v),
\end{aligned}
$$

where the arrays w_I^1 and w_I^2, and $\tilde{w}_I^1$ and $\tilde{w}_I^2$ are given in (37), (38), (39) and (40), for rate parameters γ_0 and $\tilde{\gamma}_0$, respectively, $\mathcal{K}(l^x, r^x) = \{k : (x_{k-1}, x_{k+1}) = (l^x, r^x)\}$, $\mathcal{K}(l^y, r^y) = \{k : (y_{k-1}, y_{k+1}) = (l^y, r^y)\}$, and $d_k(v_{k-1}, v_k, v_{k+1})$ denotes the summation of $L(\gamma_0; v, x)L(\tilde{\gamma}_0; v, y)p^{\text{root}}(v; \pi_0)$ over v with (v_{k-1}, v_k, v_{k+1}) fixed,

$$d_k(v_{k-1}, v_k, v_{k+1}) = \sum_{v_1,\ldots,v_{k-2}} \sum_{v_{k+2},\ldots,v_n} L(\gamma_0; v, x)L(\tilde{\gamma}_0; v, y)p^{\text{root}}(v; \pi_0), \qquad (21)$$

for $k = 2, \ldots, n-1$. The last term, (20), equals

$$\sum_{k=2}^{n-1} \sum_{v_{k-1},v_k,v_{k+1}} \log \pi(v_{k+1} \mid v_{k-1}, v_k) d_k(v_{k-1}, v_k, v_{k+1}) + \sum_{v_1,v_2} \log \pi_c(v_1, v_2) d_1(v_1, v_2),$$

where $d_1(v_1, v_2)$ denotes the summation of $L(\gamma_0; v, x)L(\tilde{\gamma}_0; v, y)p^{\text{root}}(v; \pi_0)$ over v with (v_1, v_2) fixed, and hence $d_1(v_1, v_2) = \sum_{v_3} d_2(v_1, v_2, v_3)$ where d_2 is defined in (21).

Rearranging the terms above we see that $G((\gamma, \tilde{\gamma}, \pi); (\gamma_0, \tilde{\gamma}_0, \pi_0))$ is on a simple form

$$\begin{aligned} &G((\gamma, \tilde{\gamma}, \pi); (\gamma_0, \tilde{\gamma}_0, \pi_0)) \\ &\quad = \sum_{l,r} \sum_{a,b:a\neq b} \log \gamma(b; l, a, r) w^{l,r}(a, b) - \sum_{l,r} \sum_{a,b:a\neq b} \gamma(b; l, a, r) w^{l,r}(a, a) \qquad (22) \\ &\qquad + \sum_{l,r} \sum_{a,b:a\neq b} \log \tilde{\gamma}(b; l, a, r) \tilde{w}^{l,r}(a, b) - \sum_{l,r} \sum_{a,b:a\neq b} \tilde{\gamma}(b; l, a, r) \tilde{w}^{l,r}(a, a) \\ &\qquad + \sum_{a_1,a_2,a_3} \log \pi(a_3 \mid a_1, a_2) w^{\text{root}}(a_1, a_2, a_3) + \sum_{a_1,a_2} \log \pi_c(a_1, a_2) d_1(a_1, a_2), \end{aligned}$$

where

$$\begin{aligned} w^{l,r}(a, b) &= \sum_{k=2}^{n-1} \sum_{v_k} w^1_{(l,r,x_{k-1},x_{k+1})}(a, b; v_k, x_k) d_k(l, v_k, r) \qquad (23) \\ &\quad + \sum_{k:(x_{k-1},x_{k+1})=(l,r)} \sum_{l^v,r^v} \sum_{v_k} w^2_{(l^v,r^v,l,r)}(a, b; v_k, x_k) d_k(l^v, v_k, r^v), \end{aligned}$$

$$\begin{aligned} \tilde{w}^{l,r}(a, b) &= \sum_{k=2}^{n-1} \sum_{v_k} \tilde{w}^1_{(l,r,y_{k-1},y_{k+1})}(a, b; v_k, y_k) d_k(l, v_k, r) \qquad (24) \\ &\quad + \sum_{k:(x_{k-1},x_{k+1})=(l,r)} \sum_{l^v,r^v} \sum_{v_k} \tilde{w}^2_{(l^v,r^v,l,r)}(a, b; v_k, x_k) d_k(l^v, v_k, r^v), \end{aligned}$$

for all a, b, and,

$$w^{\text{root}}(a_1, a_2, a_3) = \sum_{k=2}^{n-1} d_k(a_1, a_2, a_3), \tag{25}$$

for all (a_1, a_2, a_3).

To compute the d_k terms in (21) another recursion is needed. First, we note that the d_k's are on the form

$$d_k(v_{k-1}, v_k, v_{k+1}) \propto g_k(v_k, v_{k+1}) c_k(v_{k-1}, v_k, v_{k+1}) h_k(v_{k-1}, v_k)/\bar{h}_k, \tag{26}$$

where c_k is defined in (15), h_k is defined recursively in (15), and $\bar{h}_k$ is the average of $h_k(\cdot, \cdot, \cdot)$. The g_k terms are defined recursively, as $g_{n-1}(v_{n-1}, v_n) = 1$, and

$$g_{k-1}(v_{k-1}, v_k) = \sum_{v_{k+1}} c_k(v_{k-1}, v_k, v_{k+1}) g_k(v_k, v_{k+1})/\bar{h}_k, \tag{27}$$

for $k = n-1 \ldots, 2$. The proportionality constant in (26) equals $\prod_{k=2}^{n-1} \bar{h}_k$.

Note that the recursions (15) and (27) in fact are similar to the recursions used in the EM-algorithm for hidden Markov models (see Section 12.2.3 in Ewens and Grant, 2005).

To summarise the E-step. First, all the c_k terms in (15) are computed, and the recursion in (15) is carried out to compute the h_k and $\bar{h}_k$ terms. Second, the recursion in (27) is carried out to compute all the g_k terms, and this gives the d_k terms in (21). Third, the arrays $w_I^1(\cdot, \cdot; \cdot, \cdot)$, $w_I^2(\cdot, \cdot; \cdot, \cdot)$, $\tilde{w}_I^1(\cdot, \cdot; \cdot, \cdot)$ and $\tilde{w}_I^2(\cdot, \cdot; \cdot, \cdot)$ are computed for each of the $4^4 = 256$ possible flanking situations I. Finally, the weight matrices $w^{l,r}(\cdot, \cdot)$ and $\tilde{w}^{l,r}(\cdot, \cdot)$ in (22) are computed for the $4^2 = 16$ combinations of l, r.

