

# A Markov Chain Monte Carlo Expectation Maximization Algorithm for Statistical Analysis of DNA Sequence Evolution with Neighbor-Dependent Substitution Rates

Asger HOBOLTH

The evolution of DNA sequences can be described by discrete state continuous time Markov processes on a phylogenetic tree. We consider neighbor-dependent evolutionary models where the instantaneous rate of substitution at a site depends on the states of the neighboring sites. Neighbor-dependent substitution models are analytically intractable and must be analyzed using either approximate or simulation-based methods. We describe statistical inference of neighbor-dependent models using a Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm. In the MCMC-EM algorithm, the high-dimensional integrals required in the EM algorithm are estimated using MCMC sampling. The MCMC sampler requires simulation of sample paths from a continuous time Markov process, conditional on the beginning and ending states and the paths of the neighboring sites. An exact path sampling algorithm is developed for this purpose.

**Key Words:** EM-algorithm; Gibbs sampling; Likelihood inference; Molecular evolution; Neighbor-dependence; Path sampling.

## 1. INTRODUCTION

A fundamental task in modern molecular genetics is to gain insight into the evolutionary forces that act on DNA and protein sequences. The analysis is often based on homologous sequence data that have been obtained from the increasing number of publicly available bacterial, archael, eukaryotic, and viral genomes. Over the past 25 years, sophisticated statistical models and inferential procedures have been developed to describe and analyze homologous sequence data.

The evolution of homologous DNA sequences can be described by discrete state continuous time Markov chains on a phylogenetic tree. These continuous time Markov chains are characterized by a substitution rate matrix and a phylogenetic tree that specifies the

---

Asger Hobolth, Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh NC 27695-7566 (E-mail: [asger@daimi.au.dk](mailto:asger@daimi.au.dk)).

© 2008 American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America

*Journal of Computational and Graphical Statistics*, Volume 17, Number 1, Pages 1–25  
DOI: 10.1198/106186008X289010

relationship between the species being considered. The phylogenetic tree also specifies the expected amount of sequence evolution on each branch of the tree. The DNA sequences are observed only in the leaves, and information on substitution events (time and type) is missing. The statistical problem is to draw inference about a discrete state continuous time Markov chain on a phylogenetic tree from data observed in the leaves only. Note that a special case of this problem is to draw inference from a partially observed discrete state continuous time Markov chain.

If we assume that each site in the DNA sequence evolves independently, the size of the state space is four because the four nucleotide types are A, G, C, and T, and the Markov chain is described by a  $4 \times 4$  substitution rate matrix  $Q$ . In order to estimate the rate parameters and branch lengths from the marginal likelihood, one needs the transition probability matrix  $P(t) = \exp(Qt)$ . For protein-coding DNA, the state space is of size 61 because there are 61 sense codons.

The expectation maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) is useful in situations where finding the maximum likelihood estimate based on the full data is analytically tractable, but solving the problem based on the observed data is more complicated. Holmes and Rubin (2002), Siepel and Haussler (2004), and Yap and Speed (2005) applied the EM algorithm for estimating rate matrices from homologous DNA sequences under the independent sites assumption. Hobolth and Jensen (2005a) used the results from Louis (1982) to derive an expression of the information matrix. The EM algorithm for partially observed discrete state continuous time Markov chains has also been described by Bladt and Sørensen (2005).

It is widely accepted that sites in a DNA sequence do not evolve independently (e.g., Blake, Hess, and Nicholson-Tuell 1992; Hess, Blake and Blake 1994), but only in recent years has the independence assumption been relaxed. Relaxation of the independence assumption leads to state spaces of size  $4^n$  (or  $61^n$ ), where  $n$  is the length of the sequence. The sequence length is usually well above 100, and the transition probability matrix cannot be computed in practice. Context-dependent models of DNA sequence evolution must therefore be analyzed using either simulation-based or approximative procedures.

In this article the independent sites EM algorithm is extended in order to facilitate statistical inference in context dependent models of homologous DNA sequences. Relaxing the independent sites assumption means that the conditional means in the E-step of the EM-algorithm are no longer analytically tractable. However, the conditional means can be approximated using Markov chain Monte Carlo (MCMC) sampling. The MCMC sampler requires simulation of sample paths from a continuous time Markov process, conditional on the beginning and ending states and the paths of the neighboring sites. Such sample paths can be achieved using rejection sampling, but in order to obtain faster convergence of the resulting MCMC-EM algorithm (Wei and Tanner 1990), a novel exact path sampling algorithm for simulating sample paths from a continuous time Markov chain conditional on the beginning and ending states is derived.

Several recent studies have analyzed context-dependent evolutionary models of DNA sequences. Lunter and Hein (2004), Arndt and Hwa (2005), and Christensen, Hobolth, and Jensen (2005) analyzed neighbor-dependent models using pairs of sequences and approx-

imate maximum likelihood methods. Siepel and Haussler (2004) also analyzed neighbor-dependent models using approximate maximum likelihood, but consider multiple sequences. Hwang and Green (2004) applied a Bayesian MCMC approach to derive neighbor-dependent substitution patterns for multiple sequences. Robinson, Jones, Kishino, Goldman, and Thorne (2003) analyzed context-dependent models using pairs of sequences and Bayesian MCMC methods. In Robinson et al. (2003) the substitution rates depend not only on the nearest neighbors, but the three-dimensional protein structure is also taken into account. This article is close in spirit to Hwang and Green (2004). We also consider multiple sequences, but use maximum likelihood inference and avoid discretization of branch lengths by using the exact path sampling algorithm. Furthermore, the stationary distribution of the model is available, and this feature allows a detailed analysis of one sequence only. For more information on current methodology for neighbor-dependent models, see the review by Jensen (2005).

This article is organized as follows. In Section 2, we first motivate the need for relaxing the independent sites assumption by analyzing the stationary distribution of a single human noncoding DNA sequence under the independent sites model. Second, the neighbor-dependent model is formulated and it is shown that the stationary distribution of the neighbor-dependent model adequately describes the DNA sequence. Details of the analysis are described in the Appendix. In Section 3, the Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm is described for pairwise sequences. The full likelihood of the neighbor-dependent model is analytically tractable so that the M-step is easy to carry out. The E-step must be done using Markov chain Monte Carlo sampling and amounts to simulating a sample path from a discrete state continuous time Markov chain conditional on the beginning and ending states. Exact simulation of such sample paths is described in Section 4, and in Section 5 the MCMC-EM algorithm is extended to multiple sequences. Finally, we discuss extensions of the neighbor-dependent model and other potential applications of the exact path sampling algorithm.

## 2. NEIGHBOR-DEPENDENT NUCLEOTIDE MODELS

### 2.1 DATA AND MOTIVATION

Perhaps the most well-known example of violation of the independence assumption is the increased substitution rate of C to T in CpG dinucleotides in vertebrates (Albert et al, 2002, p. 434). The process is presumably due to methylation of cytosine in CpG followed by deamination and substitution from CpG to TpG or CpA (on the reverse strand). The CpG-methylation-deamination process leaves vertebrates with a remarkable deficiency of CpG dinucleotides. In Section 5, we analyze a multiple alignment of 741 sites and five species (human, chimpanzee, orangutan, mouse, and rat) from noncoding DNA. Table 1 summarizes the human DNA sequence in terms of a Markov chain along the sequence. The observed nucleotide counts violate the independence assumption and motivate the study of context dependent substitution processes.

The evolution of noncoding DNA is often described as a stationary homogeneous time

Table 1. The observed human noncoding DNA sequence summarized in terms of a Markov chain along the sequence. Presumably due to the increased rate of C to T substitutions in CpG dinucleotides, the observed count of CpG dinucleotides is much smaller than expected under the independent sites assumption. The residuals, defined as  $\text{residual}_i = (\text{observed}_i - \text{expected}_i) / \sqrt{\text{expected}_i}$ , confirm that the CpG cell shows the largest deviation from independence. Pearson's chi-squared test statistic (the sum of the squared residuals) is 36.13 and the  $p$  value for the independence assumption is  $3.7 \cdot 10^{-5}$  with 9 degrees of freedom. Thus, the independent sites assumption is violated, and this is mainly due to the CpG cell that accounts for more than 2/5 ( $3.84^2 = 14.75$  out of 36.13) of the total test statistic.

	Observed				Expected				Residuals			
	A	G	C	T	A	G	C	T	A	G	C	T
A	62	46	40	81	71	40	42	76	-1.05	0.91	-0.27	0.56
G	47	25	21	36	40	23	24	43	1.12	0.49	-0.52	-1.05
C	57	<b>5</b>	34	39	42	<b>24</b>	25	45	2.36	<b>-3.84</b>	1.89	-0.88
T	63	54	40	90	76	43	45	82	-1.54	1.61	-0.75	0.87

reversible continuous time Markov process. Assume that sites are independent and consider the general time reversible (GTR) model with rate matrix (e.g., Yap and Speed 2004)

$$Q = \begin{bmatrix} \cdot & \theta_{AG}\pi_G & \theta_{AC}\pi_C & \theta_{AT}\pi_T \\ \theta_{AG}\pi_A & \cdot & \theta_{GC}\pi_C & \theta_{GT}\pi_T \\ \theta_{AC}\pi_A & \theta_{GC}\pi_G & \cdot & \theta_{CT}\pi_T \\ \theta_{AT}\pi_A & \theta_{GT}\pi_G & \theta_{CT}\pi_C & \cdot \end{bmatrix}. \quad (2.1)$$

Here the off-diagonal entries, the instantaneous rates of substitutions, are all non-negative, and the diagonal elements are such that each row sums to zero. We can write the rate matrix as  $Q = S \text{diag}(\pi)$ , where

$$S = \begin{bmatrix} \cdot & \theta_{AG} & \theta_{AC} & \theta_{AT} \\ \theta_{AG} & \cdot & \theta_{GC} & \theta_{GT} \\ \theta_{AC} & \theta_{GC} & \cdot & \theta_{CT} \\ \theta_{AT} & \theta_{GT} & \theta_{CT} & \cdot \end{bmatrix}$$

is a symmetric matrix and  $\text{diag}(\pi)$  is the diagonal matrix with  $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$  on the diagonal. We observe that the detailed balance condition  $\text{diag}(\pi) Q = Q^* \text{diag}(\pi)$  is fulfilled, where superscript  $*$  denotes vector transpose, and thus  $\pi$  is the stationary distribution.

