research
reports

No. 455 February 2005
2005 / 02 / 28

Asger Hobolth
and Jens Ledet Jensen

Statistical Inference in Evolutionary Models of DNA Sequences via the EM Algorithm

department of
theoretical
statistics

university of
aarhus

ISSN 0902-9893

Statistical inference in evolutionary models of DNA sequences via the EM algorithm

Asger Hobolth[1] and Jens Ledet Jensen[2]

[1]*Bioinformatics Research Center, University of Aarhus, Denmark.*
[2]*Department of Theoretical Statistics, University of Aarhus, Denmark.*

Abstract

We describe statistical inference in continuous time Markov processes of DNA sequences related by a phylogenetic tree. The maximum likelihood estimator can be found by the EM algorithm and an expression of the information matrix is also derived. We provide closed-form analytical solutions of the EM algorithm and information matrix in the case of a reversible substitution process.

Key words: continuous time Markov chain, EM algorithm, information matrix, likelihood inference, molecular evolution.

Corresponding author: Asger Hobolth

Bioinformatics Research Center
University of Aarhus
Hoegh-Guldbergs Gade 10, Building 090
DK-8000 Aarhus C
Denmark

E-mail: asger@birc.au.dk
Phone: +45 8942 3099
Fax: +45 8942 3077

1 Introduction

The evolution of homologous DNA sequences can be described by continuous time Markov chains on a phylogenetic tree. A continuous time Markov chain is characterized by a substitution rate matrix, and the phylogenetic tree summarizes the relationship between the species in terms of branch lengths (time before divergence) and common ancestors. The DNA sequences are only observed at the tip of the leaves, and information on substitution events (time and type) and branch lengths are missing.

The 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) and Yap and Speed (2004) describe the EM algorithm for estimating unconstrained reversible substitution rate matrices from homologous DNA sequences. Here we consider general substitution rate matrices so that we may have constraints, and the substitution process need not be reversible. We draw attention to variance estimation and use established statistical theory to derive expressions for the information matrix. In case of a reversible substitution process we provide closed-form solutions.

The paper is organized as follows. In Section 2 we describe the maximum likelihood (ML) problem and the EM algorithm. The E-step is a matter of calculating conditional expectations of the log-likelihood based on the full data. Section 3 describes how to estimate the information matrix from conditional expectations of the gradient and curvature of the full likelihood score. Details in the derivation are provided in Appendix A. In Section 4 we consider continuous time Markov processes and calculate the conditional expressions needed for the E-step and information matrix. In Appendix B we assume a reversible substitution process and provide analytical solutions of the conditional expressions. In Section 5 we consider evolutionary models of DNA sequences related by a phylogenetic tree.

2 The ML problem and the EM algorithm

We denote the full data by x and observed data by $y = y(x)$. The parameter of interest is a vector θ. The full likelihood of θ based on x is denoted by $L(\theta; x)$, and the marginal likelihood of θ based on y is denoted by

$$L(\theta; y) = L(\theta; x)/L(\theta; x|y). \tag{1}$$

The EM algorithm is attractive in situations where finding the maximum likelihood estimate (MLE) $\hat{\theta}$ 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 function

$$G(\theta; \theta_0) = \mathrm{E}_{\theta_0}[\log L(\theta; x)|y] \tag{2}$$

is calculated, and in the M-step a new parameter value θ_1 is obtained as the value of θ that maximizes $G(\theta; \theta_0)$. The algorithm converges to a local maximum $\hat{\theta}$ of $L(\theta; y)$.

3 Estimating the information

3.1 General case

Following Louis (1982) we let

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

be the likelihood score and information matrix based on the full likelihood. As is shown in Appendix A the information matrix based on the observed data y is given by

$$I(\theta;y)=\mathrm{E}_\theta[I(\theta;x)|y]-\mathrm{V}_\theta[S(\theta;x)|y], \tag{3}$$

where

$$\mathrm{V}_\theta[S(\theta;x)|y]=\mathrm{E}_\theta[S(\theta;x)S^*(\theta;x)|y]-\mathrm{E}_\theta[S(\theta;x)|y]\,\mathrm{E}_\theta[S(\theta;x)|y]^*$$

is the conditional variance of the full likelihood score. Thus the information matrix based on data y can be computed from the conditional expectations of the gradient and curvature of the full likelihood score.