4.2 The M-step

The maximisation of $G((\gamma, \tilde{\gamma}, \pi); (\gamma_0, \tilde{\gamma}_0, \pi_0))$ in (22) with respect to $(\gamma, \tilde{\gamma}, \pi)$ provides the parameter values for the next iteration in the EM-algorithm. First, the maximisation of (22) with respect to $\pi(a_3 \mid a_1, a_2)$ and $\pi_c(a_1, a_2)$ under the constraints which determine the parameter space in the general and in the strand-symmetric cases, respectively, is done using the `nlm` numerical maximisation routine implemented in `R` (R Development Core Team, 2005). Further details of the actual parameterisation used in the strand-symmetric case are given in Appendix A. Second, using a given model for the rate parameters γ and $\tilde{\gamma}$, the maximisation of (22) with respect to these parameters is done almost analytically. Some specific models for the rate parameters γ and $\tilde{\gamma}$ are considered in Appendix C.

5 Data analysis

A re-analysis the 100 kb intergenic human-mouse aligned sequence data from human chromosome 10 analysed in Lunter and Hein (2004) is made here. The models in Sections 2 are used, and parameter estimates are obtained by the EM-algorithm, where Appendix C shows the formulas for updating $(\gamma, \tilde{\gamma})$ for the models considered. The results in terms of value of log-pseudo-likelihood are reported in Tables 1 and 5. The former shows the results for the types of models containing all possible neighbour dependent rate parameters, whereas the latter shows the results for the simple models (4) with only neighbour dependent rate parameters corresponding to CpG to TpG and CpA substitutions.

Model	Param.	Description	$\log L_p$	$\hat{\tau}$
I	2*192+48	general	-222099.98	-
II	2*96+25	strand-sym.	-222196.04	-
III	192+1+48	same on branches	-222198.39	1.079
IV	96+1+25	strand-sym. + same on branches	-222265.06	1.084
V	2*84+48	di-nucl.	-222424.39	-
VI	2*48+25	di-nucl. + strand-sym.	-222469.19	-
VII	84+1+48	di-nucl. + same on branches	-222478.42	1.077
VIII	48+1+25	di-nucl. + strand-sym. + same	-222512.04	1.074

Table 1: Log-pseudo-likelihood and estimated value of τ parameter for eight different models. The number of parameters reported in the table refers to the sum of the number of free parameters in the substitution processes on the two branches and the number of free parameters in the model for the distribution of the root sequence.

Since these sequence data are not within any genes (intron or exon) we expect the strand-symmetric models to be appropriate for our analysis. The models I, III, V and VII in Table 1, and 1, 3, 5 and 7 in Table 5 do not assume strand-symmetry and the results for these models are mainly reported to demonstrate that the strand-symmetric models actually fit the data well. As expected, the values of the likelihoods are not largely increased for these models, when taking into account the larger number of parameters.

When considering the models III, IV, VII and VIII, which assume the same substitution process on the two branches, again the decrease in likelihood is not dramatic, and model IV therefore provides a good fit to the data (compared to the other models in Table 1). Such a result seems to contradict the conclusion in Hwang and Green (2004) that the substitution processes for the mouse

branch and the human branch differ, but we do note here that Hwang and Green (2004) consider 19 mammalian species, and our result may be an artefact from considering only human and mouse sequences. We also see from the $\hat{\tau}$ parameter estimates in (1) that the estimated mouse branch is slightly longer than the estimated human branch.

Considering model IV, Table 2 lists the largest substitution rates. It is clearly seen that the rates involving a CpG to TpG substitution (which is seen as a CpG to CpA substitution when it happens on the other strand) are much larger than the other rates. Hence, as also Lunter and Hein (2004) concluded, the data clearly demonstrate the CpG-methylation-deamination mutational process where a CpG becomes a TpG.

subst.	subst. compl.-strand	rates (IV)	rates (VIII)
CCG $\rightarrow$ CTG	CGG $\rightarrow$ CAG	0.9952	0.8603
ACG $\rightarrow$ ATG	CGT $\rightarrow$ CAT	0.8599	0.8598
GCG $\rightarrow$ GTG	CGC $\rightarrow$ CAC	0.7346	0.8598
TCG $\rightarrow$ TTG	CGA $\rightarrow$ CAA	0.6388	0.8598
AAC $\rightarrow$ AGC	GTT $\rightarrow$ GCT	0.2442	0.2012

Table 2: The five largest estimated substitution rate parameters for model IV. The second column shows the substitution pattern on the complementary strand, and the fourth column shows the parameter estimates for model VIII.

Considering instead the smallest estimated rate parameters for model IV, there are 31 different rate parameters which are approximately equal to zero, and therefore no table with such estimates is shown. Most of these rate parameters are related to either removal of a TpG or a CpA, or addition of a CpG, and they therefore represent substitutions in the opposite direction of the predominant CpG to TpG substitution pattern. These estimated zero rates may therefore reflect that parameters in the model are poorly identified. Both Lunter and Hein (2004) and Hwang and Green (2004) assign positive prior distributions to all substitution rate parameters and therefore obtain posterior means of these parameters which are strictly positive. Such Bayesian approach may be seen as a way to stabilise the inference procedure, but in this case it also hides possible problems with the model and data.