We now move from the independent sites GTR model to the neighbor-dependent model. A change in the nucleotide sequence  $x = (x_1, \dots, x_n)$  consists of a change of one nucleotide only and the rate matrix is no longer a  $4 \times 4$  matrix, but a  $4^n \times 4^n$  matrix. Consider the rate from sequence  $x$  to sequence  $\tilde{x}$ , where  $x$  and  $\tilde{x}$  are the same except at position  $j$ . The new nucleotide is denoted  $\tilde{x}_j$ . The rate from  $x$  to  $\tilde{x}$  is determined by two main components.

First, there is the  $4 \times 4$  substitution rate matrix  $Q$ , where the rates do not depend on the neighboring codons. This component corresponds to the model one would use had there been no interaction among neighboring nucleotides. We assume that the site-independent

part of the model is reversible with stationary distribution  $\pi$  such that detailed balance  $\text{diag}(\pi)Q = Q^*\text{diag}(\pi)$  is fulfilled.

Second, there is a CpG component, determined by the parameter  $\lambda$ , that introduces dependence among nucleotides. If  $\lambda < 1$ , the component introduces higher substitution rates from CpG dinucleotides. If  $\lambda > 1$ , the component introduces lower substitution rates, and if  $\lambda = 1$  there is no CpG effect. Consider the triplet of adjacent nucleotides  $(y_1, y_2, y_3)$ , and suppose  $y_2$  undergoes a change. If  $(y_1, y_2)$  or  $(y_2, y_3)$  are CpG dinucleotides and  $\lambda < 1$  ( $\lambda > 1$ ), the substitution rate for a change should increase (decrease), and if  $\lambda = 1$  the substitution rate should remain unchanged. We therefore define the function

$$\begin{aligned} R(y_1, y_2, y_3) &= (1/\lambda)^{1_{(C,G)}(y_1,y_2)+1_{(C,G)}(y_2,y_3)} \\ &= \begin{cases} 1/\lambda & \text{if } (y_1, y_2) = (C, G) \text{ or } (y_2, y_3) = (C, G) \\ 1 & \text{otherwise,} \end{cases} \end{aligned} \quad (2.2)$$

which takes the value  $1/\lambda$  if  $y_2$  is a member of a CpG pair, and takes the value 1 otherwise. The indicator function  $1_{(C,G)}(a, b)$  is one if  $a = C$  and  $b = G$ , and zero otherwise.

The substitution rate  $\gamma$  for a change from sequence  $x$  to sequence  $\tilde{x}$  thereby depends on  $x_j, \tilde{x}_j$ , and the neighboring pairs  $x_{j-1}$  and  $x_{j+1}$ , and is given by

$$\gamma(\tilde{x}_j; x_{j-1}, x_j, x_{j+1}) = Q(x_j, \tilde{x}_j)R(x_{j-1}, x_j, x_{j+1}). \quad (2.3)$$

When  $Q$  is the rate matrix from the GTR model, the neighbor-dependent model is termed the GTR+CpG model. Note that  $\lambda = 1$  implies  $R(x_{j-1}, x_j, x_{j+1}) = 1$ , and the rate from  $x_j$  to  $\tilde{x}_j$  becomes  $Q(x_j, \tilde{x}_j)$ , which does not depend on the neighboring nucleotides. Thus, the independent sites GTR model is nested in the GTR+CpG model.

A nice feature of the GTR+CpG model is that the stationary distribution can be found. As can be proved from detailed balance on the sequence level, the stationary distribution is given by

$$P(x) = \frac{1}{Z(\lambda, \pi)} \left( \prod_{j=0}^{n-1} \lambda^{1_{(C,G)}(x_j, x_{j+1})} \pi_{x_{j+1}} \right) \lambda^{1_{(C,G)}(x_n, x_{n+1})}, \quad (2.4)$$

where  $Z(\lambda, \pi)$  is a normalizing constant and  $x_0$  and  $x_{n+1}$  are fixed flanking nucleotides.

We can use this expression for the stationary distribution to analyze the CpG effect. Indeed, we can estimate the parameters  $\lambda$  and  $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$  from a single sequence using, for example, maximum likelihood, and if  $\lambda$  is significantly smaller than 1 we may conclude that the CpG-methylation-deamination process has played a role during the evolution of the sequence. This is the topic for the next subsection.

## 2.2 ANALYSIS OF THE STATIONARY DISTRIBUTION

Define the  $4 \times 4$  matrix  $A$  with entries  $A(a, b) = \lambda^{1_{(C,G)}(a,b)} \pi_b$ . Appendix A.2 shows that, for long sequences, the normalizing constant can be well approximated by

$$Z(\lambda, \pi) \approx \mu_1^n \sum_a l_1(a),$$

Table 2. Observed and expected dinucleotide counts for the human noncoding DNA sequence. The expected counts are found using Equation (2.5). Pearson’s chi-squared test statistic is 13.47 and the  $p$  value for the stationary distribution of the GTR+CpG model is 0.097 with 8 degrees of freedom. Thus, the stationary distribution of the GTR+CpG model provides a reasonable description of the human DNA sequence.

	Observed				Expected				Residuals			
	A	G	C	T	A	G	C	T	A	G	C	T
A	62	46	40	81	68	47	40	73	-0.73	-0.17	-0.02	0.89
G	47	25	21	36	39	27	23	42	1.35	-0.34	-0.37	-0.88
C	57	5	34	39	49	5	29	52	1.20	0.00	1.00	-1.86
T	63	54	40	90	73	51	43	79	-1.21	0.44	-0.50	1.22

where  $\mu_1$  is the largest eigenvalue of  $A$  and  $l_1$  is the corresponding left eigenvector. This expression allows us to easily numerically find the maximum likelihood estimates (MLEs) of  $\lambda$  and  $\pi$  from (2.4). The MLEs are

$$\hat{\lambda} = 0.148, \quad \text{and} \quad \hat{\pi} = (\hat{\pi}_A, \hat{\pi}_G, \hat{\pi}_C, \hat{\pi}_T) = (0.287, 0.199, 0.205, 0.309)$$

and the maximum log-likelihood is  $-982.13$ . This value is significantly larger than the log-likelihood  $-996.40$  obtained under the independent sites model where  $\lambda = 1$  and  $\pi$  is given by the observed frequencies of A, G, C, and T. The likelihood ratio test statistic is  $2(996.40 - 982.13) = 28.54$  with  $5 - 4 = 1$  degree of freedom. Under the  $\chi^2(1)$  approximation of the test-statistic, the  $p$  value is  $9 \cdot 10^{-8}$ , indicating that the independent sites assumption is inadequate. In the neighbor-dependent model, the estimated value of  $\lambda = 0.148$ , and thus  $1/\lambda = 6.74$ . Therefore, the CpG component  $R$  of the substitution rate from nucleotide  $x_j$  to  $\tilde{x}_j$  is almost seven times higher if  $x_j$  is a member of a CpG pair than if  $x_j$  is not a member of a CpG pair (recall equation (2.2) and the basic model (2.3)).

Appendix A.4 derives an expression of the expected number of dinucleotides. The expected number  $E[n_{(a,b)}]$  of  $(a, b)$  dinucleotides is well approximated by

$$E[n_{(a,b)}] \approx (n-1)l_1(a)A(a,b)r_1(b)/\mu_1, \quad (2.5)$$

where  $l_1$  is the left eigenvector and  $r_1$  is the right eigenvector corresponding to the largest eigenvalue  $\mu_1$  of  $A$ . Table 2 provides a summary of how the stationary distribution of the GTR+CpG model fits the human noncoding DNA sequence. The stationary distribution of the GTR+CpG model fits the human noncoding DNA sequence much better than the independence model in Table 1.

It is worth emphasizing that it is possible to extend the GTR+CpG model to take other neighbor dependencies into account. The residuals in Table 2 naturally lend themselves for such a purpose. Diaconis and Rolles (2006), in a Bayesian setting, also modeled single DNA sequences as a Markov chain along the sequence.