3.2 Independence case

When $x_1,\ldots,x_n$ are independent and $y_i(x)=y_i(x_i)$, the function $G(\theta;\theta_0)$, the full likelihood score $S(\theta;x)$ and the full data information matrix $I(\theta;x)$ become sums and

$$I(\theta;y)=\sum_{i=1}^n \mathrm{E}_\theta[I_i(\theta;x_i)|y_i]-\sum_{i=1}^n \mathrm{V}_\theta[S_i(\theta;x_i)|y_i].$$

If $x_1,\ldots,x_n$ are also identically distributed it is useful to arrange the terms into classes determined by the value of y_i. Indicating the class by $a\in\mathcal{C}$ and letting $\nu(a)$ be the number of terms in class a we get

$$G(\theta;\theta_0)=\sum_{a\in\mathcal{C}}\nu(a)\,\mathrm{E}_{\theta_0}[\log L(\theta;x_a)|y_a], \tag{4}$$

and we also find

$$I(\theta;y)=\sum_{a\in\mathcal{C}}\nu(a)\big(\,\mathrm{E}_\theta[I_a(\theta;x_a)|y_a]-\mathrm{V}_\theta[S_a(\theta;x_a)|y_a]\big). \tag{5}$$

4 Continuous time Markov chains

4.1 Likelihood

We consider a continuous time Markov chain $\{X(t)\}_{t\in[0,T]}$ on a finite state space of size m with rate matrix Q conditioned on the initial value X_0. We only observe

$Y = y(X) = X_T$. The rate matrix is parameterized by θ and the full likelihood from a complete observation of the Markov chain is given by

$$L(\theta; \{x(t)\}_{t\in[0,T]}) = \prod_{a=1}^{m} e^{T(a)Q_\theta(a,a)} \prod_{a=1}^{m}\prod_{b\neq a} Q_\theta(a,b)^{N(a,b)},$$

where $T(a)$ is the total time spent in state a and $N(a,b)$ is the number of substitutions of a with b. Thus the sufficient statistics R is a m^2-dimensional vector given by the total time spent in a state and the number of substitutions between states

$$R = \Big(T(1), \ldots, T(m), N(1,2), \ldots, N(m, m-1)\Big)^*. \tag{6}$$

The full log-likelihood becomes

$$\log L(\theta; R) = \sum_{a=1}^{m} T(a)Q_\theta(a,a) + \sum_{a=1}^{m}\sum_{b\neq a} N(a,b)\log Q_\theta(a,b) = q(\theta)^* R,$$

where

$$q(\theta) = \Big(Q_\theta(1,1), Q_\theta(2,2), \ldots, Q_\theta(m,m), \log Q_\theta(1,2), \ldots, \log Q_\theta(m, m-1)\Big)^*,$$

and where $Q_\theta(a,a) = -\sum_{b\neq a} Q_\theta(a,b)$. The function $G(\theta;\theta_0)$ in (2) is now given by

$$G(\theta,\theta_0) = \mathrm{E}_{\theta_0}[q(\theta)^* R | x_0, x_T] = q(\theta)^* \,\mathrm{E}_{\theta_0}[R|x_0, x_T], \tag{7}$$

and we need to calculate the conditional means of the sufficient statistics.

In order to calculate the observed data information matrix (3) we find that

$$\mathrm{E}_{\theta_0}[S(\theta;x)|x_0,x_T] = \frac{\partial q(\theta)^*}{\partial\theta}\,\mathrm{E}_{\theta_0}[R|x_0,x_T], \tag{8}$$

$$\mathrm{E}_{\theta_0}[I(\theta;x)|x_0,x_T] = -\sum_{k=1}^{m^2} \frac{\partial^2 q_k(\theta)}{\partial\theta\partial\theta^*}\,\mathrm{E}_{\theta_0}[R_k|x_0,x_T], \tag{9}$$

where $q_k(\theta)$ and R_k are coordinates of $q(\theta)$ and R, and

$$\mathrm{E}_{\theta_0}[S(\theta,x)S^*(\theta,x)|x_0,x_T] = \frac{\partial q(\theta)^*}{\partial\theta}\,\mathrm{E}_{\theta_0}[RR^*|x_0,x_T]\frac{\partial q(\theta)}{\partial\theta^*}, \tag{10}$$

and so we also need to calculate the conditional variances and covariances of the sufficient statistics. In the next subsection we describe how to calculate the conditional means, variances and covariances of the sufficient statistics.