In Tables 3 and 4 we show the estimates of the root distribution parameters. Focusing on the CpG, TpG and CpA di-nucleotides, the estimated CpG di-nucleotide frequency, $\pi_c(\mathrm{C},\mathrm{G}) = 0.0888$ is high, and the estimated TpG di-nucleotide frequency $\pi_c(\mathrm{T},\mathrm{G}) = \pi_c(\mathrm{C},\mathrm{A}) = 0.0305$ is low. We also see that the estimates of the conditional frequency of G are high when resulting in a CpG di-nucleotide, i.e. $\pi(\mathrm{G} \mid \mathrm{A},\mathrm{C}) = 0.487$, and estimates of the conditional

a_1	a_2	$\pi(\mathrm{A} \mid a_1, a_2)$	$\pi(\mathrm{G} \mid a_1, a_2)$	$\pi(\mathrm{C} \mid a_1, a_2)$	$\pi(\mathrm{T} \mid a_1, a_2)$
A	A	0.363	0.133	0.193	0.310
A	G	0.363	0.249	0.159	0.229
A	C	0.144	0.487	0.212	0.157
A	T	0.387	0.094	0.164	0.355
G	A	0.369	0.155	0.231	0.245
G	G	0.337	0.225	0.192	0.247
G	C	0.126	0.483	0.235	0.155
G	T	0.307	0.107	0.241	0.344
C	A	0.323	0.197	0.203	0.278
C	G	0.218	0.244	0.221	0.317
C	C	0.140	0.436	0.225	0.199
C	T	0.266	0.151	0.237	0.346
T	A	0.346	0.109	0.184	0.361
T	G	0.330	0.228	0.168	0.274
T	C	0.166	0.319	0.276	0.238
T	T	0.324	0.095	0.216	0.363

Table 3: The root parameters $\pi(a_3 \mid a_1, a_2)$ for model IV.

a_1	$\pi_c(a_1, \mathrm{A})$	$\pi_c(a_1, \mathrm{G})$	$\pi_c(a_1, \mathrm{C})$	$\pi_c(a_1, \mathrm{T})$
A	0.103	0.040	0.058	0.090
G	0.061	0.050	0.041	0.058
C	0.031	0.089	0.050	0.040
T	0.097	0.031	0.061	0.103

Table 4: The root frequencies $\pi_c(a_1, a_2)$ for model IV.

frequency of T and A are small when resulting in a TpG or CpA di-nucleotide, respectively, i.e. $\pi(\mathrm{G} \mid \mathrm{A}, \mathrm{T}) = 0.094$. Such pattern of di-nucleotide frequencies at the root sequence seems unrealistic, and it is therefore tempting to conclude that this result and the result above about rate parameters being zero when they involve addition of a CpG or deletion of a TpG or CpA, may be due to parameters in the model being poorly identified when considering only two distantly related sequences. Implicitly Lunter and Hein (2004) avoid an increased CpG frequency at the root sequence by letting the root distribution equal the equilibrium distribution which has a decreased frequency of CpG.

Returning to Table 1, all the overlapping di-nucleotide models V, VI, VII and VIII present small pseudo-likelihood values, indicating a poor performance

of these models. From Table 2 we see that there are large differences in the parameter estimates for model IV and model VIII. In particular, the four substitution rates involving CpG to TpG or CpA substitutions are nearly identical under model VIII, whereas they clearly differ for model IV. I conclude that the neighbour dependent substitution rates depend on the nucleotides to the left and right in a non-additive way which violates model VIII. In particular, note the increased rate for CCG to CTG substitutions (seen as CGG to CAG substitution on the complementary strand) compared to the other three types of substitutions involving a CpG to CpT substitution.

Considering the results in Table 5, and comparing the log-likelihoods with the log-likelihoods in Table 1 we note large differences which show that the models (4) are too simple, and other important neighbour dependent mutational processes must exist than the ones related to the CpG-methylation-deamination process. I will not discuss this further here, but refer to Arndt and Hwa (2005) for an analysis of copies of AluSx Sines in the human genome where they discover several other important neighbour-dependent substitutional processes.

Model	Param.	Description	$\log L_p$	$\hat{\tau}$
1	2*20+48	general	-223744.18	-
2	2*10+25	strand-sym.	-223775.53	-
3	20+1+48	same on branches	-223759.56	1.074
4	10+1+25	strand-sym. + same on branches	-223788.08	1.092
5	2*14+48	di-nucl.	-223760.98	-
6	2*7+25	di-nucl. + strand-sym.	-223788.29	-
7	14+1+48	di-nucl. + same on branches	-223773.65	1.092
8	7+1+25	di-nucl. + strand-sym. + same	-223799.41	1.092

Table 5: Log-pseudo-likelihood and estimated value of τ parameter for eight simple models satisfying (4).

Comparing models 1-4 with models 5-8, again the overlapping di-nucleotide models do not perform very well, i.e. the difference in log-likelihood between models 4 and 8 is 11.33 which should be compared to $\chi^2(3)$-distribution and results in a p-value of 1%.

Considering model 4, Tables 6-8 list the parameter estimates. We observe that the estimated CpG to TpG/CpA types of substitution rates in Table 6 are much larger than the ones in Table 2, and thus see that the actual model has a significant impact on the estimates of the strength of the CpG-methylation-deamination process. Considering the neighbour indepen-

subst.	subst. compl.-strand	rates
CCG → CTG	CGG → CAG	1.9385
ACG → ATG	CGT → CAT	1.6507
GCG → GTG	CGC → CAC	1.6061
TCG → TTG	CGA → CAA	1.6415
A → G	T → C	0.1808
A → C	T → G	0.0607
A → T	T → A	0.0363
G → A	C → T	0.0000
G → C	C → G	0.0315
G → T	C → A	0.0090

Table 6: The neighbour dependent substitution rate parameters $\gamma(\mathrm{T}; l; \mathrm{C}, \mathrm{G})$ and neighbour independent substitution rate parameters $\epsilon(b; a)$ for model 4.

a_1	a_2	$\pi(\mathrm{A} \mid a_1, a_2)$	$\pi(\mathrm{G} \mid a_1, a_2)$	$\pi(\mathrm{C} \mid a_1, a_2)$	$\pi(\mathrm{T} \mid a_1, a_2)$
A	A	0.4196	0.1542	0.1408	0.2854
A	G	0.4322	0.1760	0.1823	0.2095
A	C	0.2617	0.3354	0.1694	0.2335
A	T	0.2787	0.1326	0.1619	0.4269
G	A	0.4374	0.1779	0.1571	0.2276
G	G	0.4105	0.1438	0.2124	0.2338
G	C	0.2189	0.2921	0.2108	0.2782
G	T	0.2105	0.1476	0.2225	0.4194
C	A	0.3594	0.2235	0.1496	0.2675
C	G	0.2398	0.2354	0.2040	0.3208
C	C	0.2463	0.3397	0.1433	0.2707
C	T	0.1636	0.1979	0.2262	0.4124
T	A	0.4433	0.1070	0.1236	0.3260
T	G	0.3919	0.1808	0.1620	0.2652
T	C	0.2730	0.1770	0.2100	0.3400
T	T	0.2534	0.1191	0.2080	0.4196