### 3. FULL LIKELIHOOD AND THE MCMC-EM ALGORITHM FOR PAIRWISE SEQUENCES

Consider the situation where the sequence  $x(t) = (x_1(t), \dots, x_n(t))$  is fully observed in the time period  $0 \leq t \leq T$ , and suppose the changes in the sequence occur at times  $t_1 < t_2 < \dots < t_M$  and positions  $j_1, \dots, j_M$ . Denote the full data  $x = \{x(t) : 0 \leq t \leq T\}$ . The waiting time in sequence  $x(t)$  is exponentially distributed with parameter

$$\Gamma_\theta(t) = \sum_{j=1}^n \sum_{\tilde{x}_j \neq x_j(t)} \gamma_\theta(\tilde{x}_j; x_{j-1}(t), x_j(t), x_{j+1}(t)) \quad (3.1)$$

and the rate for a change from sequence  $x(t_m)$  to sequence  $x(t_{m+1})$  is given by

$$\gamma_\theta(t_{m+1}) = \gamma_\theta(x_{j_{m+1}}(t_{m+1}); x_{j_{m+1}-1}(t_m), x_{j_{m+1}}(t_m), x_{j_{m+1}+1}(t_m)).$$

The full likelihood thereby takes the form

$$L_\theta(x) = \left( \prod_{m=0}^{M-1} e^{-\Gamma_\theta(t_m)(t_{m+1}-t_m)} \gamma_\theta(t_{m+1}) \right) e^{-\Gamma_\theta(t_M)(T-t_M)}, \quad (3.2)$$

where  $t_0 = 0$ . With the notation  $t_{M+1} = T$ , the full log-likelihood is given by

$$\log L_\theta(x) = \sum_{m=0}^{M-1} \log \gamma_\theta(t_{m+1}) - \sum_{m=0}^M \Gamma_\theta(t_m)(t_{m+1} - t_m). \quad (3.3)$$

Despite the somewhat complicated expression for the waiting times and rates, the full likelihood is actually surprisingly simple. As an illustration of the simplicity of the likelihood, consider the following example.

#### 3.1 EXAMPLE: K80+C<sub>P</sub>G MODEL

In order to illustrate the simplicity of the likelihood (3.3) and the idea behind the MCMC-EM algorithm, we consider the following situation. The Kimura (1980) model is a special case of the GTR model (2.1). The model gives one rate  $\alpha = \theta_{AG} = \theta_{CT}$  to *transitions* (substitutions within purines (A, G) or pyrimidines (C, T)), and another rate  $\beta = \theta_{AC} = \theta_{AT} = \theta_{GC} = \theta_{GT}$  to *transversions* (substitutions between purines and pyrimidines). Furthermore, the model has a uniform stationary distribution  $\pi_A = \pi_G = \pi_C = \pi_T = 1/4$ . Suppose sequence  $x(t)$  evolves according to the K80+C<sub>P</sub>G model and is fully observed from time  $t = 0$  to time  $t = 1$ . The parameter  $\lambda$  can be estimated from the stationary distribution of  $x(0)$  as described in Section 2.2. We now describe how to estimate the two remaining parameters  $\alpha$  and  $\beta$ .

The waiting times in the K80+C<sub>P</sub>G model are determined by

$$\begin{aligned} \Gamma_{\alpha,\beta}(t) &= 2n_{\text{CpG}}(t)(\alpha + 2\beta)/(4\lambda) + (n - 2n_{\text{CpG}}(t))(\alpha + 2\beta)/4 \\ &= (\alpha + 2\beta) \left( 2n_{\text{CpG}}(t)/\lambda + n - 2n_{\text{CpG}}(t) \right) / 4, \end{aligned}$$

where  $n_{\text{CpG}}(t)$  is the total number of CpG dinucleotides in sequence  $x(t)$ . The log-likelihood is, up to an additive constant, given by

$$\log L_{\alpha,\beta}(x) = n_{\text{ts}} \log \alpha + n_{\text{tv}} \log \beta - \sum_{m=0}^M \Gamma_{\theta}(t_m)(t_{m+1} - t_m),$$

where  $n_{\text{ts}}$  and  $n_{\text{tv}}$  denote the number of transitions and transversions. Let

$$\bar{n}_{\text{CpG}} = \sum_{m=0}^M (t_{m+1} - t_m) n_{\text{CpG}}(t_m)$$

be the weighted average of CpG dinucleotides. Then the log-likelihood takes the form

$$\log L_{\alpha,\beta}(x) = n_{\text{ts}} \log \alpha + n_{\text{tv}} \log \beta - (\alpha + 2\beta) \left( 2\bar{n}_{\text{CpG}}/\lambda + n - 2\bar{n}_{\text{CpG}} \right) / 4, \quad (3.4)$$

and the full likelihood is maximized for

$$\hat{\alpha} = \frac{4n_{\text{ts}}}{(2\bar{n}_{\text{CpG}}/\lambda + n - 2\bar{n}_{\text{CpG}})} \quad \text{and} \quad \hat{\beta} = \frac{4n_{\text{tv}}}{2(2\bar{n}_{\text{CpG}}/\lambda + n - 2\bar{n}_{\text{CpG}})}. \quad (3.5)$$

Thus, the likelihood based on a complete observation of the K80+CpG model is easy to analyze. The sufficient statistics are the total number of transitions and transversions and the weighted average of CpG dinucleotides, and the MLEs are simple functions of the sufficient statistics.

The EM algorithm is attractive in situations where finding the maximum likelihood estimate (MLE)  $\hat{\psi}$  based on the full data is analytically tractable, but finding the MLE based on the observed data is a more complicated problem. The algorithm is an iterative procedure. In the E-step, the conditional mean of the full log-likelihood

$$G(\psi; \psi_{(s-1)}) = E_{\psi_{(s-1)}}[\log L_{\psi}(x)|y] \quad (3.6)$$

is calculated conditional on the observed data  $y = y(x)$ . In the M-step, a new parameter value  $\psi_s$  is obtained as the value of  $\psi$  that maximizes  $G(\psi; \psi_{(s-1)})$ .

Consider the K80+CpG model from the previous example and suppose we observe only the beginning and ending sequences  $x(0)$  and  $x(T)$ . In the E-step we need to calculate the three conditional means

$$E[n_{\text{ts}}|x(0), x(T)], \quad E[n_{\text{tv}}|x(0), x(T)], \quad \text{and} \quad E[\bar{n}_{\text{CpG}}|x(0), x(T)]$$

for given parameter values  $\alpha, \beta$ . The M-step is as in (3.5) with  $n_{\text{ts}}, n_{\text{tv}}$ , and  $\bar{n}_{\text{CpG}}$  substituted by their conditional means. Unfortunately, the conditional means are only available in analytical form under the independence assumption. However, they can be simulated using a Gibbs sampling approach as described in the next section.

When the conditional mean calculation in the E-step of the EM algorithm is carried out using Markov chain Monte Carlo (MCMC) sampling, the resulting algorithm is called an MCMC-EM algorithm (Wei and Tanner 1990). The main disadvantage of having to

approximate the conditional means from Markov chain Monte Carlo sampling is that the likelihood of the observed data is no longer guaranteed to increase in every iteration of the EM algorithm. Recently, however, Caffo, Jank, and Jones (2005) proposed a strategy for recovering the likelihood-increasing property of the EM algorithm with high probability. Caffo, Jank, and Jones (2005) also described how to make efficient use of the Monte Carlo resources.

Convergence of the MCMC-EM algorithm was established by Fort and Moulines (2003) for curved exponential families and was also briefly discussed by Caffo, Jank, and Jones (2005, sec. 2.3). For more information on convergence properties of the MCMC-EM algorithm, the reader should consult Fort and Moulines (2003) and references therein.

#### 4. GIBBS SAMPLING

In Gibbs sampling, the path between nucleotide  $x_j(0)$  and  $x_j(T)$  is updated, conditional on the paths of all other nucleotides. Hwang and Green (2004) also used Gibbs sampling, but discretize time. In this article we use continuous time Gibbs sampling.

First, consider the situation on the left side of Figure 1. In this situation, the Gibbs update is a matter of simulating a sample path  $\{x_j(t) : 0 \leq t \leq T\}$  from a continuous time Markov chain with  $4 \times 4$  rate matrix given by (2.3), with fixed left neighboring nucleotide C, fixed right neighboring nucleotide T, and with beginning value  $x_j(0) = G$  and ending value  $x_j(T) = T$ .

Next, consider the more complicated situation shown on the right side of Figure 1. In this situation, the neighboring paths experience three substitutions at times  $t_1 < t_2 < t_3$ . In order to update the sample path  $\{x_j(t) : 0 \leq t \leq T\}$  with beginning value  $x_j(0)$  and ending value  $x_j(T)$ , we first determine the  $4 \times 4$  transition matrices  $P(0, t_1)$ ,  $P(t_1, t_2)$ ,  $P(t_2, t_3)$ , and  $P(t_3, T)$  in the four time intervals where there are no changes in the neighboring nucleotides. From these transition matrices and the starting and ending values of the Markov chain, we simulate the value of  $x_j(t)$  in the three change points  $t_1$ ,  $t_2$ , and  $t_3$ . Finally, we simulate the sample paths in each of the four intervals, conditional on the neighboring nucleotides and the simulated values in the change points.

From the above two examples it should be clear that in order for the Gibbs sampling approach to be applied, all we need is an algorithm for simulating sample paths from a continuous time Markov chain conditional on the beginning and ending values. One possible algorithm is based on rejection sampling, where a sample path is generated by simulating forward in time, and the resulting sample path is rejected if the simulated ending state is different from the observed ending state. Bladt and Sørensen (2005) used rejection sampling.

It is, however, well known that rejection sampling can be very slow. Nielsen (2002) considered the case when the time interval is small and the beginning and ending states are different. In this case rejection sampling is potentially very slow, but by taking advantage of the fact that at least one substitution must occur, Nielsen (2002) used the inverse transformation method to simulate the time before the first substitution.