4.2 Conditional means, variances and covariances of the sufficient statistics

In order to calculate $G(\theta;\theta_0)$ and the information matrix based on the observed data we need the conditional means, variances and covariances of the sufficient statistics. These are given by the following Theorem. The statements in the Theorem are conditional versions of Proposition 3.6-3.8 in Guttorp (1995) and can be shown using similar techniques as in that book.

Theorem (Conditional means, variances and covariances).
We have the following conditional means:

- *Time spent in state j*

$$\mathrm{E}[T(j)|X_0 = a, X_T = b] = \int_0^T P_{aj}(t)P_{jb}(T-t)\,dt/P_{ab}(T).$$

- *Number of transitions between states j and k*

$$\mathrm{E}[N(j,k)|X_0 = a, X_T = b] = Q(j,k)\int_0^T P_{aj}(t)P_{kb}(T-t)\,dt/P_{ab}(T).$$

- *Product of two times*

$$\begin{aligned}\mathrm{E}\big[T(j)T(k)|X_0 = a, X_T = b\big] = \\ \int_0^T\int_0^t \big[P_{aj}(u)P_{jk}(t-u)P_{kb}(T-t) \\ + P_{ak}(u)P_{kj}(t-u)P_{jb}(T-t)\big]\,du\,dt/P_{ab}(T).\end{aligned}$$

- *Product of two number of transitions*

$$\begin{aligned}\mathrm{E}\big[N(j_1,k_1)N(j_2,k_2)|X_0 = a, X_T = b\big] = \\ \Big\{\mathbf{1}((j_1,k_1)=(j_2,k_2))\cdot Q(j_1,k_1)\int_0^T P_{aj_1}(t)P_{k_1b}(T-t)\,dt \\ + Q(j_1,k_1)Q(j_2,k_2)\int_{t=0}^{t=T}\int_{u=0}^{u=t}\big[P_{aj_1}(u)P_{k_1j_2}(t-u)P_{k_2b}(T-t) \\ + P_{aj_2}(u)P_{k_2j_1}(t-u)P_{k_1b}(T-t)\big]\,du\,dt\Big\}/P_{ab}(T).\end{aligned}$$

- *Product of time and number of transitions*

$$\begin{aligned}\mathrm{E}[N(j,k)T(l)|X_0 = a, X_T = b] = \\ Q(j,k)\int_0^T\int_0^t \big[P_{aj}(u)P_{kl}(t-u)P_{lb}(T-t) \\ + P_{al}(u)P_{lj}(t-u)P_{kb}(T-t)\big]\,du\,dt/P_{ab}(T).\end{aligned}$$

In the case of a reversible process, the integrals in the Theorem can be written in closed form, as shown in Appendix B.

5 Models of DNA sequence evolution

5.1 Pairwise sequences

We consider data from Felsenstein (2004) page 207 where the evolution of $n = 500$ sites of two homologous DNA sequences are summarized in the frequency table ν given by Table 1.

	A	G	C	T	total
A	93	13	3	3	112
G	10	105	3	4	122
C	6	4	113	18	141
T	7	4	21	93	125
total	116	126	140	118	500

Table 1: Pairs of nucleotides for 500 sites.

