MaPhySto
The Danish National Research Foundation:
Network in Mathematical Physics and Stochastics

Research Report
no. 29 December 2004

Ole F. Christensen, Asger Hobolth and Jens L. Jensen: Pseudo-likelihood analysis of context-dependent codon substitution models

ISSN 1398-2699

Pseudo-likelihood analysis of context-dependent codon substitution models

Ole F. Christensen[1], Asger Hobolth[1] and Jens L. Jensen[2]

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

Abstract

We consider Markov processes for coding DNA sequence evolution. In context dependent models the instantaneous substitution rate for a codon depends on the neighboring codons. This makes the model analytically intractable, and previously Markov chain Monte Carlo methods have been used for statistical inference. We introduce an approximative estimation method based on pseudo-likelihood that makes inference analytically tractable. We demonstrate that the pseudo-likelihood estimates are very accurate, and from analyzing 348 human-mouse coding sequences we conclude that incorporating the `CpG` effect improves the model fits considerably.

Keywords: codon model, context dependence, `CpG` avoidance, EM-algorithm, maximum likelihood, pseudo-likelihood.

1 Introduction

For protein coding sequences, amino-acids are encoded by codons consisting of triplets of nucleotides. A commonly used model for coding sequences is the Goldman and Yang (GY) model described in Goldman and Yang (1994), where a basic assumption is that codon sites evolve independently. The `CpG` effect (Albert *et al.*, 2002, p. 434-435) refers to the fact that an excess of substitutions are observed for positions in a sequence where the nucleotides `C` and `G` are neighbors. The `CpG` effect violates the independent site assumption and motivates studying models where the substitution process is context dependent.

An extension of the GY model allowing `CpG` effects across codon boundaries is presented in Jensen and Pedersen (2000) and Pedersen and Jensen (2001). The neighbor dependence in the substitution rates implies that the transition probability for a sequence of length n can no longer be written as a product over the n codons. Thus we have to consider the sequence itself as a state in a Markov process and the corresponding rate matrix is of size $61^n \times 61^n$. The large dimension of the rate matrix makes the likelihood function intractable in practice, and Markov chain Monte Carlo (MCMC) methods are used for statistical inference. The use of MCMC methods is often time-consuming and problems such as slow convergence and poor mixing may arise.

The Jensen and Pedersen (2000) model consists of two components. The first component depends on the type of change (transition/transversion, synonymous/non-synonymous)

[1] MaPhySto — A Network in Mathematical Physics and Stochastics, funded by The Danish National Research Foundation

and the second component models the CpG effect. In this paper we consider a model similar to Jensen and Pedersen (2000), but we extend our model to allow the first component to be any reversible codon substitution model.

Statistical inference in the model may naturally be divided into two separate parts. The probability of observing two sequences x and y is

$$P(x, y) = P(x)P(y \mid x).$$

Firstly, we use the stationary distribution of sequence x, $P(x)$, to estimate the codon frequency and CpG parameters using maximum likelihood. Secondly, we construct a pseudo-likelihood (Besag, 1975) approximation of the conditional probability $P(y|x)$ that makes inference analytically tractable. The pseudo-likelihood is used to estimate the remaining parameters in the substitution rate matrix. We also present an EM-algorithm for computing the pseudo-likelihood estimates, which is useful in models with many parameters. An analysis of simulated data shows that the pseudo-likelihood estimates are very accurate. We analyst 348 human-mouse coding sequences from human chromosome 1 using different codon substitution models, and conclude that incorporating the CpG effect improves the fit considerably.

Our paper is in the spirit of Siepel and Haussler (2004) and we arrive at the same conclusion that substitution models with context dependence fit data considerably better than substitution models with site independence do. However, our paper differs from Siepel and Haussler (2004) and the follow-up paper by Jojic, Jojic, Geiger, Siepel, Haussler and Heckerman (2004) by having a proper stochastic process model for sequence evolution. Furthermore, instead of analyzing the merged alignments only, we also allow gene individual parameters.

Section 2 describes the model, and the human-mouse alignments are described in Section 3. Section 4 considers likelihood inference for one sequence, and in Section 5 we present pseudo-likelihood estimation for two sequences. In Section 6 we discuss inference for more than two sequences, and further extensions of the models and methods.

2 Context dependent codon models

We describe the evolution of a protein coding sequence as a stationary reversible time-homogeneous continuous time Markov process, where a change in the sequence consists of a change of one nucleotide only. The model is an extension of the Jensen and Pedersen (2000) model, and consists of two components.

Firstly, there is the codon 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 codons, and we call it the site independent part of the model. We consider two types of site independent reversible substitution models, namely the GY model (Goldman and Yang, 1994) and the general reversible substitution model

(REV). The GY model is given by the rate matrix

$$Q(a,b) = \begin{cases} 0 & \text{if } a \text{ and } b \text{ differ at more than one position} \\ \alpha\pi_b & \text{for synonymous transition} \\ \beta\pi_b & \text{for synonymous transversion} \\ \omega\alpha\pi_b & \text{for nonsynonymous transition} \\ \omega\beta\pi_b & \text{for nonsynonymous transversion,} \end{cases} \tag{2.1}$$

for $a \neq b$ where π_b is the codon frequency of codon b. The REV model assumes $Q(a,b) = 0$ when a and b differ at more than one position, and the reversibility condition $\pi_a Q(a,b) = \pi_b Q(b,a)$. The REV model has 263 parameters in Q and 60 free codon frequencies π_a.

Secondly, there is the `CpG` component determined by the `CpG` parameter $\lambda = (\lambda_{12}, \lambda_{23}, \lambda_{31})$. The parameter λ_{31} introduces dependence among codons. If $\lambda_{31} < 1$ the parameter introduces `CpG` avoidance across codon boundaries, if $\lambda_{31} > 1$ the parameter introduces `CpG` attraction, and if $\lambda_{31} = 1$ the model is a site independent model. Thus we can investigate whether a site independent codon model is appropriate by testing $\lambda_{31} = 1$. The parameters λ_{12} and λ_{23} introduce `CpG` avoidance (parameters less than 1) or attraction (parameters larger than 1) within codon positions (1,2) and (2,3), respectively. We note that these two parameters are confounded with the codon frequencies of codons with `CpG` at position (1,2) or (2,3). Therefore, we only include these two parameters when analyzing multiple genes with common codon frequencies, and in this case the average of the gene specific values is close to 1, with individual values modeling gene specific deviations from the general pattern.

We write a codon sequence x of n codons as $x = (x_1, \ldots, x_n)$ with $x_k = (x_k^1, x_k^2, x_k^3)$, where the upper index u in x_k^u indicates the position within the k'th codon, and let $\tilde{x}_k = (\tilde{x}_k^1, \tilde{x}_k^2, \tilde{x}_k^3)$ denote the new codon. The rate γ for such a change depends upon x_k as well as the nucleotide neighbors x_{k-1}^3 and x_{k+1}^1 and is given by

$$\begin{aligned} \gamma(\tilde{x}_k; x_{k-1}^3, x_k, x_{k+1}^1) = Q(x_k, \tilde{x}_k)&\lambda_{31}^{1_{\mathtt{CG}}(x_{k-1}^3, \tilde{x}_k^1) - 1_{\mathtt{CG}}(x_{k-1}^3, x_k^1)} \lambda_{12}^{1_{\mathtt{CG}}(\tilde{x}_k^1, \tilde{x}_k^2) - 1_{\mathtt{CG}}(x_k^1, x_k^2)} \\ &\times \lambda_{23}^{1_{\mathtt{CG}}(\tilde{x}_k^2, \tilde{x}_k^3) - 1_{\mathtt{CG}}(x_k^2, x_k^3)} \lambda_{31}^{1_{\mathtt{CG}}(\tilde{x}_k^3, x_{k+1}^1) - 1_{\mathtt{CG}}(x_k^3, x_{k+1}^1)}. \end{aligned} \tag{2.2}$$

Note that the rate depends on the neighbors through the parameter λ_{31} only. We assume that the boundary codons of a sequence are a start codon ($x_0 = \mathtt{ATG}$) and a stop codon ($x_{n+1} \in \mathtt{TAA}, \mathtt{TGA}, \mathtt{TGG}$).