Table 7: The root parameters $\pi(a_3 \mid a_1, a_2)$ for model 4.

dent substitution rates in Table 6, we see that the estimated G to A substitution rate is equal to zero. Whether this result is a real phenomenon or some artefact of either the model or of considering only human and mouse sequences is unclear to me. From Tables 7 and 8 we still see a high frequency of CpG di-nucleotides, although not as extreme as in Tables 3 and 4.

a_1	$\pi_c(a_1, \mathrm{A})$	$\pi_c(a_1, \mathrm{G})$	$\pi_c(a_1, \mathrm{C})$	$\pi_c(a_1, \mathrm{T})$
A	0.136	0.051	0.046	0.091
G	0.065	0.033	0.033	0.046
C	0.045	0.048	0.033	0.051
T	0.078	0.045	0.065	0.136

Table 8: The root frequencies $\pi_c(a_1, a_2)$ for model 4.

Finally, with a number of rates being equal to zero I did observe some rate matrices on the edge of not being complex diagonalisable, and I had to check the values of the log-pseudo-likelihood in Table 1 and Table 5 by calculating the matrix exponentials using the Pade-algorithm instead of the eigenvalue decomposition.

6 Discussion

I have constructed a pseudo-likelihood method for inference in non-reversible nucleotide substitution models with neighbour dependent substitution rates. The method is computationally faster than the MCMC methods considered previously and will make data analysis using these models more practical. Extending the pseudo-likelihood method to a general phylogenetic tree is simple in principle, but for a large tree it requires an efficient algorithm for the summations of the unobserved nodes in the tree, similar to the extension of Felsenstein's pruning algorithm in Siepel and Haussler (2004). Also, the implementation presented here has not been optimised with respect to computational speed, which could be significantly increased by implementing the recursions in Sections 3 and 4.1 and the for-loops in Appendix B in a compiled language like `C` or `Fortran` instead of `R` which is known to be relatively slow for such types of computations. However such an implementation may require a significant investment of time since the current implementation benefits much from the build in routines in `R` for numerical optimisation, root finding and complex matrix algebra.

The general model (1) is very rich and contains a large number of parameters, and here I have considered a number of specific sub-models. I have demonstrated the poor performance of the overlapping di-nucleotide types of models used by Lunter and Hein (2004). I have also demonstrated that strand-symmetry is a feature that should be included in both the substitution process part of a model and the distribution of the root sequence.

Finally, the pseudo-likelihood method presented here for non-reversible nu-

cleotide models can in principle also be used for non-reversible codon models, although the large increase in computing time due to having 61 codons instead of 4 nucleotides should be noted. However, as discussed in Christensen *et al.* (2005), for coding sequences the main future challenge seems to be the development of non-reversible models with a small number of parameters more than the improvement of the inferential procedure.

Appendix A: The parameter space for sequence distribution at the root

Here the technical details of how the parameter space for the sequence distribution at the root looks like for the strand-symmetric case described in Section 2.2 are provided. The notation here follows the one in Section 2.2.

Lemma 1. *The parameter space given by (7), (8),* $\sum_{a_1,a_2} \pi_c(a_1,a_2) = 1$ *and* $\sum_{a_3} \pi(a_3 \mid a_1,a_2) = 1$ *is* 25 *dimensional and can be determined by having the* 7 *free parameters for* π_c *:*

$$\pi_c(A,A), \quad \pi_c(A,G), \quad \pi_c(A,C), \quad \pi_c(A,T), \quad \pi_c(G,A), \quad \pi_c(G,C), \quad \pi_c(C,A),$$

the 18 *free parameters for* π *:*

$$\pi(a_3 \mid a_1,a_2), \quad a_1 \in \{G,C,T\}, a_2 \in \{A,G\}, a_3 \in \{A,G,C\},$$

and the remaining parameters given by

$$\pi_c(T,T) = \pi_c(A,A), \quad \pi_c(C,T) = \pi_c(A,G), \quad \pi_c(G,T) = \pi_c(A,C),$$
$$\pi_c(T,C) = \pi_c(G,A), \quad \pi_c(T,G) = \pi_c(C,A), \tag{28}$$

$$\pi_c(T,A) = \pi_c(A,G) + \pi_c(A,C) + \pi_c(A,T) - \pi_c(G,A) - \pi_c(C,A) \tag{29}$$

$$\pi_c(C,G) = \pi_c(G,A) + \pi_c(G,C) + \pi_c(G,T) - \pi_c(A,G) - \pi_c(T,G) \tag{30}$$

$$\pi_c(G,G) = \frac{1 - \sum_{(a_1,a_2) \notin \{(C,C)(G,G)\}} \pi_c(a_1,a_2)}{2}, \quad \pi_c(C,C) = \pi_c(G,G) \tag{31}$$

$$\pi(T \mid a_1,a_2) = 1 - \sum_{a_3 \neq T} \pi(a_3 \mid a_1,a_2) \quad \textit{for} \quad a_1 \in \{G,C,T\}, a_2 \in \{A,G\}, \tag{32}$$

$$\pi(a_3 \mid a_1,a_2) = \frac{\pi_c(\mathtt{C}a_3,\mathtt{C}a_2)\pi(\mathtt{C}a_1 \mid \mathtt{C}a_3,\mathtt{C}a_2)}{\pi_c(a_1,a_2)}$$
$$\textit{for} \quad a_1 \in \{A,G,C,T\}, a_2 \in \{C,T\}, a_3 \in \{A,G,C\} \tag{33}$$

$$\pi(T \mid a_1, a_2) = 1 - \sum_{a_3 \neq T} \pi(a_3 \mid a_1, a_2) \;\; for \;\; a_1 \in \{A, G, C, T\}, a_2 \in \{C, T\}, \quad (34)$$

$$\pi(a_3 \mid A, a_2) = \frac{\pi_c(\mathtt{C}a_3, \mathtt{C}a_2)\pi(T \mid \mathtt{C}a_3, \mathtt{C}a_2)}{\pi_c(A, a_2)} \quad for \;\; a_2 \in \{A, G\}, a_3 \in \{A, G, C, T\} \quad (35)$$

Proof. A careful inspection of (28)-(35) shows that this system of equations defines all the parameters $\pi(a_3 \mid a_1, a_2)$ and $\pi_c(a_1, a_2)$ in an iterative way; i.e. there are no loops in this system.