5.1.1 Example: Kimura model

The Kimura (1980) model has rate matrix

$$Q = Q_{\alpha,\beta} = \begin{bmatrix} -\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha-2\beta & \beta & \beta \\ \beta & \beta & -\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha-2\beta \end{bmatrix}.$$

The model is reversible with uniform stationary distribution $\pi = (1/4, 1/4, 1/4, 1/4)$. We let $T = 1$ and for a single site we get the full log-likelihood

$$\log L(\alpha, \beta; R) = q(\alpha, \beta)^* R,$$

where R is given by (6) and

$$q(\alpha,\beta)^* = (\underbrace{-\alpha-2\beta, \ldots, -\alpha-2\beta}_{\text{length } 4}, \underbrace{\log\alpha, \log\beta, \ldots, \log\alpha}_{\text{length } 12}).$$

From (4) we get

$$G(\alpha, \beta; \alpha_0, \beta_0) = q(\alpha,\beta)^* \sum_a \sum_b \nu(a,b)\, \mathrm{E}_{\alpha_0,\beta_0}[R|x_0 = a, x_T = b].$$

The conditional means of the sufficient statistics are given by the Theorem in Section 4.2 and Appendix B. We may write the function as

$$G(\alpha, \beta; \alpha_0, \beta_0) = -\alpha n - 2n\beta + n_{\text{ts}} \log\alpha + n_{\text{tv}} \log\beta,$$

where n_{ts} and n_{tv} denote the aggregated expected number of transitions and transversions conditional on the observed data and parameter values α_0, β_0. The updating procedure is given by $\alpha_1 = n_{\text{ts}}/n,\ \beta_1 = n_{\text{tv}}/(2n)$, and using the Felsenstein data we get $\hat{\alpha} = 0.153$ and $\hat{\beta} = 0.037$. In order to compute the observed data information we combine (5) and expressions (8), (9), (10), to get

$$I(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} 2172.7 & 158.9 \\ 158.9 & 23431.5 \end{bmatrix}.$$

The Kimura model is so simple that it is possible to find an analytical expression for the observed data likelihood. Following Ewens and Grant (2001) page 378 the observed data likelihood is proportional to

$$(1 + e^{-4\beta} + 2e^{-2(\alpha+\beta)})^{n_0}(1 + e^{-4\beta} - 2e^{-2(\alpha+\beta)})^{n_1}(1 - e^{-4\beta})^{n_2},$$

where n_0 is the number of sites where the nucleotides in the two sequences are the same, n_1 is the number of sites where a purine (pyrimidine) occurs in the ancestral sequence and the other purine (pyrimidine) occurs in the descendant sequence, and n_2 is the number of sites where a purine occurs in one sequence and a pyrimidine in the other. Maximization of this function leads to the above estimates, and the Hessian of minus the log-likelihood evaluated at the maximum gives the above information matrix.

5.2 Multiple sequences

Now consider the case of four sequences related by a phylogenetic tree as illustrated in Figure 1. We get the single site full log-likelihood

$$\log L(\theta; R) = \sum_{i=1}^{5} \Big\{ \sum_{a=1}^{m} T^i(a) Q^i(a,a) + \sum_{a=1}^{m} \sum_{b \neq a} N^i(a,b) \log Q^i(a,b) \Big\} = \sum_{i=1}^{5} q^i(\theta)^* R^i,$$

where $R = (R^1, \ldots, R^5)$ with R^i being the sufficient statistic on lineage i consisting of $T^i(a)$ (total time spent in state a on lineage i) and $N^i(a,b)$ (number of transitions from a to b on lineage i) and $q^i(\theta)$ is the corresponding parameterisation of the rate matrix on lineage i. Letting $y = (y^1, y^2, y^3, y^4)$ be the observed data at the tip of the leaves and letting $a(i), d(i)$ be the ancestral and descendant values at the two ends of lineage i we get

$$\begin{aligned} G(\theta; \theta_0) &= \mathrm{E}_{\theta_0}[\log L(\theta; R)|y] = \sum_{i=1}^{5} \mathrm{E}_{\theta_0}[q^i(\theta)^* R^i|y] \\ &= \sum_{i=1}^{5} \sum_{a(i),d(i)} q^i(\theta)^* \, \mathrm{E}_{\theta_0}[R^i|y, a(i), d(i)] P_{\theta_0}(a(i), d(i)|y). \end{aligned}$$

In the E-step we therefore need to calculate conditional mean values on each lineage. Conditioning on $a(i)$ and $d(i)$, the conditional mean values are determined by the Theorem in Section 4.2. For example the mean number of transitions between two states a and b on lineage 1 is given by (recall Figure 1)

$$\mathrm{E}_\theta\big[N^1(a,b)\big|y, y^1, z^1\big] = \mathrm{E}_\theta\big[N^1(a,b)\big|y^1, z^1\big].$$

Furthermore the probabilities $P_\theta(a(i), d(i)|y)$ are easily calculated using Felsenstein's peeling algorithm (Felsenstein, 1981). For multiple sites formula (4) apply.