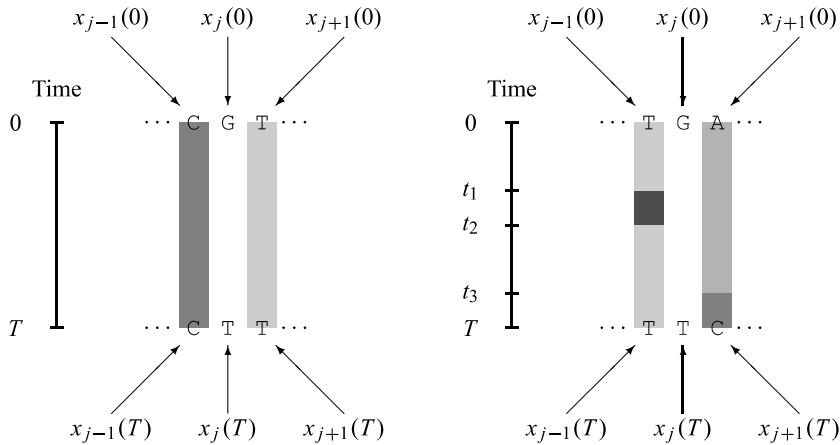


Figure 1. In Gibbs sampling, the path between  $x_j(0)$  and  $x_j(T)$  has to be updated, given the paths between all other nucleotides. For neighbor-dependent models, only the paths of the two neighboring sites are needed. Left: A situation where the neighboring paths experience no substitutions. Right: A situation where the neighboring paths experience three substitutions.

The MCMC-EM algorithm advocated in this article is applicable not only on the nucleotide level, but extends to the codon level. On the codon level, paths with two or three substitutions are often required, even on short evolutionary distances (e.g., a substitution from codon AAA to codon GGG). Nielsen's rejection sampling scheme is likely to be very slow for producing such sample paths because it only takes advantage of the fact that at least one substitution must occur, and it should be clear that a more general sampling approach is desirable. Neighbor-dependent codon models have been considered by Jensen and Pedersen (2000), Siepel and Haussler (2004), and Christensen, Hobolth, and Jensen (2005).

It should also be stressed that the exact path sampling algorithm derived in the following can be applied as a general tool for studying partially observed continuous time Markov processes on a discrete state space. Huelsenbeck, Nielsen, and Bollback (2003) discussed biological applications of continuous time Markov processes using path sampling algorithms.

#### 4.1 EXACT PATH SAMPLING ALGORITHM

We want to simulate a realization of a discrete-state Markov chain  $\{X(t) : 0 \leq t \leq T\}$  conditional on the starting state  $X(0) = a$  and final state  $X(T) = b$ . We consider the case where it is possible to make an eigenvalue decomposition of the rate matrix  $Q$ . Let  $U$  be the orthogonal matrix with eigenvalues as columns and  $D_\lambda$  the diagonal matrix of corresponding eigenvectors such that  $Q = UD_\lambda U^{-1}$ . Then the transition probability matrix is given by

$$P(t) = e^{Qt} = Ue^{tD_\lambda}U^{-1} \quad \text{and} \quad P_{ab}(t) = \sum_j U_{aj}U_{jb}^{-1}e^{t\lambda_j}. \quad (4.1)$$

First, consider the case where  $a = b$ . The probability of no substitutions in the time interval  $[0, T]$  conditional on the starting value of the Markov process  $X(0) = a$  and final value of the process  $X(T) = a$  is given by

$$p_a = \frac{e^{-q_a T}}{P_{aa}(T)}. \quad (4.2)$$

We use the notation  $q_{ab} = Q(a, b)$  for entries in the matrix  $Q$  and  $q_a = -q_{aa}$  for minus the diagonal entry in row  $a$  of matrix  $Q$ . Second, consider the probability of the first substitution of  $a$  being a substitution to  $i$ , conditional on the process starting in  $a$  and ending in  $b$ . This probability is given by

$$p_i = \int_0^T q_a e^{-q_a t} \frac{q_{ai}}{q_a} \frac{P_{ib}(T-t)}{P_{ab}(T)} dt = \int_0^T f_i(t) dt, \quad i \neq a, \quad (4.3)$$

where  $f_i(t)$  is the integrand. Using (4.1) we can rewrite the integrand as

$$f_i(t) = q_{ai} e^{-q_a t} \frac{P_{ib}(T-t)}{P_{ab}(T)} = \frac{q_{ai}}{P_{ab}(T)} \sum_j U_{ij} U_{jb}^{-1} e^{T\lambda_j} e^{-t(\lambda_j + q_a)}, \quad (4.4)$$

and so it is easy to calculate the integral in (4.3). We get

$$p_i = \frac{q_{ai}}{P_{ab}(T)} \sum_j U_{ij} U_{jb}^{-1} J_{aj}, \quad (4.5)$$

where

$$J_{aj} = \begin{cases} T e^{\lambda_j T} & \text{if } q_a + \lambda_j = 0 \\ \frac{e^{-q_a T} - e^{\lambda_j T}}{q_a + \lambda_j} & \text{if } q_a + \lambda_j \neq 0. \end{cases}$$

Putting these things together we have the following procedure for sampling a continuous time Markov chain  $\{X(t) : 0 \leq t \leq T\}$  that begins in  $X(0) = a$  and ends in  $X(T) = b$ . The procedure is illustrated in Figure 2.

1. If  $a = b$  sample  $Z \sim \text{Bernoulli}(p_a)$  where  $p_a$  is given by (4.2). If  $Z = 1$  we are done:  $X(t) = a$ ,  $0 \leq t \leq T$ .
2. If  $a \neq b$  or  $Z = 0$ , then at least one substitution happens. Calculate  $p_j$ ,  $j \neq a$ , from (4.5). Sample  $i \neq a$  from the discrete distribution  $p_j/p_{-a}$ ,  $j \neq a$ , where  $p_{-a} = \sum_{j \neq a} p_j$ .
3. Sample the waiting time  $\tau$  in state  $a$  according to the continuous density  $f_i(t)/p_i$ ,  $0 \leq t \leq T$ , where  $f_i(t)$  is given by (4.4). Set  $X(t) = a$ ,  $0 \leq t < \tau$ .
4. Repeat the procedure with new starting value  $i$  and new time interval  $T - \tau$ .

In Step 3 above, we simulate from the scaled density (4.4) by finding the cumulative distribution function and using the inverse transformation method.

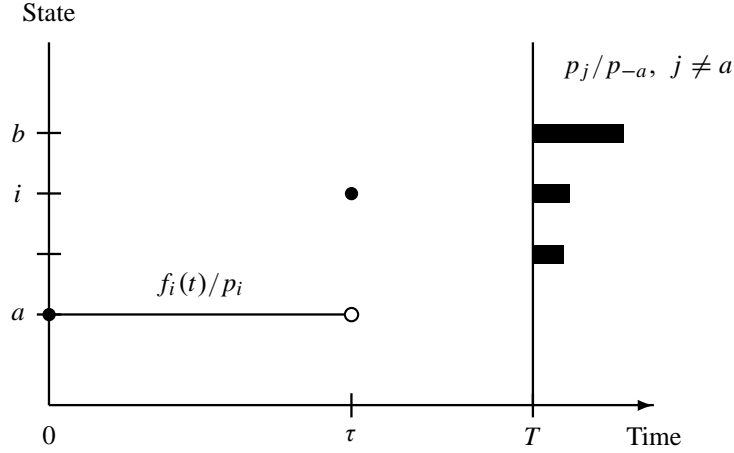


Figure 2. Algorithm for simulating the first substitution event (type and time) of a continuous time Markov process conditional on the beginning state  $a$  and ending state  $b$  of the process and that at least one substitution occurs. First, the new state  $i$  is found based on the discrete distribution  $p_j/p_{-a}$ ,  $j \neq a$ , where  $p_j$  is given by (4.3) and  $p_{-a} = \sum_{j \neq a} p_j$ . Second, the waiting time in state  $a$  is found based on the continuous density  $f_i(t)/p_i$ ,  $0 \leq t \leq T$ , where  $f_i(t)$  is given by (4.4).

## 4.2 SIMULATION STUDY: K80+CpG MODEL

A simulation study of the K80+CpG model described in Section 3.1 is carried out in order to compare dependent and independent sites models. Sequences  $x(0)$  and  $x(1)$  of length  $n = 750$  are simulated using parameter values  $\lambda = 0.15$ ,  $\alpha = 0.4$ , and  $\beta = 0.2$ . The parameter value of  $\lambda$  introduces CpG-deficiency. The ratio of  $\alpha/\beta = 2$  (the so-called transition-to-transversion rate ratio) makes it twice as likely to make a transition (such as, e.g., A→G) compared to a transversion (e.g., A→C or A→T).

The observed number of CpG dinucleotides in sequence  $x(0)$  is 7, and based on the stationary distribution (2.4) we obtain  $\hat{\lambda} = 0.125$  and a 95%-confidence interval [0.054, 0.246] for  $\lambda$ . The maximum log-likelihood is  $-1009.48$  while the log-likelihood obtained under the independent sites model with  $\lambda = 1$  is  $n \log(1/4) = -1039.72$ . These findings show that we can detect lack of independence from a single sequence analysis.

The parameters  $\alpha$  and  $\beta$  do not enter in the stationary distribution, but can be estimated from a pairwise analysis of sequences  $x(0)$  and  $x(1)$ . The independent sites Kimura (1980) model is so tractable that it is possible to find an analytical expression for the data likelihood. Following Ewens and Grant (2001, p. 378) the data likelihood is proportional to

$$(1 + e^{-\beta} + 2e^{-(\alpha+\beta)/2})^{n_0} (1 + e^{-\beta} - 2e^{-(\alpha+\beta)/2})^{n_1} (1 - e^{-\beta})^{n_2},$$

where  $n_0$  is the number of sites where the nucleotides in sequences  $x(0)$  and  $x(1)$  are the same,  $n_1$  is the number of sites where a purine (pyrimidine) occurs in sequence  $x(0)$  and the other purine (pyrimidine) occurs in sequence  $x(1)$ , and  $n_2$  is the number of sites where a purine occurs in one sequence and a pyrimidine in the other. For the simulated data we have  $n_0 = 627$ ,  $n_1 = 55$  and  $n_2 = 68$ . Maximization of the independent sites data log-

likelihood function leads to the estimates  $\hat{\alpha}_0 = 0.3418$  and  $\hat{\beta}_0 = 0.2001$ . Furthermore, the log-likelihood evaluated at the independent sites maximum likelihood estimates is  $-419.25$ .