First, we need to investigate whether all restrictions on the parameters π and π_c are satisfied. It is trivially seen that $\sum_{a_1,a_2} \pi_c(a_1, a_2) = 1$ and that $\sum_{a_3} \pi(a_3 \mid a_1, a_2) = 1$ for all $(a_1, a_2) \notin \{(\mathrm{A}, \mathrm{A}), (\mathrm{A}, \mathrm{G})\}$. We also easily see that both (7) and (8) are satisfied. What remains is to investigate whether $\sum_{a_3} \pi(a_3 \mid \mathrm{A}, \mathrm{A}) = 1$ and $\sum_{a_3} \pi(a_3 \mid \mathrm{A}, \mathrm{G}) = 1$. Here we use (35), (34), (33), (32), (28) and (31) to obtain

$$\begin{aligned}
&\sum_{a_3} \pi(a_3 \mid \mathrm{A}, a_2) \\
&\quad = \left(\sum_{a_3} \pi_c(\mathtt{C}a_3, \mathtt{C}a_2) - \sum_{\tilde{a}_3 \neq \mathrm{T}} \sum_{a_3} \pi(\tilde{a}_3 \mid \mathtt{C}a_3, \mathtt{C}a_2)\pi_c(\mathtt{C}a_3, \mathtt{C}a_2) \right) / \pi_c(\mathrm{A}, a_2) \\
&\quad = \left(\sum_{a_3} \pi_c(\mathtt{C}a_3, \mathtt{C}a_2) - \sum_{\tilde{a}_3 \neq \mathrm{T}} \sum_{a_3} \pi(a_3 \mid \mathtt{C}\tilde{a}_3, a_2)\pi_c(\mathtt{C}\tilde{a}_3, a_2) \right) / \pi_c(\mathrm{A}, a_2) \\
&\quad = \left(\sum_{a_3} \pi_c(\mathtt{C}a_3, \mathtt{C}a_2) + \pi_c(\mathrm{A}, a_2) - \sum_{\tilde{a}_3} \pi_c(\mathtt{C}\tilde{a}_3, a_2) \right) / \pi_c(\mathrm{A}, a_2) \\
&\quad = 1 + \left(\sum_{\tilde{a}} \pi_c(a_2, \tilde{a}) - \sum_{\tilde{a}} \pi_c(\tilde{a}, a_2) \right) / \pi_c(\mathrm{A}, a_2),
\end{aligned}$$

for $a_2 \in \{\mathrm{A}, \mathrm{G}\}$. Hence using (29) we see that $\sum_{a_3} \pi(a_3 \mid \mathrm{A}, \mathrm{A}) = 1$, and similarly using (30) we see that $\sum_{a_3} \pi(a_3 \mid \mathrm{A}, \mathrm{G}) = 1$, which completes the first half of the proof.

Second, we must show that all the equations (28)-(35) are really needed. Equations (28) and (31)-(35) are derived directly from the restrictions $1 = \sum_{a_1,a_2} \pi_c(a_1, a_2)$, $1 = \sum_{a_3} \pi(a_3 \mid a_1, a_2)$, (7) and (8), and the considerations in the first half of the proof shows the necessity of (29) and (30) to ensure that $\sum_{a_3} \pi(a_3 \mid \mathrm{A}, \mathrm{A}) = 1$ and $\sum_{a_3} \pi(a_3 \mid \mathrm{A}, \mathrm{G}) = 1$. This completes the proof. □

Appendix B: E-step for sequence to sequence likelihood

Here we consider the E-step for the pseudo-likelihood describing the sequence to sequence evolution. We assume that sequences $v = (v_1, \ldots, v_n)$ and $x = (x_1, \ldots, x_n)$ have been observed, and consider the evolution from v to x under the model (1) with substitution parameters γ. In the E-step we must calculate

$$G(\gamma; \gamma_0) = \mathrm{E}_{\gamma_0}[\log L_c(\gamma; e^{vx}) \mid v, x],$$

where $L_c(\gamma; e^{vx})$ is the complete observation pseudo-likelihood for the evolution from v to x.

The derivation follows Section 5 and Appendix C in Christensen *et al.* (2005) and we will therefore skip some of the details. Using that $L_c(\gamma; e^{vx})$ is on an exponential family form with a $2 \times (4 \times 3 + 4) \times 4^4 = 8192$-dimensional sufficient statistics

$$(N_I^1(a,b), T_I^1(a), N_I^1(a,b), T_I^1(a)) \quad : \quad a \neq b, \quad I = (l^v, r^v, l^x, r^x),$$

where $N_I^1(a,b)$ and $T_I^1(a)$ are the number of substitutions from a to b and the time spend in a, respectively, for flanking situation I and in the time interval $[0; 1/2]$, and $N_2^1(a,b)$ and $T_2^1(a)$ are defined similarly for the time interval $[1/2; 1]$, we have that

$$\begin{aligned} &G(\gamma; \gamma_0) = \mathrm{E}_{\gamma_0}[\log L_c(\gamma; e^{vx}) \mid v, x] \\ &= \sum_I \sum_{a,b:a\neq b} \log \gamma(b; l^v, a, r^v)\, \mathrm{E}_{\gamma_0}[N_I^1(a,b) \mid v, x] - \gamma(b; l^v, a, r^v)\, \mathrm{E}_{\gamma_0}[T_I^1(a) \mid v, x] \\ &+ \sum_I \sum_{a,b:a\neq b} \log \gamma(b; l^x, a, r^x)\, \mathrm{E}_{\gamma_0}[N_I^2(a,b) \mid v, x] - \gamma(b; l^x, a, r^x)\, \mathrm{E}_{\gamma_0}[T_I^2(a) \mid v, x], \end{aligned}$$

which in Christensen *et al.* (2005) corresponds to the first formula in Section 5.2.1 and formula (C3).