In order to compute the information matrix (5) we first calculate the full likelihood score and full information matrix

$$S(\theta; R) = \sum_{i=1}^{5} \frac{\partial q^i(\theta)^*}{\partial \theta} R^i \quad \text{and} \quad I(\theta; R) = -\sum_{i=1}^{5} \sum_{k=1}^{m^2} \frac{\partial^2 q^i_k(\theta)}{\partial \theta \partial \theta^*} R^i_k.$$

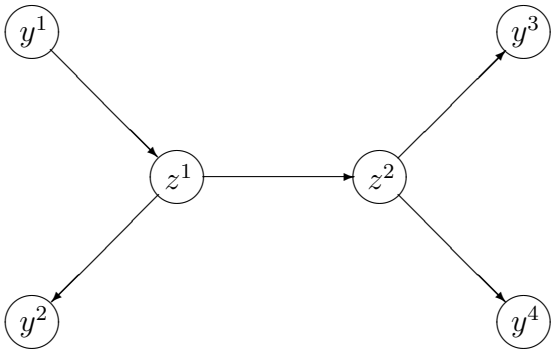

Figure 1: Phylogenetic tree relating four species. The observed data $y = (y^1, y^2, y^3, y^4)$ for a single site are the nucleotides at the tip of the leaves. The complete data consists of all substitution events at the branches of the tree. The (unobserved) nucleotides at the two inner nodes are denoted $z = (z^1, z^2)$.

The first term in (5), determined by $\mathrm{E}_\theta[I(\theta; R)|y]$, involves the same conditional means as described in the E-step above. In order to calculate the second term in (5), determined by $\mathrm{V}_\theta[S(\theta; R)|y]$, we note that

$$S(\theta; R)S^*(\theta; R) = \sum_{i=1}^{5}\sum_{j=1}^{5} \frac{\partial q^i(\theta)^*}{\partial \theta} R^i (R^j)^* \frac{\partial q^j(\theta)}{\partial \theta^*}.$$

When $i = j$ we find

$$\mathrm{E}_\theta[R^i(R^i)^*|y] = \sum_{a(i),d(i)} \mathrm{E}_\theta[R^i(R^i)^*|y, a(i), d(i)] P_\theta(a(i), d(i)|y),$$

and in the case $i \neq j$ we find

$$\mathrm{E}_\theta[R^i(R^j)^*|y] = \sum_{a(i),d(i),a(j),d(j)} \mathrm{E}_\theta[R^i|y, a(i), d(i)]\, \mathrm{E}_\theta[R^j|y, a(j), d(j)]^* P_\theta(a(i), d(i), a(j), d(j)|y).$$

Thus the second term is calculated from conditional means, variances and covariances as determined by the Theorem in Section 4.2 and from the probabilities of the inner nodes as determined by Felsenstein's peeling algorithm.

These considerations hold for any number of sequences related by a phylogenetic tree.

5.2.1 Example: Hasegawa-Kishino-Yano (HKY) model

The Hasegawa, Kishino and Yano (1985) model has rate matrix

$$Q = Q(\alpha, \beta) = \begin{bmatrix} \cdot & \alpha\pi_{\mathtt{G}} & \beta\pi_{\mathtt{C}} & \beta\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & \cdot & \beta\pi_{\mathtt{C}} & \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \beta\pi_{\mathtt{G}} & \cdot & \alpha\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \beta\pi_{\mathtt{G}} & \alpha\pi_{\mathtt{C}} & \cdot \end{bmatrix}.$$