A crucial feature of our model is that the stationary distribution can be determined explicitly. In Section 4 we use the stationary distribution to estimate codon frequencies and `CpG` parameters.

Theorem *If (Q, π) obeys detailed balance*

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

the model defined in (2.2) is reversible with stationary distribution

$$P(x) = \frac{1}{Z(\lambda, \pi)} \lambda_{12}^{2\sum_{k=1}^n 1_{\mathtt{CG}}(x_k^1, x_k^2)} \lambda_{23}^{2\sum_{k=1}^n 1_{\mathtt{CG}}(x_k^2, x_k^3)} \lambda_{31}^{2\sum_{k=2}^n 1_{\mathtt{CG}}(x_{k-1}^3, x_k^1)} \prod_{k=1}^n \pi_{x_k}, \tag{2.3}$$

where $Z(\lambda, \pi)$ is a normalizing constant.

Proof. The claim follows since detailed balance

$$\begin{aligned} &P(x)\gamma(\tilde{x}_k; x^3_{k-1}, x_k, x^1_{k+1}) \\ &\quad = P(x_1, \ldots, x_{k-1}, \tilde{x}_k, x_{k+1}, \ldots, x_n)\gamma(x_k; x^3_{k-1}, \tilde{x}_k, x^1_{k+1}), \end{aligned}$$

is fulfilled. □

An important parameter in a substitution model is the branch-length parameter τ which is the expected number of substitutions per codon (on a given branch). For the models with CpG considered here, the branch-length parameter is not easily expressed in terms of the other parameters in the model, and we therefore use an approximation which we derive in Appendix B.

3 Data

We consider a set of homologous human-mouse coding sequence alignments from human chromosome 1. The alignments were obtained from the NCBI data base from stringent human-mouse alignments on human chromosome 1 made using the 'build 34' human assembly vs. the 'NCBI Mouse Build 32' assembly. The Ensembl gene predictions ('ensGene' annotation file) were used to extract the coding parts of the alignments, and these exonic alignments within a gene were combined to one protein coding alignment. Alignments where one of the sequences had gaps were removed, as were alignments with length not a multiplum of three, and alignments where one of the sequences had internal stop codons. The resulting data consists of 348 human-mouse alignments with no internal stop codons and no gaps.

4 Analysis of one sequence

We now consider parameter estimation for the set of human sequences in the alignments described in Section 3. From the stationary distribution (2.3) we obtain the log-likelihood

$$l(\lambda, \pi) = -\log Z(\lambda, \pi) + n^{\mathtt{CpG}}_{12} \log \lambda^2_{12} + n^{\mathtt{CpG}}_{23} \log \lambda^2_{23} + n^{\mathtt{CpG}}_{31} \log \lambda^2_{31} + \sum_a n_a \log \pi_a, \qquad (4.1)$$

where $n^{\mathtt{CpG}}_{12}$, $n^{\mathtt{CpG}}_{23}$ and $n^{\mathtt{CpG}}_{31}$ is the number of CpG's at codon positions $(1,2)$, $(2,3)$ and $(3,1)$, respectively, and n_a is the number of times codon a appears in the sequence. An approximation of the normalizing constant $Z(\lambda, \pi)$ is presented in Appendix A.

The number of codons in a gene is typically too small to obtain reliable estimates of all 61 codon frequencies. Furthermore it is not possible to identify both π and $(\lambda_{12}, \lambda_{23})$ from one gene. Since the CpG effect is our main interest we use the same codon frequencies π for all 348 genes, but allow gene specific CpG parameters $(\lambda_{12}, \lambda_{23}, \lambda_{31})$. The common frequencies π and gene specific CpG parameters λ are obtained by maximizing the log-likelihood function numerically. The log-likelihood for a set of codon sequences is the sum of the individual log-likelihoods (4.1).

In the top row of Figure 1 we show the distribution of $(\hat{\lambda}_{12}, \hat{\lambda}_{23}, \hat{\lambda}_{31})$. The distributions of $\hat{\lambda}_{12}$ and $\hat{\lambda}_{23}$ are centered around 1 whereas the distribution of $\hat{\lambda}_{31}$ is centered around 0.50. In order to investigate whether the variation in the CpG parameter estimates $\hat{\lambda}_{12}$ is due to random fluctuation or gene specific CpG effect we tested $\lambda_{12} = 1$ against the alternative $\lambda_{12} \neq 1$. A similar test was carried out for λ_{23} and also for the CpG parameter across codon boundaries λ_{31}. P-values were calculated using the usual $\chi^2(1)$-approximation for this likelihood ratio test. The distribution of the p-values is shown in the middle row of Figure 1. The p-values of λ_{12} and λ_{23} are not uniformly distributed, implying that these parameters should be included in the model. The p-values of λ_{31} show the great importance of including CpG avoidance across codon boundaries. In the bottom row of Figure 1 we consider the correlation between the parameters. All three parameters are proportional to one another, and we use the linear relationship to collapse the three parameters into one single parameter, letting $\lambda_{31} = 0.5\lambda_{12} = 0.5\lambda_{23}$.

In Table 1 we summaries four different parameterizations of the stationary distribution for the 348 genes. For all parameterizations we assume common codon frequencies. Model A is the site independent model. Model B is the context dependent model with one single parameter λ_{31} describing CpG avoidance across codon boundaries in the merged set of genes. In model D we take CpG avoidance into account by fitting all three CpG parameters $(\lambda_{12}, \lambda_{23}, \lambda_{31})$ for each gene. This corresponds to the model analyzed above with results summarized in Figure 1. Finally model C is the collapsed model where $\lambda_{31} = 0.5\lambda_{12} = 0.5\lambda_{23}$ for each gene. The differences in log-likelihood with model A as reference are shown in the left plot of Figure 2. Bearing in mind that the difference between model A and B is only one single parameter, the remarkable difference in likelihood between these two models is perhaps the most striking feature. However, there is also a considerable gain in likelihood when modeling gene specific CpG avoidance as in model C. Recall that model D has many more free parameters than model C, and therefore the smaller difference in likelihood between these two models justifies the constraint $\lambda_{31} = 0.5\lambda_{12} = 0.5\lambda_{23}$.

Model	Parameters	N	Description
A	$\pi, \lambda_{12} = \lambda_{23} = \lambda_{31} = 1$	60	Merged genes, independent site
B	$\pi, \lambda_{12} = \lambda_{23} = 1, \lambda_{31}$	60+1	Merged genes, context dependent
C	$\pi, (\lambda_{31} = 0.5\lambda_{12} = 0.5\lambda_{23})_i$	60+1·348	Gene specific, constrained λ's
D	$\pi, (\lambda_{12}, \lambda_{23}, \lambda_{31})_i$	60+3·348	Gene specific, free λ's

Table 1: Models for the stationary distribution used in this paper. The models range from the simple site independent merged gene model to the general context dependent model with gene specific CpG effect. N: Number of free parameters in the stationary distribution.

The same conclusions are obtained from the Akaike Information Criterion (AIC) (Akaike, 1973) given in the right hand plot of Figure 2. The AIC supports the use of gene specific