The formula above provides some intuition about the form of the E-step, but since I in Section 4.1 need to sum out the sequence v, I instead prefer a less compact formula here. The first formula in Section 5.2.1, and formula (C3), (C4) and (C5) in Christensen *et al.* (2005) viewed in the present context

(although the notation w_I^1 and w_I^2 has a different meaning here) become,

$$\begin{aligned} G(\gamma;\gamma_0) &= \mathrm{E}_{\gamma_0}[\log L_c(\gamma; e^{vx}) \mid v, x] \\ &= \sum_I \sum_{a,b:a\neq b} \log\gamma(b; l^v, a, r^v) \sum_{k:(v_{k-1},v_{k+1},x_{k-1},x_{k+1})=I} w_I^1(a,b;v_k,x_k) \\ &- \sum_I \sum_{a,b:a\neq b} \gamma(b; l^v, a, r^v) \sum_{k:(v_{k-1},v_{k+1},x_{k-1},x_{k+1})=I} w_I^1(a,a;v_k,x_k) \\ &+ \sum_I \sum_{a,b:a\neq b} \log\gamma(b; l^x, a, r^x) \sum_{k:(v_{k-1},v_{k+1},x_{k-1},x_{k+1})=I} w_I^2(a,b;v_k,x_k) \\ &- \sum_I \sum_{a,b:a\neq b} \gamma(b; l^x, a, r^x) \sum_{k:(v_{k-1},v_{k+1},x_{k-1},x_{k+1})=I} w_I^2(a,a;v_k,x_k) \end{aligned} \tag{36}$$

where $I = (l^v, r^v, l^x, r^x)$,

$$w_I^1(a,b;\tilde{a},\tilde{b}) = Q_0^{l^v,r^v}(a,b)\frac{\int_0^{1/2} P_0^I(0,t,\tilde{a},a)P_0^I(t,1,b,\tilde{b})dt}{P_0^I(0,1,\tilde{a},\tilde{b})}, \tag{37}$$

for $a \neq b$,

$$w_I^1(a,a;\tilde{a},\tilde{b}) = \frac{\int_0^{1/2} P_0^I(0,t,\tilde{a},a)P_0^I(t,1,a,\tilde{b})dt}{P_0^I(0,1,\tilde{a},\tilde{b})}, \tag{38}$$

$$w_I^2(a,b;\tilde{a},\tilde{b}) = Q_0^{l^x,r^x}(a,b)\frac{\int_{1/2}^1 P_0^I(0,t,\tilde{a},a)P_0^I(t,1,b,\tilde{b})dt}{P_0^I(0,1,\tilde{a},\tilde{b})}, \tag{39}$$

for $a \neq b$,

$$w_I^2(a,a;\tilde{a},\tilde{b}) = \frac{\int_{1/2}^1 P_0^I(0,t,\tilde{a},a)P_0^I(t,1,a,\tilde{b})dt}{P_0^I(0,1,\tilde{a},\tilde{b})}. \tag{40}$$

Here the 4×4 rate matrices $Q^{l,r}$ are defined in (10), and

$$P_0^I(0,1,\tilde{a},\tilde{b}) = \sum_c P_0^I(0,1/2,\tilde{a},c)P^I(1/2,1,c,\tilde{b}). \tag{41}$$

Assume that Q^{l^v,r^v} is complex diagonalisable with complex eigenvalues and complex valued eigenvectors. Let V be the matrix with eigenvectors as columns, D_λ the diagonal matrix of eigenvalues, and $W = V^{-1}$ the inverse of the eigenvector matrix. It follows that

$$P^I(0,t) = \exp(Q^{l^v,r^v}t) = V\exp(tD_\lambda)W, \;\; 0 \le t \le 1/2.$$

Similarly, assume that Q^{l^x,r^x} is complex diagonalisable, let U be the matrix with eigenvectors as columns, D_μ the diagonal matrix of eigenvalues, and $O = U^{-1}$ the inverse of the eigenvector matrix. Then we obtain

$$P^I(1/2, 1/2+t) = \exp(Q^{l^x,r^x}t) = U\exp(tD_\mu)O, \;\; 0 \le t \le 1/2.$$

Now we can find (41) and thereby the integrals in the formulas for w_I^1 and w_I^2.

Note that for $0 \le t \le 1/2$ we have

$$\begin{aligned} P^I(t,1,b,\tilde{b}) &= \sum_{\tilde{c}} P^I(t,1/2,b,\tilde{c})P^I(1/2,1,\tilde{c},\tilde{b}) \\ &= \sum_i V_{bi}e^{(1/2-t)\lambda_i}\sum_{\tilde{c}} W_{i\tilde{c}}\sum_m U_{cm}e^{\mu_m/2}O_{m\tilde{b}}, \end{aligned}$$

and we get

$$\begin{aligned} &\int_0^{1/2} P^I(0,t,\tilde{a},a)P^I(t,1,b,\tilde{b})dt \\ &= \int_0^{1/2}\sum_j V_{\tilde{a}j}\exp(t\lambda_j)W_{ja}\sum_i V_{bi}e^{(1/2-t)\lambda_i}\sum_c W_{ic}\sum_m U_{cm}e^{\mu_m/2}O_{m\tilde{b}}dt \\ &= \sum_j V_{\tilde{a}j}W_{ja}\sum_i V_{bi}J^1_{ji}\Big\{\sum_c W_{ic}\sum_m U_{cm}e^{\mu_m/2}O_{m\tilde{b}}\Big\}, \end{aligned} \tag{42}$$

where

$$J^1_{ji} = \begin{cases} \frac{1}{2}\exp(\lambda_i/2) & \text{if } \lambda_j = \lambda_i \\ (\exp(\lambda_j/2) - \exp(\lambda_i/2))/(\lambda_j - \lambda_i) & \text{if } \lambda_j \ne \lambda_i. \end{cases}$$

Similarly,

$$\begin{aligned} &\int_{1/2}^1 P^I(0,t,\tilde{a},a)P^I(t,1,b,\tilde{b})dt \\ &= \sum_j\Big\{\sum_c\sum_m V_{\tilde{a}m}e^{\lambda_m/2}W_{mc}U_{cj}\Big\}O_{ja}\sum_i U_{bi}J^2_{ji}O_{i\tilde{b}}, \end{aligned} \tag{43}$$

where

$$J^2_{ji} = \begin{cases} \frac{1}{2}\exp(\mu_i/2) & \text{if } \mu_j = \mu_i \\ (\exp(\mu_j/2) - \exp(\mu_i/2))/(\mu_j - \mu_i) & \text{if } \mu_j \ne \mu_i. \end{cases}$$

In the above equations, the reader should recognise that the formulas involve complex numbers. In the actual implementation, it was very convenient that `R` (R Development Core Team, 2005) handles matrix arithmetics for complex numbers.