The dependent sites K80+CpG model can be analyzed using the MCMC-EM algorithm outlined in Section 3.1. The MCMC-EM algorithm works by updating the two parameters  $\alpha$  and  $\beta$  using Equation (3.5) with  $n_{ts}$ ,  $n_{tv}$ , and  $n_{CpG}$  replaced by the conditional means

$$E[n_{ts}|x(0), x(1)], \quad E[n_{tv}|x(0), x(1)] \quad \text{and} \quad E[\bar{n}_{CpG}|x(0), x(1)],$$

calculated under the current parameter values of  $\alpha$  and  $\beta$ . The conditional means are estimated by simulating sample paths for each site, conditional on the paths of the neighboring sites. This is the exact Gibbs sampler described previously in this section. A Monte Carlo sample is obtained when the sample path for every single site has been simulated.

The initial values of  $\alpha$  and  $\beta$  are the independent sites estimates and the Monte Carlo sample size is 10 (iterations 1–4), 50 (5–8), 200 (9–12), and 500 (13–16). As can be seen from Table 3, the algorithm seems to stabilize rather quickly. From the results of iteration 14–16, the maximum likelihood estimates are  $(\hat{\alpha}, \hat{\beta}) = (0.3404, 0.1881)$ , correct to two decimal places. Using a prespecified number of Monte Carlo sample sizes does not make efficient use of computational resources and does not ensure the likelihood-increasing property of the EM algorithm. Caffo, Jank, and Jones (2005) described a method that deals with these two issues. We use Caffo, Jank, and Jones' method in the more complicated situation of multiple sequences and a general time reversible model with CpG effect. The GTR+CpG model for multiple sequences is considered in the next section.

The increase in data log-likelihood for the substitution process can be obtained using the formula

$$\frac{L_{\hat{\alpha}_0, \hat{\beta}_0}(y)}{L_{\hat{\alpha}, \hat{\beta}}(y)} = E_{\hat{\alpha}, \hat{\beta}} \left[ \frac{L_{\hat{\alpha}_0, \hat{\beta}_0}(x)}{L_{\hat{\alpha}, \hat{\beta}}(x)} \mid y \right],$$

where  $y = (x(0), x(1))$  is the observed data,  $L(y)$  is the data likelihood, and  $L(x)$  is the full likelihood given by (3.4). The conditional expectation is easily calculated from the Monte Carlo samples and we obtain a data log-likelihood difference of 0.15 between the independent and dependent sites models. This difference is not very large, showing that the CpG-effect cannot be detected from the substitution pattern only. Thus, the simulation study shows that for short pairs of sequences from closely related species, the CpG effect is easier to detect from the stationary distribution than from the substitution pattern.

## 5. MCMC-EM ALGORITHM FOR MULTIPLE SEQUENCES

In this section we analyze a multiple alignment of 741 sites and five species (human, chimpanzee, orangutan, mouse, and rat) from noncoding DNA using the GTR+CpG model described in Section 2 and the MCMC-EM algorithm described in Section 3. In Figure 3 we show the phylogenetic tree that relates the five species. The multiple alignment was obtained from [www.nics.nih.gov/data](http://www.nics.nih.gov/data) and is a subset of the data analyzed by Hwang and Green (2004).

Table 3. Parameter estimates for the K80+CpG model for two simulated sequences. The model has two rate parameters  $\alpha$  and  $\beta$ . The first column in the table shows the number of iterations used in the MCMC-EM algorithm and the second column shows the Monte Carlo sample size used within each iteration.

Iteration	Sample Size	Rate parameters	
		$\alpha$	$\beta$
0		0.3418	0.2001
1	10	0.3524	0.1904
2	10	0.3400	0.1904
3	10	0.3412	0.1868
4	10	0.3304	0.1824
5	50	0.3368	0.1864
6	50	0.3328	0.1904
7	50	0.3396	0.1916
8	50	0.3316	0.1856
9	200	0.3360	0.1884
10	200	0.3388	0.1888
11	200	0.3412	0.1872
12	200	0.3404	0.1888
13	500	0.3384	0.1888
14	500	0.3392	0.1876
15	500	0.3408	0.1884
16	500	0.3412	0.1884

We use the estimation procedure advocated by Christensen, Hobolth, and Jensen (2005) and estimate the CpG parameter  $\lambda$  and frequencies  $\pi$  from the stationary distribution using the human sequence as reported in Section 2.2.

### 5.1 GTR+CpG MODEL FOR PAIRWISE SEQUENCES

Consider the GTR+CpG model for pairwise sequences. We wish to estimate the six free parameters  $\theta = (\theta_{AG}, \theta_{AC}, \theta_{AT}, \theta_{GC}, \theta_{GT}, \theta_{CT})$  of the model using the MCMC-EM algorithm. The waiting times (3.1) in the GTR+CpG model are determined by

$$\begin{aligned}
\Gamma_{\theta}(t) = & n_A(t)(\theta_{AG}\pi_G + \theta_{AC}\pi_C + \theta_{AT}\pi_T) \\
& + (n_G(t) - n_{CpG}(t))(\theta_{AG}\pi_A + \theta_{GC}\pi_C + \theta_{GT}\pi_T) \\
& + (n_C(t) - n_{CpG}(t))(\theta_{AC}\pi_A + \theta_{GC}\pi_G + \theta_{CT}\pi_T) \\
& + n_T(t)(\theta_{AT}\pi_A + \theta_{GT}\pi_G + \theta_{CT}\pi_C) \\
& + n_{CpG}(t)(\theta_{AG}\pi_A + \theta_{GC}\pi_C + \theta_{GT}\pi_T + \theta_{AC}\pi_A + \theta_{GC}\pi_G + \theta_{CT}\pi_T)/\lambda, \quad (5.1)
\end{aligned}$$

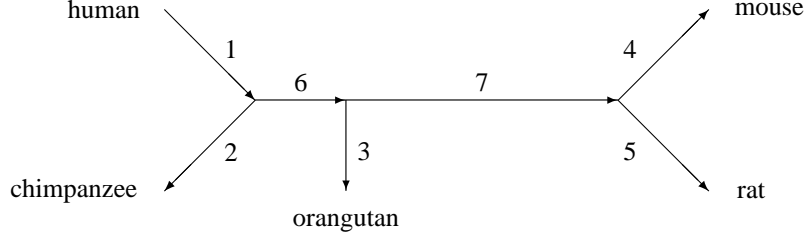


Figure 3. Unrooted phylogenetic tree relating the five species in the multiple alignment. The GTR+C<sub>PG</sub> model is time reversible and thus we can choose any of the leaves to be the root. The human leaf is chosen as the root. The numbering of the seven branches is also shown.

and the full log-likelihood (3.3) becomes, up to an additive constant,

$$\log L_{\theta}(x) = n_{AG} \log \theta_{AG} + n_{AC} \log \theta_{AC} + \cdots + n_{TC} \log \theta_{CT} - \sum_{m=0}^M \Gamma_{\theta}(t_m)(t_{m+1} - t_m). \quad (5.2)$$

From (5.1) and (5.2) it follows that the sufficient statistics of the model are determined by the number of substitutions between any two different states  $n_{AG}, \dots, n_{TC}$  and the weighted average number of nucleotides  $\bar{n}_A, \bar{n}_G, \bar{n}_C, \bar{n}_T$  and the weighted average number of C<sub>PG</sub> dinucleotides  $\bar{n}_{C_{PG}}$  in the sequence where, for example,

$$\bar{n}_A = \sum_{m=0}^M (t_{m+1} - t_m) n_A(t).$$

Another interpretation of  $\bar{n}_A$  is that it is the aggregated total time spent in state A. Note that the last term in (5.2) is linear in the weighted average number of nucleotides and C<sub>PG</sub> dinucleotides and in the parameters. Adding terms and introducing a shorter notation we can write

$$\log L_{\theta}(x) = f_{AG} \log \theta_{AG} + \cdots + f_{CT} \log \theta_{CT} - g_{AG} \theta_{AG} - \cdots - g_{CT} \theta_{CT}, \quad (5.3)$$

where, for example,  $f_{AG} = n_{AG} + n_{GA}$  and

$$g_{AG} = \bar{n}_A \pi_G + (\bar{n}_G - \bar{n}_{C_{PG}}) \pi_A + \bar{n}_{C_{PG}} \pi_A / \lambda = \bar{n}_A \pi_G + \bar{n}_G \pi_A + (1/\lambda - 1) \bar{n}_{C_{PG}} \pi_A.$$

For pairwise sequences, the M-step of the MCMC-EM algorithm is straightforward. From (5.3) it follows immediately that  $\theta_{AG}$  is updated by  $\theta_{AG} = f_{AG}/g_{AG}$ , and updating the remaining parameters follows in a similar fashion.

## 5.2 GTR+C<sub>PG</sub> MODEL FOR MULTIPLE SEQUENCES

For multiple sequences, the analysis is somewhat more complicated because we must also estimate the branch lengths. For the five sequences considered in Figure 3, we thus

have 13 parameters; the 6 rate parameters  $\theta = (\theta_{AG}, \theta_{AC}, \theta_{AT}, \theta_{GC}, \theta_{GT}, \theta_{CT})$  and the 7 branch length parameters  $\tau = (\tau_1, \dots, \tau_7)$ . More generally, an unrooted phylogenetic tree with  $I$  leaves has  $2I - 3$  branches.