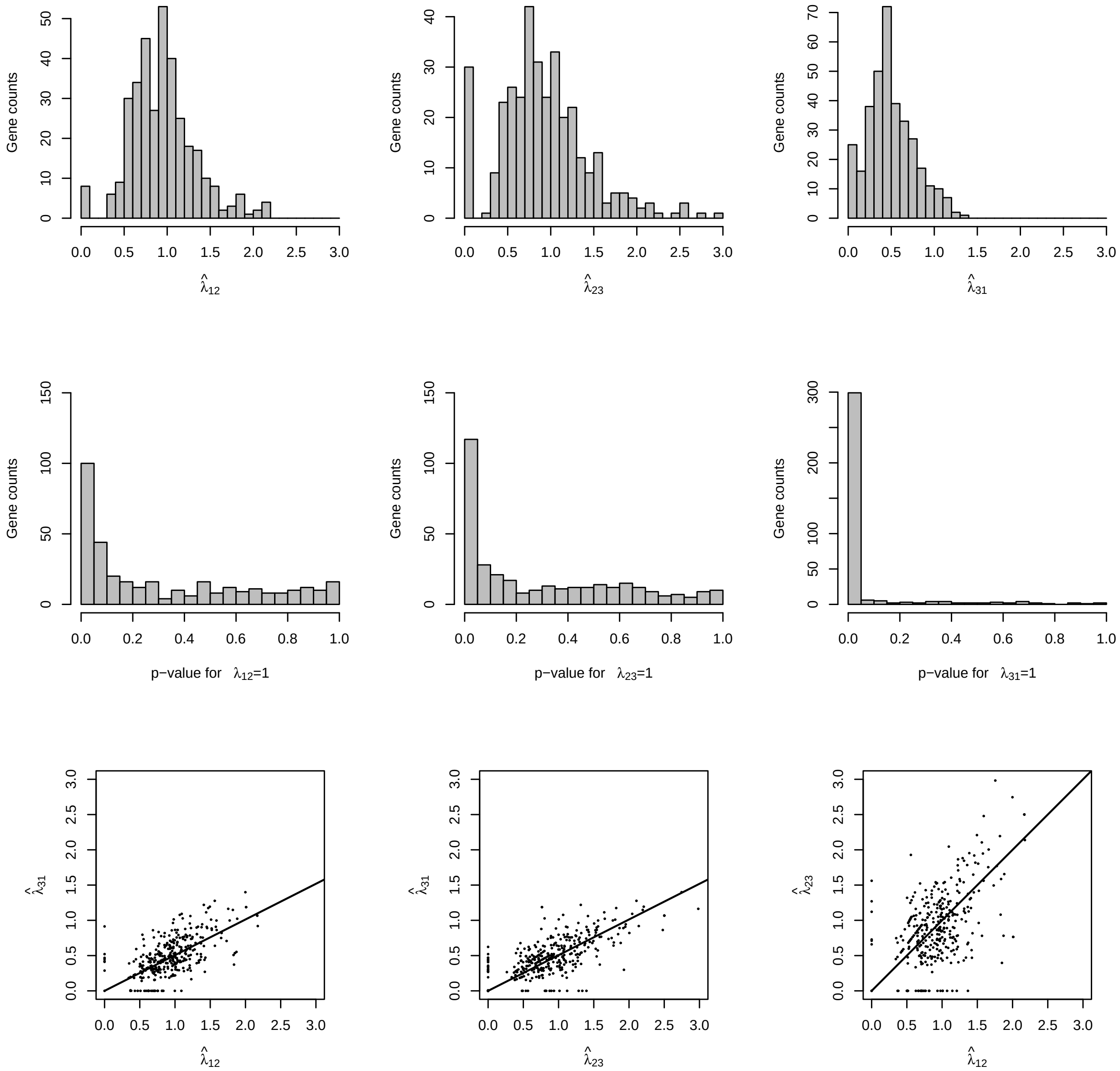

Figure 1: Top row: Histograms of $\hat{\lambda}_{12}, \hat{\lambda}_{23}$ and $\hat{\lambda}_{31}$. The distributions of $\hat{\lambda}_{12}$ and $\hat{\lambda}_{23}$ are centered around 1, while the distribution of $\hat{\lambda}_{31}$ is well below 1 (average 0.50). Middle row: Histograms of p-values for testing $\lambda_{12} = 1, \lambda_{23} = 1$ and $\lambda_{31} = 1$. The histograms of the p-values for testing $\lambda_{12} = 1$ and $\lambda_{23} = 1$ show that these are not uniformly distributed, implying that λ_{12} and λ_{23} should be included in the model. The histogram of the p-values for testing $\lambda_{31} = 1$ show the great importance of including `CpG` avoidance across codon boundaries. Bottom row: Scatter plot of $(\hat{\lambda}_{12}, \hat{\lambda}_{31})$ and line with slope 0.50, $(\hat{\lambda}_{23}, \hat{\lambda}_{31})$ and line with slope 0.50, and $(\hat{\lambda}_{12}, \hat{\lambda}_{23})$ and identity line.

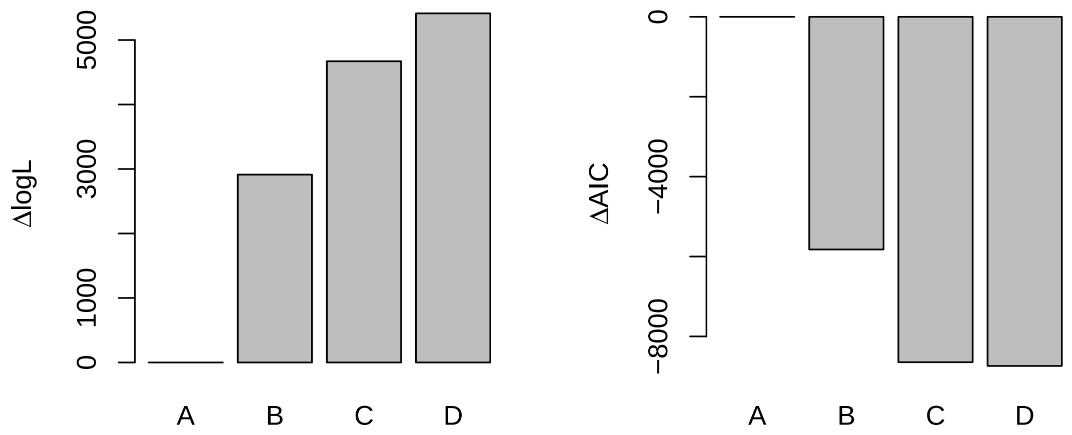

Figure 2: Left: Log likelihoods of various models with respect to the site independent model on the set of merged genes. We use the same codon frequencies for all genes, but the codon frequencies vary between models. A: Site independent model on the set of merged genes. B: Context dependent model on the set of merged genes. C: Context dependent model with one gene specific `CpG` parameter determined by $\lambda_{31} = 0.5\lambda_{12} = 0.5\lambda_{23}$. D: Context dependent model with three gene specific `CpG` parameters $(\lambda_{12}, \lambda_{23}, \lambda_{31})$. Right: Similar barplot showing the Akaike Information Criterion (AIC) in place of the log likelihood ($AIC = -2\log L + 2N$, where N is the number of parameters - see Table 1).

`CpG` parameters and only slightly supports the use of free `CpG` parameters compared to constrained `CpG` parameters. In the following we use the maximum likelihood estimates obtained from model C. Using this model instead of model D, we obtain more robust estimates of the `CpG` effect and we avoid outliers from genes with no observed `CpG` dinucleotides at a given position.

5 Analysis of two sequences

Consider two homologous coding sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ of length n codons, and assume the boundary codons are a start codon ($x_0 = y_0 =$ `ATG`) and stop codon ($x_{n+1}, y_{n+1} \in$ `TAA`, `TGA`, `TGG`). We consider the situation where the codon frequencies and `CpG` parameters are estimated as described in Section 4. We let the rate matrix Q be scaled in such a way that x has evolved into y in a time period of length one.

In this section we describe how to estimate the remaining substitution rate parameters from a pseudo-likelihood that approximates the conditional likelihood of y given x.

5.1 Pseudo-likelihood function

Besag (1975) introduced a pseudo-likelihood function, in the context of a random field, being a product of conditional likelihoods, each term representing the conditional likelihood for the observation at a particular site given the observations at the neighboring sites. Our pseudo-likelihood is in the same spirit although different in details. We use the probability of codon k being changed from x_k to y_k when the evolutionary history of the two flanking nucleotides are fixed.