Appendix C: The M-step for specific models

The maximisation of $G((\gamma, \tilde{\gamma}, \pi); (\gamma_0, \tilde{\gamma}_0, \pi_0))$ in (22) with respect to $(\gamma, \tilde{\gamma})$ is considered here for some specific models for the rate parameters γ and $\tilde{\gamma}$.

M-step for the general branch-specific substitution model

For the general model where γ and $\tilde{\gamma}$ vary freely, maximising (22) with respect to these two parameter vectors gives

$$\gamma(b; l, a, r) = w^{l,r}(a, b)/w^{l,r}(a, a),$$

$$\tilde{\gamma}(b; l, a, r) = \tilde{w}^{l,r}(a, b)/\tilde{w}^{l,r}(a, a),$$

for $B \neq A$.

M-step for the strand-symmetric branch-specific substitution model

For the strand-symmetric model, the constraints (2) hold both for γ and $\tilde{\gamma}$, and the parameters, $\gamma(b; l, a, r)$, $\tilde{\gamma}(b; l, a, r)$, $a \in \{\mathrm{A}, \mathrm{G}\}$, $b \neq a$, $l, r \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$, vary freely. The maximisation with respect to these parameters gives

$$\gamma(b; l, a, r) = (w^{l,r}(a, b) + w^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a, \mathtt{C}b))/(w^{l,r}(a, a) + w^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a, \mathtt{C}a)),$$

$$\tilde{\gamma}(b; l, a, r) = (\tilde{w}^{l,r}(a, b) + \tilde{w}^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a, \mathtt{C}b))/(\tilde{w}^{l,r}(a, a) + \tilde{w}^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a, \mathtt{C}a)),$$

for $b \neq a$.

M-step for models with the same substitution process on branches

For a model where the substitution process is the same on both branches $\tilde{\gamma} = \tau\gamma$, where $\tau > 0$ is the speed of evolution on the vy branch relative to the vx branch. The maximisation with respect to τ gives that the following equation should be satisfied

$$0 = (1/\tau) \sum_{a,B,l,r:a\neq b} \tilde{w}^{l,r}(a, b) - \sum_{a,b,l,r:a\neq b} \gamma(b; l, a, r)\tilde{w}^{l,r}(a, a)$$

For the general substitution model, the maximisation with respect to γ gives

$$\gamma(b; l, a, r) = (w^{l,r}(a, b) + \tilde{w}^{l,r}(a, b))/(w^{l,r}(a, a) + \tau\tilde{w}^{l,r}(a, a)).$$

Combining these two formulas and re-arranging the terms, we see that τ is found as the solution to the equation

$$0 = \sum_{a,b,l,r:a\neq b} w^{l,r}(a,b) - \frac{(w^{l,r}(a,b) + \tilde{w}^{l,r}(a,b))w^{l,r}(a,a)}{(w^{l,r}(a,a) + \tau\tilde{w}^{l,r}(a,a))},$$

which has to solved numerically. To solve such one-dimensional equations I throughout this paper have used the `R`-function `uniroot` (R Development Core Team, 2005).

Similarly, for the strand-symmetric model, the maximisation gives that τ is the solution to the equation

$$0 = \sum_{a,b,l,r:a\neq b} w^{l,r}(a,b) - \frac{(w^{l,r}(a,b) + \tilde{w}^{l,r}(a,b))(w^{l,r}(a,a) + w^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}a))}{w^{l,r}(a,a) + \tau\tilde{w}^{l,r}(a,a) + w^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}a) + \tau\tilde{w}^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}a)},$$

and γ is given by

$$\gamma(b;l,a,r) = \frac{w^{l,r}(a,b) + \tilde{w}^{l,r}(a,b) + w^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}b) + \tilde{w}^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}B)}{w^{l,r}(a,a) + \tau\tilde{w}^{l,r}(a,a) + w^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}a) + \tau\tilde{w}^{\mathtt{C}r,\mathtt{C}l}(\mathtt{C}a,\mathtt{C}a)},$$

for $a \in \{\mathrm{A},\mathrm{G}\}$, $b \neq a$.

M-step for the overlapping di-nucleotide models

For the overlapping di-nucleotide model (2), $\gamma(b;l,a,r) = \nu^{\mathrm{left}}(b;l,a)+\nu^{\mathrm{right}}(b;a,r)$ and $\tilde{\gamma}(b;l,a,r) = \tilde{\nu}^{\mathrm{left}}(b;l,a)+\tilde{\nu}^{\mathrm{right}}(b;a,r)$. Only the procedure for the general model where the parameters vary freely is stated here, but for the models with strand-symmetry and/or the same substitution process on both branches the derivations are not complicated either.

The maximisation with respect to $\nu^{\mathrm{left}}(b;l,a)$ gives that this parameter solves the equation

$$0 = \sum_r \frac{w^{l,r}(a,b)}{\nu^{\mathrm{left}}(b;l,a) + \nu^{\mathrm{right}}(b;a,r)} - \sum_r w^{l,r}(a,a), \tag{44}$$

given the other parameters. Similarly, the maximisation with respect to the parameter $\nu^{\mathrm{right}}(b;a,r)$ gives the equation

$$0 = \sum_l \frac{w^{l,r}(a,b)}{\nu^{\mathrm{left}}(b;l,a) + \nu^{\mathrm{right}}(b;a,r)} - \sum_l w^{l,r}(a,a). \tag{45}$$

For every a and b, the new parameter values of $\nu^{\text{left}}(b; l, a)$, $l \in \{\text{A}, \text{G}, \text{C}, \text{T}\}$ and $\nu^{\text{right}}(b; a, r)$, $r \in \{\text{A}, \text{G}, \text{C}, \text{T}\}$ are found using the following iterative procedure. Given the parameters $\nu^{\text{right}}(b; a, r)$, $r \in \{\text{A}, \text{G}, \text{C}, \text{T}\}$, (44) is solved for $\nu^{\text{left}}(b; l, a)$ for every l, and given the parameters $\nu^{\text{left}}(b; l, a)$, $l \in \{\text{A}, \text{G}, \text{C}, \text{T}\}$, (45) is solved for $\nu^{\text{right}}(b; a, r)$ for every r. The procedure is iterated until convergence.