Let  $\theta_j$ ,  $j = 1, \dots, J$ , refer to the  $J = 6$  rate parameters. From (5.3) it follows that the full log-likelihood for a tree with  $I$  leaves becomes

$$\log L_{\theta, \tau}(x) = \sum_{i=1}^{2I-3} \sum_{j=1}^J \left( f_{ij} \log(\tau_i \theta_j) - g_{ij} \tau_i \theta_j \right). \quad (5.4)$$

Here  $f_{ij} = f_{ij}(x)$  is a linear function of the number of substitutions between any two different states  $n_{AG}, \dots, n_{TC}$  on lineage  $i$  and  $g_{ij} = g_{ij}(x)$  is a linear function of the weighted average number of nucleotides  $\bar{n}_A, \bar{n}_G, \bar{n}_C, \bar{n}_T$  and CpG dinucleotides  $\bar{n}_{CpG}$  in the sequence on lineage  $i$ . Note that time and rate are confounded. In order to be able to identify the parameters we let  $\theta_{AC} = 1$ .

In the M-step, we need to maximize (5.4) with respect to  $\theta$  and  $\tau$  and with  $f_{ij}$  and  $g_{ij}$  substituted with their conditional means. Given the branch lengths, the rate parameters are easy to maximize. The complete log-likelihood (5.4) is maximized for

$$\theta_j = \frac{\sum_{i=1}^{2I-3} f_{ij}}{\sum_{i=1}^{2I-3} \tau_i g_{ij}}.$$

Similarly, the branch lengths are easy to maximize when the rate parameters are known. The branch lengths are maximized for

$$\tau_i = \frac{\sum_{j=1}^J f_{ij}}{\sum_{j=1}^J \theta_j g_{ij}}.$$

Within the M-step, we iterate between updating the rate parameters for given branch lengths and updating the branch lengths for given rate parameters. This iterative algorithm is called Zellner's two-stage procedure, and convergence properties are described by, for example, Lauritzen (1996) and Drton (2004, Appendix A).

In the E-step, we need to calculate the expected number of substitutions between any two nucleotides and the weighted average number of nucleotides and CpG dinucleotides on each branch, conditionally on the observed sequences in the leaves. We find these expectations using Monte Carlo sampling. The Gibbs sampling procedure now consists of updating the sample path for a single site conditional on the paths of the neighboring sites and the observed states in the leaves.

The sample path simulation consists of three parts. First, the transition matrices between the nodes are calculated along the same lines as described in connection with Figure 1. Based on these transition matrices, the states of the inner nodes are simulated. Second, the states of the change points on each edge are simulated, and finally the sample paths between change points are simulated.

### 5.3 PARAMETER ESTIMATES AND CONFIDENCE INTERVALS

In order to estimate the parameters, we use the method advocated by Caffo, Jank, and Jones (2005). Caffo, Jank, and Jones (2005) described a method to efficiently use com-

putational resources and at the same time ensure the likelihood-increasing property of the EM algorithm with high probability.

Denote the parameters in the model  $\psi = (\theta, \tau)$  and let  $\tilde{\psi}_{(s-1)}$  be the current MCMC-EM parameter estimate and  $\{x_{s,k} : k = 1, \dots, m_s\}$  the current Monte Carlo sample. The Monte Carlo sample is obtained after  $m_s$  sweeps of the Gibbs sampler conditional on the observed data  $y = y(x)$  (the five sequences in the leaves) and with parameter value  $\tilde{\psi}_{(s-1)}$ . Recall from Equation (5.4) that the sufficient statistics for a sample consists of the terms  $f_{ij}$  and  $g_{ij}$ , which are functions of the substitutions between any two different states and the weighted average of single nucleotides and CpG dinucleotides in the sample. Plots of the autocorrelations indicate that the sufficient statistics are approximately independent between sweeps, and we therefore apply Caffo, Jank, and Jones' methodology developed for independent samples from the model conditional on the observed data  $y = y(x)$  and parameter value  $\tilde{\psi}_{(s-1)}$ .

Let  $\tilde{\psi}_{(s,m_s)}$  be the proposed new MCMC-EM parameter estimate based on the random sample  $\{x_{(s,k)} : k = 1, \dots, m_s\}$ . Caffo, Jank, and Jones (2005) described a method to decide if the proposed MCMC-EM estimate should be accepted or if the Monte Carlo sample size  $m_s$  should be increased. Recall from (3.6) that  $G(\cdot, \psi_{(s-1)})$  is the full log-likelihood conditional on the observed data and the parameter estimate  $\psi_{(s-1)}$ . The new MCMC-EM parameter estimate should be accepted if the data log-likelihood is increased which corresponds to evidence that

$$\Delta G(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)}) \equiv G(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)}) - G(\tilde{\psi}_{(s-1)}, \tilde{\psi}_{(s-1)}) > 0.$$

A consistent estimate of  $\Delta G(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)})$  is given by

$$\Delta \tilde{G}(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)}) \equiv \tilde{G}(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)}) - \tilde{G}(\tilde{\psi}_{(s-1)}, \tilde{\psi}_{(s-1)}) = \sum_{k=1}^{m_s} \Lambda_k / m_s, \quad (5.5)$$

where

$$\Lambda_k = \log L_{\tilde{\psi}_{(s,m_s)}}(x_{(s,k)}) - \log L_{\tilde{\psi}_{(s-1)}}(x_{(s,k)}), \quad k = 1, \dots, m_s.$$

The full log-likelihoods are given by Equation (5.4). Since the MCMC-EM algorithm is based on a Monte Carlo estimation of the conditional expectations, we should only require that (5.5) is positive with high probability. Caffo, Jank, and Jones (2005) argued that this condition amounts to

$$\Delta \tilde{G}(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)}) > z_\alpha \text{ASE}. \quad (5.6)$$

Here  $z_\alpha$  is such that  $P(Z > z_\alpha) = \alpha$ , where  $Z$  is a standard normal random variable, and  $\text{ASE} = \hat{\sigma} / \sqrt{m_s}$ , where  $\hat{\sigma}$  is the sample variance of  $\Lambda_k$ ,  $k = 1, \dots, m_s$ . We follow Caffo, Jank, and Jones (2005) and let  $\alpha = 0.3$ .

If condition (5.6) is fulfilled, the new proposed MCMC-EM parameter estimate is accepted, and the algorithm moves to the next iteration. If the condition is not fulfilled, we generate new Monte Carlo samples, append them to the existing samples, and obtain a new parameter estimate by using the larger Monte Carlo sample. This latter process is repeated

Table 4. Parameter estimates for the GTR+CpG model on a tree with five lineages. The model has five relative rate parameters ( $\theta_{AC} = 1$ ) and seven branch length parameters. Numbering of the branches is indicated in Figure 3. The first column in the table shows the number of iterations used in the MCMC-EM algorithm and the second column shows the Monte Carlo sample size used within each iteration. The last row shows the standard deviations and was calculated from the observed information matrix.

Iteration	Sample size	Rate parameters				Branch lengths							
		$\theta_{AG}$	$\theta_{AT}$	$\theta_{GC}$	$\theta_{GT}$	$\theta_{CT}$	$\tau_1$	$\tau_2$	$\tau_3$	$\tau_4$	$\tau_5$	$\tau_6$	$\tau_7$
0		4.56	0.34	0.60	1.06	3.09	0.0045	0.0022	0.0067	0.0384	0.0216	0.0103	0.1585
1	10	4.74	0.38	0.58	1.01	3.37	0.0039	0.0018	0.0047	0.0355	0.0212	0.0095	0.1474
2	10	4.87	0.36	0.58	1.11	3.38	0.0041	0.0015	0.0052	0.0326	0.0206	0.0092	0.1445
3	634	4.74	0.32	0.52	1.02	3.33	0.0041	0.0017	0.0048	0.0339	0.0223	0.0102	0.1527
4	846	4.72	0.29	0.50	1.06	3.33	0.0042	0.0016	0.0049	0.0345	0.0226	0.0100	0.1527
5	1128	4.73	0.30	0.51	1.05	3.32	0.0042	0.0016	0.0049	0.0346	0.0225	0.0100	0.1524
s.d.		1.54	0.19	0.28	0.41	1.04	0.0024	0.0017	0.0035	0.0115	0.0088	0.0043	0.0445

until the condition is fulfilled. We follow Caffo, Jank, and Jones (2005) and let the next additional sample size be  $m_s/3$ . For MCMC-EM iteration  $s$ , let  $m_{s,\text{start}}$  be the starting Monte Carlo sample size and  $m_{s,\text{end}}$  be the ending Monte Carlo sample size. The initial sample size is  $m_{1,\text{start}} = 10$  and the subsequent starting values are  $m_{s,\text{start}} = m_{(s-1),\text{end}}$ .

The estimated parameter values of the MCMC-EM algorithm are shown in Table 4. We are in the fortunate situation where reasonable starting values for the MCMC-EM algorithm can be provided. This means that the algorithm converges very quickly (a similar situation was reported for the Gibbs sampler by Jensen and Pedersen (2000)). As expected, the DNA sequences from human, chimpanzee, and orangutan are closely related, the sequences from mouse and rat are closely related, and the two clades are separated by a relatively long branch. Furthermore, the parameter estimates suggest that a strand-symmetric model would be appropriate. Strand-symmetry (e.g., Yap and Speed 2004) is fulfilled when  $\pi_A = \pi_T$ ,  $\pi_G = \pi_C$ ,  $\theta_{AG} = \theta_{CT}$  and  $\theta_{AC} = \theta_{GT}$  (recall that  $\theta_{AC} = 1$ ).