To describe our pseudo-likelihood in detail let us first consider the full likelihood of the evolution changing x to y when all the substitution events are known. This is a product over the codon sites where each term is the contribution from the evolutionary events at this site. The term for site k corresponds to a Markov process for one site where the events at the two flanking nucleotide positions are fixed and we denote it by $L_{e_{k-1},e_{k+1}}(e_k|x_k)$, with e_{k-1} denoting the events at the left flanking nucleotide and e_{k+1} denoting the events at the right flanking nucleotide. In this site process we next find the marginal distribution for the codon y_k and write the corresponding k'th site marginal likelihood as $L_{e_{k-1},e_{k+1}}(y_k|x_k)$. We wish to use this term in our pseudo-likelihood. We do, however, not know e_{k-1} and e_{k+1} and need to approximate these. To this end we use the approximation that if $x^3_{k-1} = y^3_{k-1}$ then there are no substitutions in e_{k-1}, and if $x^3_{k-1} \neq y^3_{k-1}$ there is exactly one substitution in e_{k-1} taking place at time $t = 1/2$. A similar approximation is used for e_{k+1}. Denoting these approximate evolutionary events by $\tilde{e}_{k-1}$ and $\tilde{e}_{k+1}$ our pseudo-likelihood becomes

$$\prod_{k=1}^{n} L_{\tilde{e}_{k-1},\tilde{e}_{k+1}}(y_k \mid x_k). \tag{5.1}$$

Let us now consider in more detail each term in the pseudo-likelihood. The rate of a change (2.2) for a single codon depends on the left flanking nucleotide being a `C` or not and the right flanking nucleotide being a `G` or not. Thus there is a total of four different flanking situations. Let $c = 1$ ($c = 0$) indicate the presence (absence) of a `C` at the left flanking position, and index $g = 1$ ($g = 0$) denote the presence (absence) of a `G` at the right flanking position. Knowing the values at the flanking positions, the rate matrix for a single codon is given by the 61×61 rate matrix

$$Q^{cg}(a,b) = Q(a,b)\lambda_{31}^{c(\mathbf{1}_{\mathtt{G}}(b^1)-\mathbf{1}_{\mathtt{G}}(a^1))+g(\mathbf{1}_{\mathtt{C}}(b^3)-\mathbf{1}_{\mathtt{C}}(a^3))}\lambda_{12}^{\mathbf{1}_{\mathtt{CG}}(b^1,b^2)-\mathbf{1}_{\mathtt{CG}}(a^1,a^2)}\lambda_{23}^{\mathbf{1}_{\mathtt{CG}}(b^2,b^3)-\mathbf{1}_{\mathtt{CG}}(a^2,a^3)}. \tag{5.2}$$

If we define $c^x_k = \mathbf{1}_{\mathtt{C}}(x^3_{k-1})$, $c^y_k = \mathbf{1}_{\mathtt{C}}(y^3_{k-1})$, $g^x_k = \mathbf{1}_{\mathtt{G}}(x^1_{k+1})$, and $g^y_k = \mathbf{1}_{\mathtt{G}}(y^1_{k+1})$ to be the flanking situations in the x and y sequences, respectively, and if we use the approximate evolutionary events $\tilde{e}_{k-1}$ and $\tilde{e}_{k+1}$, the rate matrix for codon k is $Q^{c^x_k,g^x_k}$ for $0 \le t \le \frac{1}{2}$ and $Q^{c^y_k,g^y_k}$ for $\frac{1}{2} \le t \le 1$. The k'th term in the pseudo-likelihood (5.1) is therefore

$$L_{\tilde{e}_{k-1},\tilde{e}_{k+1}}(y_k|x_k) = \Big[\exp\left(Q^{c^x_k g^x_k}/2\right)\exp\left(Q^{c^y_k g^y_k}/2\right)\Big]_{x_k,y_k}. \tag{5.3}$$

Since there are only 16 different values of $I = (c^x, c^y, g^x, g^y) \in \{0,1\}^4$ we can rearrange the terms in (5.1) and use (5.3) to get the pseudo-likelihood

$$L_p(Q;\nu) = \prod_I \prod_{a,b} \left\{ \left[\exp\left(Q^{c^x g^x}/2\right) \exp\left(Q^{c^y g^y}/2\right)\right]_{a,b} \right\}^{\nu_I(a,b)}, \tag{5.4}$$

where $\nu_I(a,b)$ counts the number of codons with a substitution of b for a with the approximate flanking evolutionary events given by I.

The matrix exponentials in (5.3) can be computed using a symmetrization and an eigenvalue decomposition as in Schadt and Lange (2002). The computation uses the stationary frequencies for Q^{cg} which are

$$\pi_a^{cg} \propto \pi_a \lambda_{12}^{2\mathbf{1}_{\texttt{CG}}(a^1,a^2)} \lambda_{23}^{2\mathbf{1}_{\texttt{CG}}(a^2,a^3)} \lambda_{31}^{2(c\mathbf{1}_{\texttt{G}}(a^1)+g\mathbf{1}_{\texttt{C}}(a^3))}. \tag{5.5}$$

Further details can be found in Appendix C.

Parameter estimation can either be achieved by maximizing the pseudo-likelihood function (5.4) numerically or alternatively by an EM algorithm. When the number of parameters is small, as in the GY-model, numerical maximization is perhaps the easiest way of obtaining estimates, but in cases with many free parameters, such as the REV-model, the EM-algorithm becomes useful.

5.2 EM-algorithm for pseudo-likelihood estimation

The pseudo-likelihood function (5.4) has the same form as the likelihood function for sixteen independent processes, each of which corresponds to a model with independent codon evolution. The codon evolution for one of these sixteen processes is given by the rate matrix

$$Q^I(a,b,t) = \begin{cases} Q^{c^x g^x}(a,b) & 0 \le t \le 1/2 \\ Q^{c^y g^y}(a,b) & 1/2 < t \le 1, \end{cases} \tag{5.6}$$

with Q^{cg} defined in (5.2). This identification allows us to construct an EM-algorithm for estimating the parameters. An EM-algorithm for estimating substitution matrices (Holmes and Rubin, 2002; Yap and Speed, 2004) iterates between two steps. In the E-step the quantity

$$G(Q;Q_0,\nu) = \mathrm{E}_{Q_0}[\log L_c(Q;\nu)]$$

is calculated given a current estimate Q_0, data $\nu = \{\nu_I : I \in \{0,1\}^4\}$ and where L_c is the complete data likelihood. In the M-step $G(Q;Q_0,\nu)$ is maximized as a function of the parameters in Q.

5.2.1 E-step

The pseudo-likelihood is a product of sixteen terms and the function $G(Q;Q_0,\nu)$ then becomes a sum of terms

$$G(Q;Q_0,\nu) = \sum_I G(Q^I;Q_0^I,\nu_I),$$

where each term corresponds to an independent codon model with transition rates (5.6). In Appendix C (C.3) we show that each individual term can be written as

$$\begin{aligned} G(Q^I; Q_0^I, \nu_I) = &\sum_{a,b:a\neq b} \log Q^{c^x g^x}(a,b) w_I^1(a,b) + \sum_a Q^{c^x g^x}(a,a) w_I^1(a,a) \\ &+ \sum_{a,b:a\neq b} \log Q^{c^y g^y}(a,b) w_I^2(a,b) + \sum_a Q^{c^y g^y}(a,a) w_I^2(a,a), \end{aligned}$$

where the weights $w_I^1(\cdot,\cdot)$ and $w_I^2(\cdot,\cdot)$ are defined in (C.4) and (C.5) with ν^* replaced by ν^I and depend on the current estimate of the rate matrix Q_0 and the data ν_I only. Using that

$$Q^{cg}(a,a) = -\sum_{b:ab\neq a} Q^{cg}(a,b)$$

and substituting the expression (5.2) for Q^{cg} we get, up to a constant,

$$\begin{aligned} G(Q^I; Q_0^I, \nu_I) = &\sum_{a,b:a\neq b} \log Q(a,b) \Big[\sum_I (w_I^1(a,b) + w_I^2(a,b)) \Big] \\ &- \sum_{a,b:a\neq b} Q(a,b) \Big[\sum_I (k^{c^x g^x}(a,b) w_I^1(a,a) + k^{c^y g^y}(a,b) w_I^2(a,a)) \Big], \end{aligned}$$

where $k^{cg}(a,b) = \lambda_{31}^{c(\mathbf{1}_{\mathtt{G}}(b^1)-\mathbf{1}_{\mathtt{G}}(a^1))+g(\mathbf{1}_C(b^3)-\mathbf{1}_C(a^3))} \lambda_{12}^{\mathbf{1}_{\mathtt{CG}}(b^1,b^2)-\mathbf{1}_{\mathtt{CG}}(a^1,a^2)} \lambda_{23}^{\mathbf{1}_{\mathtt{CG}}(b^2,b^3)-\mathbf{1}_{\mathtt{CG}}(a^2,a^3)}$. To simplify notation, we observe that the form of $G(Q; Q_0, \nu)$ is

$$G(Q; Q_0, \nu) = \sum_{a,b:a\neq b} \log Q(a,b) w(a,b) - \sum_{a,b:a\neq b} Q(a,b) w(a,a), \qquad (5.7)$$

which is similar to the independent site estimating function, cf. Yap and Speed (2004, page 20, top equation). To summaries, the E-step is a matter of calculating the weight matrices $\tilde{w}_I$ and $\bar{w}_I$ for each of the sixteen cases and add appropriately scaled versions of these matrices to obtain the weight matrix w in (5.7).