The model is reversible with stationary distribution $(\pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}})$. We consider the case where we have the HKY rate matrix on each lineage, but lineage specific branch lengths. Thus the rate matrix Q^1 on lineage 1 is given as above and furthermore we have the constraints

$$Q^i = \tau_i Q^1, \quad i = 2, \ldots, 5.$$

The parameters are therefore $\theta = (\alpha, \beta, \tau_2, \ldots, \tau_5)$ and we get

$$q^i(\theta)^* = \Big(\underbrace{-\tau_i(\alpha\pi_{\mathtt{G}} + \beta(\pi_{\mathtt{C}} + \pi_{\mathtt{T}})), \ldots, -\tau_i(\beta(\pi_{\mathtt{A}} + \pi_{\mathtt{G}}) + \alpha\pi_{\mathtt{C}})}_{\text{length } 4}, \underbrace{\log(\tau_i\alpha\pi_{\mathtt{G}}), \ldots, \log(\tau_i\alpha\pi_{\mathtt{C}})}_{\text{length } 12} \Big),$$

$i = 1, \ldots, 5$, where $\tau_1 = 1$. The function $G(\theta; \theta_0)$ is, up to an additive constant, given by

$$G(\theta; \theta_0) = \sum_{i=1}^{5} \Big(- c_i\tau_i\alpha - \tilde{c}_i\tau_i\beta + k_i \log\alpha + \tilde{k}_i \log\beta + (k_i + \tilde{k}_i)\log\tau_i \Big),$$

where c_i, $\tilde{c}_i$, k_i and $\tilde{k}_i$ are conditional means dependent on θ_0. Differentiating with respect to α, β and $\tau_i, i = 2, \ldots, 5$, we get the updating scheme

$$\alpha = \frac{\sum_{i=1}^5 k_i}{\sum_{i=1}^5 c_i\tau_i}, \qquad \beta = \frac{\sum_{i=1}^5 \tilde{k}_i}{\sum_{i=1}^5 \tilde{c}_i\tau_i} \tag{11}$$

and

$$\tau_i = \frac{k_i + \tilde{k}_i}{c_i\alpha + \tilde{c}_i\beta}, \qquad i = 2, \ldots, 5. \tag{12}$$

In each M-step we iterate between (11) and (12) to obtain new parameter values of $(\alpha, \beta, \tau_2, \ldots, \tau_5)$. In the literature this iterative algorithm is called Zellner's two-stage procedure, and convergence properties are described in e.g. Lauritzen (1996, Appendix A4) and Drton (2004, Appendix A).

For illustration we consider a multiple alignment of homologous non-coding sequences from human, dog, mouse and rat. The alignment was obtained from the UCSC Genome Browser and can be seen at `www.daimi.au.dk/~asger/EMdata.html`. We use the human sequence to estimate the equilibrium frequencies, and the EM-algorithm described above to estimate the remaining parameters and obtain

$$\hat{\alpha} = 0.331, \quad \hat{\beta} = 0.071, \quad \hat{\tau}_2 = 1.174, \quad \hat{\tau}_3 = 0.315, \quad \hat{\tau}_4 = 0.359, \quad \hat{\tau}_5 = 1.931.$$

Furthermore we find the information matrix

$$I(\theta) = \begin{bmatrix} 765.89 & 78.40 & 58.46 & 73.52 & 71.17 & 44.28 \\ 78.40 & 10511.36 & 161.81 & 181.20 & 171.49 & 150.00 \\ 58.46 & 161.81 & 20.63 & 0.80 & -0.69 & 0.99 \\ 73.52 & 181.20 & 0.80 & 84.35 & 25.22 & 0.66 \\ 71.17 & 171.49 & -0.69 & 25.22 & 77.04 & 0.06 \\ 44.28 & 150.00 & 0.99 & 0.66 & 0.06 & 11.85 \end{bmatrix}.$$

We use the delta method (e.g. Oehlert, 1992) to obtain parameter estimates and standard deviations of the transition to transversion rate ratio $\kappa = \alpha/\beta$ and the size of the whole tree in terms of expected number of substitutions

$$\Gamma = \Big(2\alpha(\pi_{\mathtt{A}}\pi_{\mathtt{G}} + \pi_{\mathtt{C}}\pi_{\mathtt{T}}) + 2\beta(\pi_{\mathtt{A}} + \pi_{\mathtt{G}})(\pi_{\mathtt{C}} + \pi_{\mathtt{T}})\Big) \cdot \Big(1 + \sum_{i=2}^{5} \tau_i\Big),$$

and obtain

	estimate	std.dev.
κ	4.686	0.838
Γ	0.515	0.081

Thus the transition to transversion rate ratio is significantly larger than one, and we expect no substitutions in about half of the alignment columns.