Caffo, Jank, and Jones (2005) suggested terminating the MCMC-EM algorithm when

$$\Delta \tilde{G}(\tilde{\psi}_{(s,m_s)}, \tilde{\psi}_{(s-1)}) + z_\gamma \text{ASE} \quad (5.7)$$

is smaller than a prespecified constant and with  $\gamma = 0.05$ . Caffo, Jank, and Jones (2005) use a termination constant as low as  $10^{-5}$ , but we found it sufficient to use a termination constant of  $10^{-1}$ .

In order to determine the uncertainty of the parameter values we follow Louis (1982) and let

$$S(\psi; x) = \frac{\partial \log L_\psi(x)}{\partial \psi} \quad \text{and} \quad I(\psi; x) = -\frac{\partial^2 \log L_\psi(x)}{\partial \psi \partial \psi^*}$$

be the likelihood score and information matrix based on the full likelihood. Superscript  $*$  denotes vector or matrix transpose and all vectors are column vectors. Louis (1982) showed that the information matrix based on the observed data  $y = y(x)$  and evaluated at  $\psi = \hat{\psi}$  is given by

$$I(\psi; y) = E_\psi [I(\psi; x)|y] - E_\psi [S(\psi; x)S^*(\psi; x)|y].$$

Thus the information matrix based on data  $y$  can be computed from the conditional mean values of the full likelihood quantities. In Table 4, the standard deviations of the rate parameters and branch lengths are calculated from the observed information matrix.

In Equation (A.5) in the Appendix, the expected number of substitutions per site on a branch is derived. The expected number of substitutions on a branch depends linearly on the entries in the substitution rate matrix. Using this linear dependency and the delta-method (e.g., Oehlert 1992), we can obtain the expected number of substitutions on each branch and the corresponding standard deviation. The values are reported in Table 5. The expected number of substitutions correspond very well to the numbers that were obtained in the simulations.

Let  $L_\psi(y)$  denote the data likelihood. Furthermore, let  $\hat{\psi}_0$  be the maximum likelihood estimates under the independent sites GTR model and  $\hat{\psi}$  the estimates under the GTR+CpG

Table 5. Expected number of substitutions per site  $\nu_i$  on each of the seven branches and the corresponding standard deviations.

	$\nu_1$	$\nu_2$	$\nu_3$	$\nu_4$	$\nu_5$	$\nu_6$	$\nu_7$
estimate	0.0058	0.0024	0.0069	0.0475	0.0309	0.0141	0.2125
s.d.	0.0031	0.0022	0.0041	0.0097	0.0088	0.0052	0.0245

model. The increase in data log-likelihood is calculated using the formula

$$\frac{L_{\hat{\psi}_0}(y)}{L_{\hat{\psi}}(y)} = E_{\hat{\psi}} \left[ \frac{L_{\hat{\psi}_0}(x)}{L_{\hat{\psi}}(x)} \mid y \right].$$

The conditional expectation is easily calculated using the Gibbs sampler, and we obtain a data log-likelihood difference of 0.88 between the two models. This difference is not very large and is probably due to the limited amount of sequence data. For longer alignments, the context dependent model is expected to fit the substitution pattern much better than the independent sites model.

## 6. DISCUSSION

The MCMC-EM algorithm for estimating the instantaneous rates of neighbor-dependent substitution models, as developed in this article, provides a powerful tool for analyzing substitution patterns in homologous DNA sequences. The approach can be extended to analyze more general context dependent models where the substitution rate at a site depends not only on the nearest neighbor, but also on sites further apart.

An important feature of the proposed neighbor-dependent model is the analytical expression of the stationary distribution. The relation between the instantaneous rates and the stationary distribution makes it possible to test for simple aspects of the model. In particular, we found in Section 2 that the neighbor-dependent model can adequately describe the single human DNA sequence, and that an independent sites model would not be appropriate.

The requirement that the stationary distribution should be accessible also has its drawbacks. We find the stationary distribution using the detailed balance condition, which also implies that the process is reversible. While the reversibility assumption is tractable, it is not likely to be fulfilled for DNA sequence evolution. The model (2.3) increases the rate away from CpG sites, but does not directly model the CpG-methylation-deamination process where only rates from CpG to TpG or CpA should be increased. The CpG-methylation-deamination process violates the reversibility assumption and in order to ensure reversibility, it is only taken into account as an increase away from CpG sites.

Time reversibility is used in a crucial way to obtain the stationary distribution, but it should be emphasized that time reversibility is not used in the MCMC-EM algorithm. The MCMC-EM algorithm therefore also applies to nonreversible neighbor-dependent models.

For nonreversible processes, we require a rooted phylogenetic tree, and in most cases the root sequence is not available. It seems appropriate to use a Markov chain to model the root sequence. Recall that in this article the stationary distribution is a Markov chain along the sequence, and with the assumption that the root is in stationarity, the Markov assumption is exact.

Hwang and Green (2004) considered an unrestricted neighbor-dependent model and used a Bayesian procedure to estimate the parameters. The change from one nucleotide (four possible types) to another (three possible types) depends on the flanking neighboring situation ( $4 \cdot 4$  possible types), so the model has a total of  $4 \cdot 3 \cdot 4 \cdot 4 = 192$  free parameters. Generally, this model is not reversible and Hwang and Green (2004) used a second-order Markov chain along the root sequence. The dataset analyzed in Hwang and Green (2004) is huge; it consists of 19 species and the alignment is of length approximately 1.7 mega bases. The model considered in this article has much fewer parameters than Hwang and Green's unrestricted neighbor-dependent model and is thus appropriate for smaller datasets.

Continuous time Markov chains on evolutionary trees are used in a wide range of applications in molecular evolution and are becoming increasingly popular. Examples include comparative gene finding (e.g. Pedersen and Hein 2003; Hobolth and Jensen 2005b), phylogeny reconstruction (e.g., Guindon and Gascuel 2003; Ren and Yang 2005), alignment programs (e.g., Redelings and Suchard 2005; Lunter et al., 2005) and detection of selection (e.g., Clark et al., 2003). All these applications make the independent sites assumption and it would be interesting to investigate if the performance could be improved by allowing neighbor-dependent effects.

## A. NORMALIZING CONSTANT, EXPECTED DINUCLEOTIDE COUNTS AND EXPECTED NUMBER OF SUBSTITUTIONS ON A BRANCH

### A.1 NORMALIZING CONSTANT

Assume  $x_0 \neq C$  and  $x_{n+1} \neq G$  are fixed flanking nucleotides such that the stationary distribution (2.4) is given by

$$P(x) = \frac{1}{Z(\lambda, \pi)} \pi_{x_1} \prod_{j=1}^{n-1} \lambda^{1_{(C,G)}(x_j, x_{j+1})} \pi_{x_{j+1}}.$$

Define the  $4 \times 4$  matrix  $A$  with entries  $A(a, b) = \lambda^{1_{(C,G)}(a,b)} \pi_b$ . Then the stationary distribution can be written as

$$P(x) = \frac{1}{Z} \pi_{x_1} \prod_{j=1}^{n-1} A(x_j, x_{j+1}), \quad (\text{A.1})$$

and

$$Z = \sum_{x=(x_1, \dots, x_n)} \pi_{x_1} \prod_{j=1}^{n-1} A(x_j, x_{j+1}) = \sum_{x_1, x_n} \pi_{x_1} A^{n-1}(x_1, x_n).$$

The two nonzero eigenvalues of  $A$  are given by

$$\mu_1 = \frac{1}{2}(1 + \sqrt{1 - 4(1 - \lambda)\pi_C\pi_G}) \quad \text{and} \quad \mu_2 = \frac{1}{2}(1 - \sqrt{1 - 4(1 - \lambda)\pi_C\pi_G}),$$

with corresponding right eigenvectors

$$r_i = (1, 1, 1 + \frac{\mu_i - 1}{\pi_C}, 1),$$

and left eigenvectors

$$l_i = \frac{1}{\pi_A + \frac{\pi_G(1 - \pi_G - (1 - \lambda)\pi_C)}{\mu_i - \pi_G} + \pi_C + \mu_i - 1 + \pi_T} \\ \times (\pi_A, \frac{\pi_G(1 - \pi_G - (1 - \lambda)\pi_C)}{\mu_i - \pi_G}, \pi_C, \pi_T).$$

The eigenvectors are normalized such that  $\sum_a l_i(a)r_i(a) = 1$ . We get

$$A = \mu_1 r_1^* l_1 + \mu_2 r_2^* l_2, \quad \text{and} \quad A^n = \mu_1^n r_1^* l_1 + \mu_2^n r_2^* l_2,$$

and thereby, for large  $n$ ,

$$Z = \sum_{a,b} \pi_a (\mu_1^{n-1} r_1(a) l_1(b) + \mu_2^{n-1} r_2(a) l_2(b)) \\ \approx \sum_{a,b} \pi_a \mu_1^{n-1} r_1(a) l_1(b) \\ = \mu_1^{n-1} \left( \sum_a \pi_a r_1(a) \right) \left( \sum_b l_1(b) \right).$$