We now derive the M-step for the context dependent GY-model and REV-model.

5.2.2 M-step for REV-model

In the case of a reversible substitution process we have detailed balance

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

and the derivative of (5.7) with respect to $Q(a,b)$ is

$$-(w(a,a) + \pi_a w(b,b)/\pi_b) + (w(a,b) + w(b,a))/Q(a,b).$$

Given a current estimate Q_0 we update Q by

$$Q(a,b) = \frac{w(a,b) + w(b,a)}{w(a,a) + \pi_a w(b,b)/\pi_b},$$

where Q_0 enters through $w(\cdot,\cdot)$.

5.2.3 M-step for GY-model

In the case of the GY-model (2.1) we get up to a constant

$$\begin{aligned} G(\alpha,\beta,\omega) &= G((\alpha,\beta,\omega);(\alpha_0,\beta_0,\omega_0),\nu) \\ &= -\alpha \sum_{a,b:s,ts} \pi_b w(a,a) - \beta \sum_{a,b:s,tv} \pi_b w(a,a) - \omega\alpha \sum_{a,b:ns,ts} \pi_b w(a,a) \\ &\quad - \omega\beta \sum_{a,b:ns,tv} \pi_b w(a,a) + \log\alpha \sum_{a,b:s,ts} w(a,b) + \log\beta \sum_{a,b:s,tv} w(a,b) \\ &\quad + \log(\omega\alpha) \sum_{a,b:ns,ts} w(a,b) + \log(\omega\beta) \sum_{a,b:ns,tv} w(a,b) + \sum_{a,b:a\neq b} \log(\pi_b) w(a,b), \end{aligned}$$

where e.g. $\{a,b : s,ts\}$ is the set of pairs (a,b) that differ at one position and where the substitution of a with b is a synonymous transition. Introducing a shorter notation, $G(\alpha,\beta,\omega)$ is on the form

$$-\alpha k_1 - \beta k_2 - \omega\alpha k_3 - \omega\beta k_4 + \log\alpha c_1 + \log\beta c_2 + \log(\omega\alpha)c_3 + \log(\omega\beta)c_4 + c_5,$$

where $k_1,\ldots,k_4,c_1,\ldots,c_5$ are constants. Differentiating with respect to the three parameters, and setting the partial derivatives equal to zero we obtain the updating

$$\hat{\alpha} = (c_1+c_3)/(k_1+\hat{\omega}k_3), \quad \hat{\beta} = (c_2+c_4)/(k_2+\hat{\omega}k_4),$$

with

$$\hat{\omega} = \frac{-((c_1-c_4)k_2k_3 + (c_2-c_3)k_1k_4) + \sqrt{D}}{2(c_1+c_2)k_3k_4},$$

where $D = ((c_1-c_4)k_2k_3 + (c_2-c_3)k_1k_4)^2 + 4(c_1+c_2)(c_3+c_4)k_1k_2k_3k_4$.

5.3 Investigation of the accuracy of the pseudo-likelihood approximation

For the GY model with CpG we investigate the accuracy of the pseudo-likelihood approximation by a simulation study. We let $\lambda_{12} = \lambda_{23} = 1$, and the frequency parameters $\pi_a = 1/61$, and for different combinations of α, β, ω and λ_{31} we simulate sequences of length 1000 codons from the model. The parameter values we consider are the 48 combinations of branch-length $\tau = 0.04, 0.2, 1$, transition/transversion ratio $\kappa = \alpha/\beta = 2, 4$, non-synonymous/synonymous rate $\omega = 0.1, 2$ and CpG avoidance parameter $\lambda_{31} = 0.01, 0.05, 0.2, 0.6$. In the estimation we use the true values of π and λ_{31} and estimate (τ,κ,ω) using both the approximative estimation procedure in Section 5 and the MCMC-EM procedure in Jensen (2005). Figure 3 shows the parameter estimates obtained by pseudo likelihood versus the exact estimates obtained by MCMC-EM. The estimates obtained by pseudo likelihood are accurate for a range of realistic parameter values. As expected, the estimates are most accurate when the two sequences are not too distant, i.e. the estimated branch-length is small.

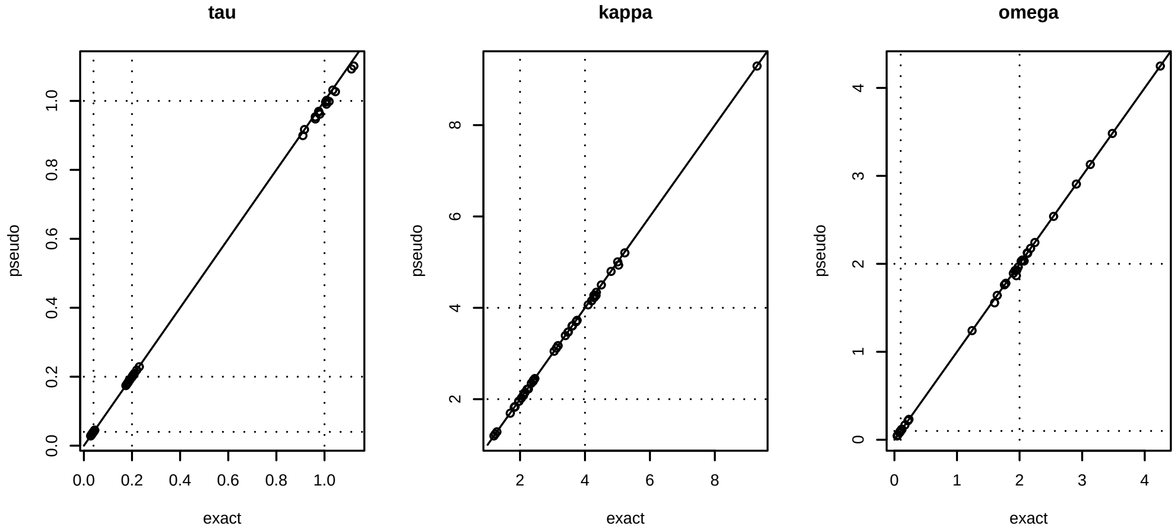

Figure 3: Comparison of estimates obtained by pseudo likelihood with estimates obtained by the MCMC-EM procedure in Jensen (2005). Dotted lines represent values used for the simulated data. Left : Branch-length τ. Middle : The transition/transversion ratio κ. Right : The non-synonymous/synonymous rate ω.

5.4 Results for data example: human-mouse alignments

For the 348 human-mouse alignments in Section 3 we investigate the effect of including CpG in the GY model and the REV model, respectively.

For the GY model Figure 4 shows the results of this comparison. For the model without CpG we used π equal to total observed frequencies, and for the model with CpG we used π, λ estimated in Section 4 under the model $\lambda_{31} = 0.5\lambda_{12} = 0.5\lambda_{23}$. We see that the differences in the parameter estimates are not large, but in general the transition/transversion ratio κ is slightly increased when including CpG (79.6% of cases), the non-synonymous/synonymous rate ω is slightly decreased (92.8% of cases) and the branch length parameter τ is slightly decreased (87.1% of cases).

The log-likelihood from (5.4) is increased for 71.6% of the alignments, and decreased for the remaining ones, when including CpG; the average increase in the log-likelihood is 2.20.

The REV model has too many parameters to be fitted on a single short alignment. We therefore consider a model where individual alignments follow the same REV model, but have individual branch lengths. The results obtained are that the branch length is decreased (78.4% of cases) when including CpG. Also, the log-likelihood is increased for 75% of the alignments when including CpG; the average increase in the log-likelihood is 2.90.