Appendix A: Observed data information matrix

From the definition of conditional distributions (1) we get the gradient vector

$$\begin{aligned} S(\theta; y) = \frac{\partial \log L(\theta; y)}{\partial \theta} &= \frac{\partial \log L(\theta; x)}{\partial \theta} - \frac{\partial \log L(\theta; x|y)}{\partial \theta} \\ &= S(\theta; x) - \frac{1}{L(\theta; x|y)} \frac{\partial L(\theta; x|y)}{\partial \theta} \end{aligned} \tag{13}$$

and the second order matrix

$$\begin{aligned} \frac{\partial^2 \log L(\theta; y)}{\partial \theta \partial \theta^*} = \frac{\partial^2 \log L(\theta; x)}{\partial \theta \partial \theta^*} &- \frac{1}{L(\theta; x|y)} \frac{\partial^2 L(\theta; x|y)}{\partial \theta \partial \theta^*} \\ &+ \frac{1}{L(\theta; x|y)} \frac{\partial L(\theta; x|y)}{\partial \theta} \frac{1}{L(\theta; x|y)} \frac{\partial L(\theta; x|y)}{\partial \theta^*}. \end{aligned} \tag{14}$$

Under the usual regularity conditions we have

$$\mathrm{E}_\theta \Big[\frac{\partial \log L(\theta; x|y)}{\partial \theta} \Big| y \Big] = 0 \quad \text{and} \quad \mathrm{E}_\theta \Big[\frac{1}{L(\theta; x|y)} \frac{\partial^2 L(\theta; x|y)}{\partial \theta \partial \theta^*} \Big| y \Big] = 0,$$

so that from (13) we get

$$S(\theta; y) = \mathrm{E}_\theta[S(\theta; x)|y], \tag{15}$$

and so that the second term in (14) equals zero. Using (13) and (15) we can express (14) as

$$\begin{aligned} I(\theta; y) &= -\,\mathrm{E}_\theta \Big[\frac{\partial^2 \log L(\theta; y)}{\partial \theta \partial \theta^*} \Big| y \Big] \\ &= \mathrm{E}_\theta[I(\theta; x)|y] - \mathrm{E}_\theta[S(\theta; x) S^*(\theta; x)|y] + S(\theta; y) S^*(\theta; y) \\ &= \mathrm{E}_\theta[I(\theta; x)|y] - \mathrm{V}_\theta[S(\theta; x)|y], \end{aligned}$$

where

$$\mathrm{V}_\theta[S(\theta;x)|y] = \mathrm{E}_\theta[S(\theta;x)S^*(\theta;x)|y] - \mathrm{E}_\theta[S(\theta;x)|y]\,\mathrm{E}_\theta[S(\theta;x)|y]^*.$$

The information matrix based on data y need only be evaluated at $\theta = \hat{\theta}$ where $S(\theta;y) = 0$.

Appendix B: Closed form expressions of integrals

Let $\pi = (\pi_1, \ldots, \pi_m)$ denote the stationary distribution of the continuous time Markov process. In case of a reversibility we have detailed balance

$$Q(b,a) = \pi_a Q(a,b)/\pi_b,$$

which can also be written as

$$D_\pi Q = Q^* D_\pi,$$

where D_π is the diagonal matrix with π along its diagonal. Consider the matrix

$$S = D_\pi^{1/2} Q D_\pi^{-1/2}.$$

The matrix S is symmetric since

$$S^* = D_\pi^{-1/2} Q^* D_\pi^{1/2} = D_\pi^{-1/2}(Q^* D_\pi) D_\pi^{-1/2} = D_\pi^{-1/2}(D_\pi Q) D_\pi^{-1/2} = S,$$

and therefore it has real eigenvalues and real orthogonal eigenvectors. Let V be the real orthogonal matrix with eigenvectors as columns and D_λ the diagonal matrix of corresponding eigenvalues. It follows that