Note that

$$\sum_a \pi_a r_1(a) = \pi_A + \pi_G + \pi_C + \mu_1 - 1 + \pi_T = \mu_1,$$

and thereby

$$Z = \mu_1^n \left( \sum_b l_1(b) \right). \quad (\text{A.2})$$

## A.2 EXPECTED DINUCLEOTIDE COUNTS

From (A.1) we get the expected dinucleotide count

$$\mathbb{E}[n_{(a,b)}] = \sum_{x=(x_1, \dots, x_n)} P(x) \sum_{j=1}^{n-1} 1_{a,b}(x_j, x_{j+1}) \\ = \frac{1}{Z} \sum_{j=1}^{n-1} \sum_{x=(x_1, \dots, x_n)} \pi_{x_1} \prod_{k=1}^{n-1} A(x_k, x_{k+1}) 1_a(x_j) 1_b(x_{j+1}). \quad (\text{A.3})$$

Using similar calculations as above, the last term can be approximated by

$$\begin{aligned}
& \sum_{x=(x_1, \dots, x_n)} \pi_{x_1} \prod_{k=1}^{n-1} A(x_k, x_{k+1}) 1_a(x_j) 1_b(x_{j+1}) \\
&= \sum_{x_j, x_{j+1}} 1_a(x_j) 1_b(x_{j+1}) \left( \sum_{x_1, \dots, x_{j-1}} \pi_{x_1} \prod_{k=1}^{j-1} A(x_k, x_{k+1}) \right) A(x_j, x_{j+1}) \\
&\quad \times \left( \sum_{x_{j+2}, \dots, x_n} \prod_{k=j+1}^{n-1} A(x_k, x_{k+1}) \right) \\
&\approx \sum_{x_j, x_{j+1}} 1_a(x_j) 1_b(x_{j+1}) \left( \mu_1^j l_1(x_j) \right) A(x_j, x_{j+1}) \\
&\quad \times \left( \mu_1^{n-j-1} r_1(x_{j+1}) \sum_{x_n} l_1(x_n) \right) \\
&= \mu_1^{n-1} l_1(a) r_1(b) A(a, b) \left( \sum_c l_1(c) \right)
\end{aligned}$$

Note that this expression does not depend on  $j$  and from (A.2) and (A.3) we get

$$E[n_{(a,b)}] = (n-1) l_1(a) A(a, b) r_1(b) / \mu_1. \quad (\text{A.4})$$

### A.3 EXPECTED NUMBER OF SUBSTITUTIONS ON A BRANCH

The expected number of substitutions per site  $v$  on a branch is given by

$$\begin{aligned}
v &= \frac{1}{n} \sum_{x=(x_1, \dots, x_n)} \sum_{j=1}^n \sum_{\tilde{x}_j \neq x_j} P(x) \gamma(\tilde{x}_j; x_{j-1}, x_j, x_{j+1}) \\
&= \frac{1}{n} \sum_{x=(x_1, \dots, x_n)} \sum_{j=1}^n \frac{1}{Z} \pi_{x_1} \left( \prod_{k=1}^{n-1} A(x_k, x_{k+1}) \right) \\
&\quad \times (1/\lambda)^{1_{(c,g)}(x_{j-1}, x_j) + 1_{(c,g)}(x_j, x_{j+1})} \sum_{\tilde{x}_j \neq x_j} Q(x_j, \tilde{x}_j).
\end{aligned}$$

From  $A(a, b) (1/\lambda)^{1_{\text{CpG}}(a,b)} = \pi_b$  and similar calculations as previously in this Appendix we get

$$v \approx \frac{1}{\mu_1} \left( \sum_a l_1(a) \right) \left( \sum_b \sum_{c \neq b} \pi_b Q(b, c) \right). \quad (\text{A.5})$$

Note that the last term is the branch length had there been no CpG effect.

## ACKNOWLEDGMENTS

I am grateful to Ole F. Christensen, the Associate Editor and three referees for helpful comments and suggestions. I would like to thank Jeff Thorne for numerous illuminating and fruitful discussions. I am financially supported by the Danish Research Council grant 21-04-0375 and the National Institute of Health grant R01 GM070806.

[Received April 2006. Revised April 2007.]

## REFERENCES

- Albert, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002), *Molecular Biology of the Cell*, New York: Garland Science.
- Arndt, P.F., and Hwa, T. (2005), "Identification and Measurement of Neighbour-Dependent Nucleotide Substitution Processes," *Bioinformatics*, 21, 2322–2328.
- Bladt, M., and Sørensen, M. (2005), "Statistical Inference for Discretely Observed Markov Jump Processes," *Journal of the Royal Statistical Society, Ser. B*, 67, 395–410.
- Blake, R.D., Hess, S.T., and Nicholson-Tuell, J. (1992), "The Influence of Nearest Neighbors on the Rate and Pattern of Spontaneous Point Mutations," *Journal of Molecular Evolution*, 34, 189–200.
- Caffo, B.S., Jank, W., and Jones, G.L. (2005), "Ascent-Based Monte Carlo Expectation-Maximization," *Journal of the Royal Statistical Society, Ser. B*, 67, 235–251.
- Christensen, O.F., Hobolth, A., and Jensen, J.L. (2005), "Pseudo-Likelihood Analysis of Context-Dependent Codon Substitution Models," *Journal of Computational Biology*, 12, 1166–1182.
- Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., Ferreira, S., Wang, G., Zheng, X., White, T.J., Sninsky, J.J., Adams, M.D., Cargill, M. (2003), "Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios," *Science*, 302, 1960–1963.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–22.
- Diaconis, P., and Rolles, S.W.W. (2006), "Bayesian Analysis for Reversible Markov Chains," *The Annals of Statistics*, 34, 1270–1292.
- Drton, M. (2004), "Maximum Likelihood Estimation in Gaussian AMP Chain Graph Models and Gaussian Ancestral Graph Models," unpublished Ph.D. thesis, Department of Statistics, University of Washington.
- Ewens, W.J., and Grant, G.R. (2001), *Statistical Methods in Bioinformatics*, New York: Springer.
- Fort, G., and Moulines, E. (2003), "Convergence of the Monte Carlo Expectation Maximization for Curved Exponential Families," *The Annals of Statistics*, 31, 1220–1259.
- Guindon, S., and Gascuel, O. (2003), "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood," *Systematic Biology*, 52, 696–704.
- Hess, S.T., Blake, J.D., and Blake, R.D. (1994), "Wide Variations in Neighbour-Dependent Substitution Rates," *Journal of Molecular Biology*, 236, 1022–1033.
- Holmes, I., and Rubin, G.M. (2002), "An Expectation Maximization Algorithm for Training Hidden Substitution Models," *Journal of Molecular Biology*, 317, 757–768.
- Hobolth, A., and Jensen, J.L. (2005a), "Statistical Inference in Evolutionary Models of DNA Sequences via the EM Algorithm," *Statistical applications in Genetics and Molecular Biology*, 4, 18.
- (2005b), "Applications of Hidden Markov Models for Characterization of Homologous DNA Sequences with a Common Gene," *Journal of Computational Biology*, 12, 186–203.
- Huelsenbeck, J.P., Nielsen, R., and Bollback, J.P. (2003), "Stochastic Mapping of Morphological Characters," *Systematic Biology*, 52, 131–158.
- Hwang, D.G., 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 Mathematical Sciences, Aarhus University.
- Jensen, J.L., and Pedersen, A.K. (2000), "Probabilistic Models of DNA Sequence Evolution with Context Dependent Rates of Substitution," *Advances in Applied Probability*, 32, 499–517.

- Kimura, M. (1980), "A Simple Method for Estimating Evolutionary Rate in a Finite Population due to Mutational Production of Neutral and Nearly Neutral Base Substitution through Comparative Studies of Nucleotide Sequences," *Journal of Molecular Evolution*, 16, 111–120.
- Lauritzen, S.L. (1996), *Graphical Models*, Oxford, UK: Clarendon Press.
- Louis, A.T. (1982), "Finding the Observed Information Matrix When using the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 44, 226–233.
- Lunter, G.A., and Hein, J. (2004), "A Nucleotide Substitution Model with Nearest-Neighbour Interactions," *Bioinformatics*, special issue for ISMB2004, 20, i216–i223.
- Lunter, G. A., Miklos, I., Drummond, A.J., Jensen, J.L., and Hein, J. (2005), "Bayesian Coestimation of Phylogeny and Sequence Alignment," *BMC Bioinformatics*, 6, 83.
- Nielsen, R. (2002), "Mapping Mutations on Phylogenies," *Systematic Biology*, 51, 729–739.
- Oehlert, G.W. (1992), "A Note on the Delta Method," *The American Statistician*, 46, 27–29.
- Pedersen, J.S., and Hein, J. (2003), "Gene Finding with a Hidden Markov Model of Genome Structure and Evolution," *Bioinformatics*, 19, 219–227.
- Redelings, B. D., and Suchard, M.A. (2005), "Joint Bayesian Estimation of Alignment and Phylogeny," *Systematic Biology*, 54, 401–418.
- Ren, F., and Yang, Z. (2005), "An Empirical Examination of the Utility of Codon-Substitution Models in Phylogeny Reconstruction," *Systematic Biology*, 54, 808–818.
- Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N., and Thorne, J.L. (2003), "Protein Evolution with Dependence Among Codons due to Tertiary Structure," *Molecular Biology and Evolution*, 20, 1692–1704.
- Siepel, A., and Haussler, D. (2004), "Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood," *Molecular Biology and Evolution*, 21, 468–488.
- Wei, G.C.G., and Tanner, M.A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.
- Yap, V.B., and Speed, T.P. (2004), "Modeling DNA Base Substitution in Large Genomic Regions from Two Organisms," *Journal of Molecular Evolution*, 58, 12–18.
- (2005), "Estimating Substitution Matrices," in *Statistical Methods in Molecular Evolution*, ed. R. Nielsen, New York: Springer.