In Figure 5 we compare the total log-likelihood for the different models. In addition to the models discussed previously, we also consider the model where all alignments follow

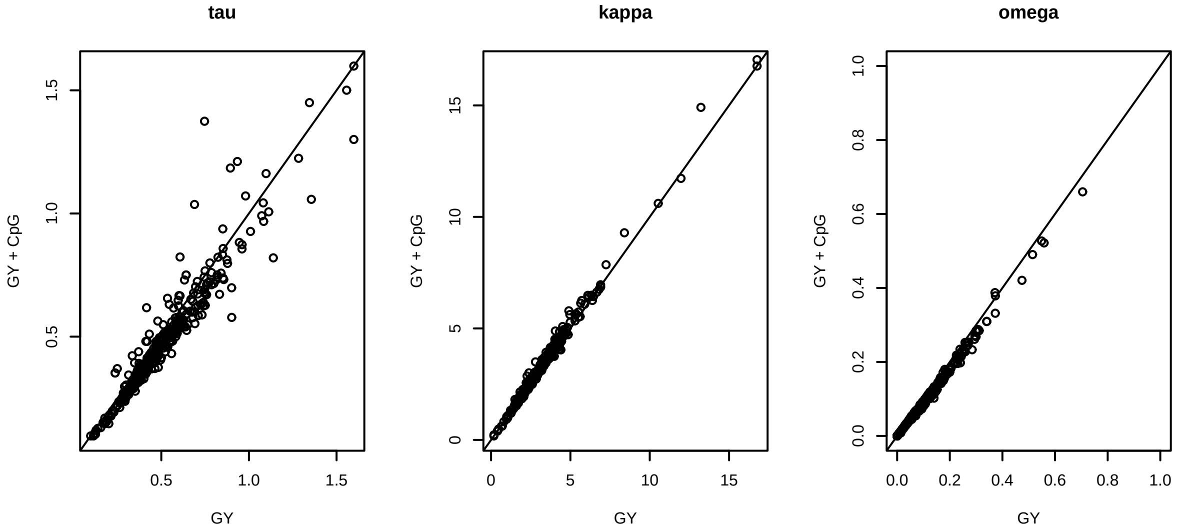

Figure 4: Comparison for the GY model with and without `CpG` effect. Left : The transition/transversion ratio κ. Middle : the non-synonymous/synonymous rate ω. Right : branch-length τ.

the same REV model, with and without `CpG` effect. From Figure 5 it is clear that including `CpG` results in a remarkable total increase in the log-likelihood for all three types of models. Another striking observation is the poor performance of the GY model with gene-specific parameters (where the total number of parameters is $3 \times 348 = 1044$), compared to the REV model (263 parameters) and the REV model with individual branch-length parameters ($263 + 348 - 1 = 610$ parameters). This seems to suggest that it is more important to model the overall substitution patterns of many genes compared to modeling individual gene specific substitution patterns. That result may throw some doubts on the routine use of the GY model as the standard codon model, but we note that the main use of the GY model has been to detect genes with an unusual substitution pattern, i.e. $\omega > 1$ corresponding to positive selection. Finally, we note that here model comparison is used in a somewhat informal way, since most of the models compared are not nested. However, we believe that the conclusions from the comparison are still valid, due to the large differences in likelihood.

6 Discussion

We have provided a pseudo-likelihood method for inference in codon models with `CpG` effects, which should make these models much more useful in practice. The method is very accurate for two sequences and for the range of parameter values in the data set we have considered.

The pseudo-likelihood function (5.4) can be extended to multiple species, and here

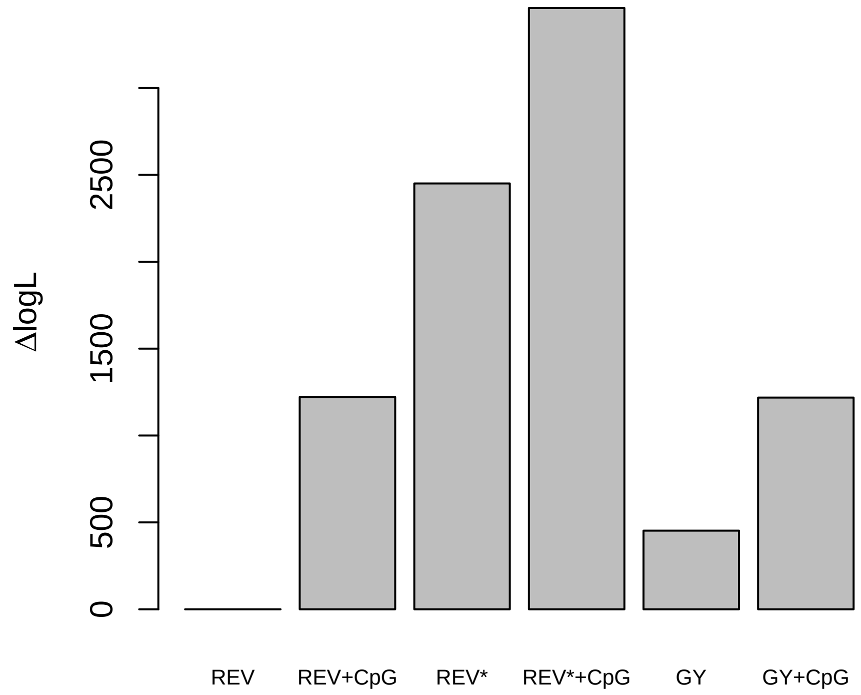

Figure 5: Log-likelihood values for three types of models, with and without CpG effects : The general reversible model, assuming same model for all genes (REV); The general reversible model, but allowing gene specific branch length (REV*); The Goldman and Yang model with all parameters being gene specific (GY).

we discuss the extension to three sequences. Assume that we have observed three codon sequences $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, x_n)$ and $z = (z_1, \dots, z_n)$, and assume that the parameters π and λ in the stationary distribution have already been estimated from sequence x. We will denote the unobserved sequence at the inner node $v = (v_1, \dots, v_n)$. Following the derivation in Section 5.1 the pseudo likelihood given x becomes

$$L_p(Q;\nu) = \prod_I \prod_{a,b,c} \Big(\sum_{c^v,g^v} \sum_d L_{c^x,g^x;c^v,g^v}(d \mid a) L_{c^v,g^v;c^y,g^y}(b \mid d) L_{c^v,g^v;c^z,g^z}(c \mid d) \\ \times P(c^v \mid c^x, c^y, c^z) P(g^v \mid g^x, g^y, g^z) \Big)^{\nu_I(a,b,c)},$$

where $I = (c^x, g^x, c^y, g^y, c^z, g^z) \in \{0,1\}^6$, $\nu_I(a,b,c)$ is the three dimensional array consisting of observed codon substitutions given flanking situation I, and the functions $L_{c^x,g^x;c^v,g^v}$, $L_{c^v,g^v;c^y,g^y}$ and $L_{c^v,g^v;c^z,g^z}$ are as in (5.3). Furthermore,

$$P(c^v|c^x,c^y,c^z) = P(\mathbf{1}_{\texttt{C}}(v_k^3) = c^v \mid \mathbf{1}_{\texttt{C}}(x_k^3) = c^x, \mathbf{1}_{\texttt{C}}(y_k^3) = c^y, \mathbf{1}_{\texttt{C}}(z_k^3) = c^z)$$

is the probability of presence ($c^v = 1$) or absence ($c^v = 0$) of a C at the third position at the inner node given the third positions at the leaves, and $P(g^v|g^x,g^y,g^z)$ is similarly defined, as the probability of presence ($g^v = 1$) or absence ($g^v = 0$) of a G at he first position at the inner node given the first positions at the leaves. In practice we will approximate $P(c^v|c^x,c^y,c^z)$ and $P(g^v|g^x,g^y,g^z)$ by estimating parameters under for example an independent nucleotide site model, and calculating these probabilities under this model.

An explanation of CpG effect in vertebrates is given by the CpG-methylation-deamination process; which is a non-reversible phenomenon. Non-reversible context dependent nucleotide models are considered in Arndt, Burge and Hwa (2003), Lunter and Hein (2004) and Hwang and Green (2004). It seems worth studying non-reversible context dependent models for codons, and develop analytically approximative inference methods either along the lines in this paper or along the lines in Lunter and Hein (2004).