$$P(T) = e^{QT} = D_\pi^{-1/2} V e^{TD_\lambda} V^* D_\pi^{1/2}.$$

Conditional means are now found from

$$\int_0^T P_{ab}(t) P_{cd}(T-t)\,dt = \Big(\frac{\pi_b \pi_d}{\pi_a \pi_c}\Big)^{1/2} \sum_i V_{ai} V_{bi} \sum_j V_{cj} V_{dj} J_{ij}$$

where

$$J_{ij} = \begin{cases} T e^{\lambda_i T} & \lambda_i = \lambda_j \\ \frac{e^{\lambda_i T} - e^{\lambda_j T}}{\lambda_i - \lambda_j} & \lambda_i \neq \lambda_j. \end{cases}$$

Conditional variances and covariances are found from

$$\begin{aligned} &\int_0^T \int_0^t P_{ab}(u) P_{cd}(t-u) P_{ef}(T-t)\,du\,dt \\ &\quad = \Big(\frac{\pi_b \pi_d \pi_f}{\pi_a \pi_c \pi_e}\Big)^{1/2} \sum_i V_{ai} V_{bi} \sum_j V_{cj} V_{dj} \sum_k V_{ek} V_{fk} I_{ijk} \end{aligned}$$

where

$$I_{ijk} = \begin{cases} \frac{1}{2}T^2 e^{\lambda_i T} & \lambda_i = \lambda_j = \lambda_k \\ \dfrac{e^{\lambda_k T} - e^{\lambda_i T}}{(\lambda_i - \lambda_k)^2} + \dfrac{T e^{\lambda_i T}}{\lambda_i - \lambda_k} & \lambda_i = \lambda_j, \lambda_i \neq \lambda_k \\ \dfrac{T e^{T\lambda_k}}{\lambda_i - \lambda_j} - \dfrac{(e^{\lambda_j T} - e^{\lambda_k T})}{(\lambda_i - \lambda_j)(\lambda_j - \lambda_k)} & \lambda_i \neq \lambda_j, \lambda_i = \lambda_k \\ \dfrac{e^{\lambda_i T} - e^{\lambda_k T}}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)} - \dfrac{T e^{T\lambda_k}}{\lambda_i - \lambda_j} & \lambda_i \neq \lambda_j, \lambda_j = \lambda_k \\ \dfrac{e^{\lambda_i T} - e^{\lambda_k T}}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)} - \dfrac{(e^{\lambda_j T} - e^{\lambda_k T})}{(\lambda_i - \lambda_j)(\lambda_j - \lambda_k)} & \lambda_i \neq \lambda_j, \lambda_k \neq \lambda_i, \lambda_k \neq \lambda_j. \end{cases}$$

If the Markov process is not reversible it may not be possible to eigendecompose the rate matrix Q, and even if it is the eigendecomposition may involve complex eigenvalues and eigenvectors, which complicates evaluating the integrals.

References

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, **39**, 1–22.

Drton, M. (2004). *Maximum Likelihood Estimation in Gaussian AMP Chain Graph Models and Gaussian Ancestral Graph Models.* Ph.D. thesis, Department of Statistics, University of Washington

Ewens, W.J. and Grant, G.R. (2001). *Statistical Methods in Bioinformatics.* Springer, New York.

Felsenstein, J. (2004) *Inferring Phylogenies.* Sinauer Associates, Sunderland, MA, U.S.A.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences. *J. Mol. Evol.* **17**, 368–376.

Guttorp, P. (1995). *Stochastic modeling of scientific data.* Chapman & Hall, Suffolk, Great Britain.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.

Holmes, I. and Rubin, G.M. (2002). An Expectation Maximization algorithm for training hidden substitution models. *Journal of Molecular Biology* **317**, 757–768.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.

Lauritzen, S.L. (1996). *Graphical Models.* Clarendon Press, Oxford, UK.

Louis, A.T. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, **44**, 226–233.

Oehlert, G.W. (1992). A note on the delta method. *The American Statistician*, **46**, 27–29.

Yap, V.B. and Speed, T.P. (2004) Estimating Substitution Matrices. In: *Statistical Methods in Molecular Evolution* (ed. Rasmus Nielsen), to appear.