A similar procedure is used to obtain the new $\tilde{\nu}^{\text{left}}(b; l, a)$ and $\tilde{\nu}^{\text{right}}(b; a, r)$ parameters.

M-step for the simple models with only neighbour dependent CpG to TpG and CpA substitutions

Considering the simple model (4) where the only neighbour dependent rate parameters are the ones involving CpG to TpG and CpA substitutions, then the procedure for the most general model is stated here. For the models with strand-symmetry and/or the same substitution process on both branches the derivations are not complicated either.

Maximising (22) with respect to $\gamma(\text{A}; \text{C}, \text{G}, r)$, $\gamma(\text{T}; l, \text{C}, \text{G})$, $\tilde{\gamma}(\text{A}; \text{C}, \text{G}, r)$, $\tilde{\gamma}(\text{T}; l, \text{C}, \text{G})$, $\epsilon(b; a)$, $\tilde{\epsilon}(b; a)$ gives

$$\gamma(\text{A}; \text{C}, \text{G}, r) = \frac{w^{\text{C},r}(\text{G}, \text{A})}{w^{\text{C},r}(\text{G}, \text{G})}, \quad \gamma(\text{T}; l, \text{C}, \text{G}) = \frac{w^{l,\text{G}}(\text{C}, \text{T})}{w^{l,\text{G}}(\text{C}, \text{C})},$$

$$\tilde{\gamma}(\text{A}; \text{C}, \text{G}, r) = \frac{\tilde{w}^{\text{C},r}(\text{G}, \text{A})}{\tilde{w}^{\text{C},r}(\text{G}, \text{G})}, \quad \tilde{\gamma}(\text{T}; l, \text{C}, \text{G}) = \frac{\tilde{w}^{l,\text{G}}(\text{C}, \text{T})}{\tilde{w}^{l,\text{G}}(\text{C}, \text{C})},$$

$$\epsilon(\text{A}; \text{G}) = \frac{\sum_{l \neq \text{C},r} w^{l,r}(\text{G}, \text{A})}{\sum_{l \neq \text{C},r} w^{l,r}(\text{G}, \text{G})}, \quad \tilde{\epsilon}(\text{A}; \text{G}) = \frac{\sum_{l \neq \text{C},r} \tilde{w}^{l,r}(\text{G}, \text{A})}{\sum_{l \neq \text{C},r} \tilde{w}^{l,r}(\text{G}, \text{G})},$$

$$\epsilon(\text{T}; \text{C}) = \frac{\sum_{l,r \neq \text{G}} w^{l,r}(\text{C}, \text{T})}{\sum_{l,r \neq \text{G}} w^{l,r}(\text{C}, \text{C})}, \quad \tilde{\epsilon}(\text{T}; \text{C}) = \frac{\sum_{l,r \neq \text{G}} \tilde{w}^{l,r}(\text{C}, \text{T})}{\sum_{l,r \neq \text{G}} \tilde{w}^{l,r}(\text{C}, \text{C})},$$

and

$$\epsilon(b; a) = \frac{\sum_{l,r} w^{l,r}(a, b)}{\sum_{l,r} w^{l,r}(a, a)}, \quad \tilde{\epsilon}(b; a) = \frac{\sum_{l,r} \tilde{w}^{l,r}(a, b)}{\sum_{l,r} \tilde{w}^{l,r}(a, a)},$$

for $b \neq a$ and $(a, b) \notin \{(\text{G}, \text{A}), (\text{C}, \text{T})\}$.

References

Arndt, P. F. and Hwa, T. (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**, 2322–2328.

Arndt, P. F., Burge, C. B. and Hwa, T. (2003a). DNA sequence evolution with neighbour-dependent mutation. *J. Comput. Biol.* **10**, 313–322.

Arndt, P. F., Petrov, D. and Hwa, T. (2003b). Distrinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **20**, 1887–196.

Christensen, O. F., Hobolth, A. and Jensen, J. L. (2005). Pseudo-likelihood analysis of codon substitution models with neighbor dependent rates. *J. Comput. Biol.* **12**, 1166–1182.

Duret, L. and Galtier, N. (2000). The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.* **17**, 1620–1625.

Ewens, W. J. and Grant, G. R. (2005). *Statistical methods in bioinformatics*. Springer, New York.

Galtier, N., Gascuel, O. and Jean-Marie, A. (2005). Markov models in molecular evolution. In: *Statistical Methods in Molecular Evolution* (ed. R. Nielsen), Springer, New York, 3–24.

Hobolth, A. (2006). A Monte Carlo expectation maximisation algorithm for statistical analysis of DNA sequence evolution with neighbour-dependent substitution rates. Submitted for publication.

Hwang, D. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *PNAS* **101**, 13994–14001.

Jensen, J. L. (2005). Context dependent DNA evolutionary models. *Research Report 458*, Department of Theoretical Statistics, Aarhus University.

Jensen, J. L. and Pedersen, A.-M. K. (2000). Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob.* **32**, 499–517.

Jojic, V., Jojic, N. Meek, C., Geiger, D., Siepel, A., Haussler, D. and Heckerman, D. (2004). Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics* **20**, i161–i168.

Lunter, G. and Hein, J. (2004). A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* **20**, i216–i223.

Pedersen, A.-M. K. and Jensen, J. L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* **18**, 763–776.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Schadt, E. and Lange, K. (2002). Codon and rate variation models in molecular phylogeny. *Mol. Biol. Evol.* **19**, 1534–1549.

Siepel, A. and Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488.

Yap, V. B. and Speed, T. P. (2005). Estimating substitution matrices. In: *Statistical Methods in Molecular Evolution* (ed. R. Nielsen), Springer, New York, 407–438.