Codon models with context dependencies arising from global properties of the sequence, such as dependence among codons due to tertiary structure of the proteins (Robinson, Jones, Kishino, Goldman and Thorne, 2003) are not covered by the techniques in this paper, since we only consider neighboring dependence. Analytically tractable inference techniques for such models still need to be developed.

Acknowledgements

The authors wish to thank Mikkel Schierup for encouragement and support throughout this project. Ole F. Christensen acknowledges financial support from SJVF grant 2052-01-0032.

Appendix A: Normalizing constant for stationary distribution

In order to calculate the normalizing constant $Z(\lambda,\pi)$ in (2.3) we write the stationary distribution as a Markov chain along the sequence and apply Jensen and Pedersen (2000) Section 4. First we rewrite the stationary measure

$$\begin{aligned} P(x) &= \frac{1}{Z}\pi_{x_1}\lambda_{12}^{2\times 1_{\texttt{CG}}(x_1^1,x_1^2)}\lambda_{23}^{2\times 1_{\texttt{CG}}(x_1^2,x_1^3)}\left[\prod_{k=2}^{n}\pi_{x_k}\lambda_{12}^{2\times 1_{\texttt{CG}}(x_k^1,x_k^2)}\lambda_{23}^{2\times 1_{\texttt{CG}}(x_k^2,x_k^3)}\lambda_{31}^{2\times 1_{\texttt{CG}}(x_{k-1}^3,x_k^1)}\right] \\ &= \frac{1}{Z}\phi(x_1)\prod_{k=2}^{n}S(x_{k-1},x_k), \end{aligned}$$

with obvious definitions of $\phi(x_1)$ and $S(x_{k-1},x_k)$. Now let s be the largest eigenvalue of S and r and l the corresponding right and left eigenvectors so that for all a we have

$$\sum_b S(a,b)r(b) = sr(a), \qquad \sum_b S(b,a)l(b) = sl(a).$$

If we normalize l such that $\sum_b l(b)r(b) = 1$ it follows from Jensen and Pedersen (2000) Section 4 that we can approximate Z by

$$\begin{aligned} Z &\approx s^{n-1} \sum_{a,b} \frac{r(a)}{r(b)} \phi(a) l(b) r(b) \\ &= s^{n-1} \Big(\sum_a r(a)\phi(a) \Big) \Big(\sum_b l(b) \Big). \end{aligned}$$

Appendix B: Formula for the branch-length parameter

As noted in Section 2 the branch-length parameter τ is not easily expressed in terms of the other parameters in the model, and we here derive an approximation. The branch-length parameter τ is defined as the average number of substitutions per codon, and for the model with `CpG` dependence, the branch length is

$$\begin{aligned} \tau &= \frac{1}{n} \sum_{x=(x_1,\ldots,x_n)} \sum_{j=1}^{n} \sum_{\tilde{x}_j} P(x) \gamma(\tilde{x}_j; x^3_{j-1}, j_k, x^1_{j+1}), \\ &= \frac{1}{n} \sum_{j=1}^{n} \sum_{x_1,\ldots,x_n} \phi(x_1) \prod_{k=2}^{n} S(x_{k-1}, x_k) \sum_{\tilde{x}_j} Q^{\mathbf{1}_{\mathtt{C}}(x^3_{j-1})\mathbf{1}_{\mathtt{G}}(x^1_{j+1})}(x_j, \tilde{x}_j)/Z \\ &= \frac{1}{n} \sum_{j=1}^{n} \psi_j, \end{aligned}$$

where Q^{cg} is defined in (5.2) and $\phi(a)$ and $S(a,b)$ are defined as in Appendix A, and with an obvious definition of ψ_j. Dividing the sum into three parts concerning $(x_1, \ldots, x_{j-2})$, (x_{j-1}, x_j, x_{j+1}) and $(x_{j+2}, \ldots, x_n)$, respectively, we get

$$\begin{aligned} \psi_j = \sum_{x_1,\ldots,x_n} \phi(x_1) \prod_{k=2}^{j-1} S(x_{k-1}, x_k) S(x_{j-1}, x_j) S(x_j, x_{j+1}) \sum_{\tilde{x}_j} Q^{\mathbf{1}_{\mathtt{C}}(x^3_{j-1})\mathbf{1}_{\mathtt{G}}(x^1_{j+1})}(x_j, \tilde{x}_j) \\ \times \prod_{k=j+2}^{n} S(x_{k-1}, x_k)/Z. \end{aligned}$$

Making a similar approximation as in Appendix A we obtain

$$\begin{aligned} \psi_j &\approx \sum_{x_1, x_{j-1}, x_j, x_{j+1}, x_n} s^{j-2} r(x_1)\phi(x_1) l(x_{j-1}) S(x_{j-1}, x_j) S(x_j, x_{j+1}) \sum_{\tilde{x}_j} Q^{\mathbf{1}_{\mathtt{C}}(x^3_{j-1})\mathbf{1}_{\mathtt{G}}(x^1_{j+1})}(x_j, \tilde{x}_j) \\ &\qquad \times s^{n-j-1} r(x_{j+1}) l(x_n) / (s^{n-1} \sum_{a,b} r(a)\phi(a) l(b)) \\ &= \sum_{x_{j-1}} l(x_{j-1}) \sum_{x_j, x_{j+1}} S(x_{j-1}, x_j) S(x_j, x_{j+1}) r(x_{j+1}) \sum_{\tilde{x}_j} Q^{\mathbf{1}_{\mathtt{C}}(x^3_{j-1})\mathbf{1}_{\mathtt{G}}(x^1_{j+1})}(x_j, \tilde{x}_j)/s^2. \end{aligned}$$

This term is independent of j, and hence we obtain

$$\tau \approx \sum_{a_1} l(a_1) \sum_{a_2,a_3} S(a_1,a_2)S(a_2,a_3)r(a_3) \sum_b Q^{\mathbf{1}_{\mathtt{C}}(a_1^3)\mathbf{1}_{\mathtt{G}}(a_3^1)}(a_2,b)/s^2.$$

Appendix C: E-step for inhomogeneous Markov model

Here we derive the E-step for estimating substitution rate matrices for an inhomogeneous reversible continuous time Markov process. Observations are n^* independent pairs $(x_k, y_k), k = 1, \dots, n^*$, of a continuous time Markov process with rate matrix

$$Q^*(A,B,t) = \begin{cases} Q^1(A,B) & 0 \le t \le 1/2 \\ Q^2(A,B) & 1/2 < t \le 1 \end{cases} \tag{C.1}$$

so that the continuous time Markov chain changes rate matrix at time $1/2$ from Q^1 to Q^2. We assume that Q^1 and Q^2 are reversible with stationary distribution π and η, respectively.

Suppose the Markov chain at site k experience m_k substitutions. Denote the substitution times $t_{k,\ell}$ and the new states $s_{k,\ell}, \ell = 1, \dots, m_k$. With $s_{k,0} = x_k$, $s_{k,m_k} = y_k$ and $t_{k,0} = 0$ the likelihood of observing the complete data given x is,

$$L_c(Q^*) = \prod_{k=1}^{n^*} \Bigg(\Bigg\{ \prod_{\ell=1}^{m_k} Q^*(s_{k,\ell-1}, s_{k,\ell}, t_{k,\ell}) e^{\int_{t_{k,\ell-1}}^{t_{k,\ell}} Q^*(s_{k,\ell-1},s_{k,\ell-1},t)dt} \Bigg\} e^{\int_{t_{k,m_k}}^{1} Q^*(y_k,y_k,t)dt} \Bigg).$$

In the E-step we calculate the function

$$\begin{aligned} G(Q^*;Q_0^*) &= \mathrm{E}_{Q_0^*}[\log L_c(Q^*) \mid x, y] \qquad\qquad \text{(C.2)} \\ &= \sum_{a,b} \nu^*(a,b) \sum_{A,B:A\neq B} \mathrm{E}_{Q_0^*}\Bigg[\sum_{\ell=1}^{m_k} \mathbf{1}_{\{s_{k,\ell-1}=A, s_{k,\ell}=B\}} \log Q^*(A,B,t_{k,\ell}) \Big| x_k = a, y_k = b \Bigg] \\ &\qquad + \sum_A \mathrm{E}_{Q_0^*}\Bigg[\sum_{\ell=1}^{m_k+1} \mathbf{1}_{\{s_{k,\ell-1}=A\}} \int_{t_{k,\ell-1}}^{t_{k,\ell}} Q^*(A,A,t)dt \Big| x_k = a, y_k = b \Bigg], \end{aligned}$$

where the expectation is with respect to $(m_k, t_{k,1}, \dots, t_{k,m_k-1}, s_{k,1} \dots, s_{k,m_k-1})$, and where ν^* is the substitution table based on $(x_k, y_k), k = 1, \dots, n^*$.

Note that for $A \neq B$ we have

$$\begin{aligned} &\mathrm{E}_{Q_0^*}\Bigg[\sum_{\ell=1}^{m_k} \mathbf{1}_{\{s_{k,\ell-1}=A, s_{k,\ell}=B\}} \log Q^*(A,B,t_{k,\ell}) \Big| x_k = a, y_k = b \Bigg] \\ &\quad = \int_0^1 \log Q^*(A,B,t) P_0(0,t,a,A) Q_0^*(A,B,t) P_0(t,1,B,t)dt / P_0(0,1,a,b) \\ &\quad = \log Q^1(A,B) Q_0^1(A,B) \frac{\int_0^{1/2} P_0(0,t,a,A) P_0(t,1,B,b)dt}{P_0(0,1,a,b)} \\ &\qquad + \log Q^2(A,B) Q_0^2(A,B) \frac{\int_{1/2}^1 P_0(0,t,a,A) P_0(t,1,B,b)dt}{P_0(0,1,a,b)}, \end{aligned}$$

and similarly we obtain

$$\begin{aligned}
&\mathrm{E}_{Q_0^*}\left[\sum_{\ell=1}^{m_k+1} \mathbf{1}_{\{s_{k,\ell-1}=A\}} \int_{t_{k,\ell-1}}^{t_{k,\ell}} Q^*(A,A,t)dt \middle| x_k = a, y_k = b\right] \\
&\quad = \int_0^1 Q^*(A,A,t)P_0(0,t,a,A)P_0(t,1,A,b)dt/P_0(0,1,a,b) \\
&\quad = Q^1(A,A)\frac{\int_0^{1/2} P_0(0,t,a,A)P_0(t,1,A,b)dt}{P_0(0,1,a,b)} + Q^2(A,A)\frac{\int_{1/2}^{1} P_0(0,t,a,A)P_0(t,1,A,b)dt}{P_0(0,1,a,b)}.
\end{aligned}$$

Therefore (C.2) takes the form

$$\begin{aligned}
G(Q^*;Q_0^*) = &\sum_{A,B:A\neq B} \log Q^1(A,B)w^1(A,B) + \sum_A Q^1(A,A)w^1(A,A) \\
&+ \sum_{A,B:A\neq B} \log Q^2(A,B)w^2(A,B) + \sum_A Q^2(A,A)w^2(A,A), \qquad \text{(C.3)}
\end{aligned}$$

where

$$w^1(A,B) = \sum_{a,b} \nu^*(a,b)Q_0^1(A,B)\frac{\int_0^{1/2} P_0(0,t,a,A)P_0(t,1,B,b)dt}{P_0(0,1,a,b)}, \qquad \text{(C.4)}$$

$$w^1(A,A) = \sum_{a,b} \nu^*(a,b)\frac{\int_0^{1/2} P_0(0,t,a,A)P_0(t,1,B,b)dt}{P_0(0,1,a,b)}, \qquad \text{(C.5)}$$

and with $w^2(A,B)$ and $w^2(A,A)$ defined similarly with the integral being from 1/2 to 1.

We now explain how to calculate the denominator in (C.4) and (C.5). For convenience we drop the subscript. Firstly note that

$$P(0,1,a,b) = \sum_k P(0,1/2,a,k)P(1/2,1,k,b). \qquad \text{(C.6)}$$

Remember that Q^1 is reversible with stationary distribution π. Let V be the real orthogonal matrix with eigenvectors as columns and D_λ the diagonal matrix of eigenvalues of the symmetric matrix $D_\pi^{1/2}Q^1D_\pi^{-1/2}$. It follows that

$$P(0,t) = \exp(Q^1t) = D_\pi^{-1/2}V\exp(tD_\lambda)V^{\mathrm{T}}D_\pi^{1/2}, \quad 0 \le t \le 1/2.$$

Remember that Q^2 is reversible with stationary distribution η. Let U be the real orthogonal matrix with eigenvectors as columns and D_μ the diagonal matrix of eigenvalues of the symmetric matrix $D_\eta^{1/2}Q^2D_\eta^{-1/2}$. It follows that

$$P(1/2,1/2+t) = \exp(Q^2t) = D_\eta^{-1/2}U\exp(tD_\mu)U^{\mathrm{T}}D_\eta^{1/2}, \quad 0 \le t \le 1/2.$$

Now we can find (C.6) and thereby the denominator in (C.5) and (C.4).

Now consider the nominator in (C.4) and (C.5). Note that for $0 \leq t \leq 1/2$ we have

$$
\begin{aligned}
P(t,1,B,b) &= \sum_k P(t,1/2,B,k)P(1/2,1,k,b) \\
&= \Big(\frac{\eta_b}{\pi_B}\Big)^{1/2} \sum_l \sum_m V_{Bl} U_{bm} e^{(\lambda_l+\mu_m)/2} e^{-t\lambda_l} \sum_k \big(\frac{\pi_k}{\eta_k}\big)^{1/2} V_{kl} U_{km},
\end{aligned}
$$

and we get

$$
\begin{aligned}
&\int_0^{1/2} P(0,t,a,A)P(t,1,B,b)dt \\
&= \int_0^{1/2} \Big\{ \sum_j \Big(\frac{\pi_A}{\pi_a}\Big)^{1/2} V_{aj} \exp(t\lambda_j) V_{Aj} \Big\} \\
&\qquad \times \Big\{ \Big(\frac{\eta_b}{\pi_B}\Big)^{1/2} \sum_l \sum_m V_{Bl} U_{bm} e^{(\lambda_l+\mu_m)/2} e^{-t\lambda_l} \sum_k \Big(\frac{\pi_k}{\eta_k}\Big)^{1/2} V_{kl} U_{km} \Big\} dt \\
&= \Big(\frac{\pi_A \eta_b}{\pi_a \pi_B}\Big)^{1/2} \sum_j V_{aj} V_{Aj} \sum_l \sum_m V_{Bl} U_{bm} e^{\mu_m/2} J_{jl} \sum_k \Big(\frac{\pi_k}{\eta_k}\Big)^{1/2} V_{kl} U_{km}, \qquad \text{(C.7)}
\end{aligned}
$$

where

$$
J_{jl} = \begin{cases} \frac{1}{2}\exp(\lambda_l/2) & \text{if } \lambda_j = \lambda_l \\ (\exp(\lambda_j/2) - \exp(\lambda_l/2))/(\lambda_j - \lambda_l) & \text{if } \lambda_j \neq \lambda_l. \end{cases}
$$

Fast implementation of (C.7) is achieved by first calculating

$$
Y_{bk} = \sum_m U_{bm} e^{1\mu_m/2} U_{km}, \quad Z_{bl} = \sum_k Y_{bk} \big(\frac{\pi_k}{\eta_k}\big)^{1/2} V_{kl}.
$$

Then

$$
\int_0^{1/2} P(0,t,a,A)P(t,1,B,b)dt = \Big(\frac{\pi_A \eta_b}{\pi_a \pi_B}\Big)^{1/2} \sum_n V_{an} V_{An} \sum_l V_{Bl} Z_{bl} J_{nl}.
$$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd international symposium on information theory* (eds. B. Petrov and F. Csaki), Akademiai Kiado, 267-281, Budapest.

Albert, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002). *Molecular biology of the cell*. Garland Science, New York.

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

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736.

Holmes, I. and Rubin, M. G. (2002). An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.* **317**, 753–764.

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

Jensen, J. L. (2005). Time discretized DNA evolutionary models. In preparation.

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

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

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

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

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. *Mol. Biol. Evol.* **20**, 1692–1704.

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

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

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