research
reports

No. 458 May 2005
2005 / 05 / 18

Jens Ledet Jensen

Context dependent DNA evolutionary models

department of
theoretical
statistics

university of
aarhus

ISSN 0902-9893

Context dependent DNA evolutionary models

Jens Ledet Jensen
Department of Mathematical Sciences, University of Aarhus
Ny Munkegade, DK-8000 Aarhus C, Denmark

Contents

1 Introduction

This paper is about stochastic models for the evolution of DNA. For a set of aligned DNA sequences, connected in a phylogenetic tree, the models should be able to explain - in probabilistic terms - the differences seen in the sequences. From the estimates of the parameters in the model one can start to make biologically interpretations and conclusions concerning the evolutionary forces at work.

In parallel with the increase in computing power, models have become more complex. Starting with Markov processes on a space with 4 states, and extended to Markov processes with 64 states, we are today studying models on spaces with 4^n (or 64^n) number of states with n well above one hundred, say. For such models it is no longer possible to calculate the transition probability analytically, and often Markov chain Monte Carlo is used in connection with likelihood analysis. This is also the approach taken in this paper, and a time discretization of the process is presented in order to make the calculations more feasible. Apart from the time discretization we introduce a set of simple estimating equations, together with an EM type algorithm, for finding the parameter estimates. A detailed derivation of the asymptotic properties of the estimates is also given.

Before describing in more detail the content of the paper we very briefly explain the structure of a DNA sequence.

1.1 DNA sequence

The hereditary information in an organism is carried by DNA (deoxyribonucleic acid) molecules. Such a molecule has two complementary chains bound together in a helix. Each chain is a string of four *nucleotides*: A, G, C, and T. The names of these are adenine, guanine, cytosine, and thymine. The four nucleotides are grouped into two purines: A and G, and two pyrimidines: C and T. In the two complementary chains of the DNA molecule, A always forms a pair with T, and G forms a pair with C. The bond between G and C is stronger than the bond between A and T. A precise description of a nucleotide will not be given here, it suffices for us to know that the DNA molecule is a string of letters from a four letter alphabet.

To obtain a protein, part of the DNA molecule is transcribed into mRNA (messenger RNA). The part that is transcribed need not be a noninterrupted part of the DNA. Instead, there are blocks, known as exons and introns, that go into the mRNA or are left out, respectively. The mRNA is next translated into a sequence of amino acids. This involves a reading frame whereby the nucleotides are put together three by three, called codons, and each codon is translated into an amino acid. Translation stops when a stop codon is encountered. There are three stop codons: TAA, TAG, and TGA. There are only 20 amino acids, so that some of the 61 nonstop codons encode the same amino acid. Two codons that give the same amino acid are called synonymous and nonsynonymous otherwise. To see the code, that relates amino acids to codons, write *genetic code* in Google.

It is possible to have more than one reading frame so that a mRNA molecule translates into two or three proteins. Also, one has parts of the DNA molecule that are transcribed, but not translated into proteins.

The stochastic models in this paper are targeted towards the analysis of a short stretch of DNA (typical corresponding to a gene) from two or more species. Before using the models the sequences are aligned. Mathematically, an alignment consists in placing the sequences in an array, where each entry is either a nucleotide or a "gap". A row corresponds to one of the sequences, and a column contains a set of nucleotides that have all developed along the phylogenetic tree through mutations from the same ancestral nucleotide at the root of the tree. Gaps in a column imply either the insertion or the deletion of a nucleotide during evolution. For the models in this paper we discard the columns that contain a gap and, thus, we consider only the point mutations where a single nucleotide is replaced by another nucleotide.

The first stochastic models in this field assumed that the nucleotides along the DNA sequence evolved independently of one another. For a protein coding part of a DNA sequence these models were supplemented by models with independent codons, but where nucleotides within a codon were dependent. We give a short review of the independent sites models in Section 2 below. The main emphasis in this paper is on models with dependence between codons. Attention is restricted to short range dependence, where the evolution at one site depends on the two neighbouring sites. The "contect dependent" of the title is in this paper a synonym for "neighbour dependent".

1.2 Overview of paper

As mentioned above we start in Section 2 with a review of the classical independent sites models before turning to a review of some of the recent papers dealing with neighbour dependence. Papers that introduce a new model or a new estimation procedure has been included, although the list is not complete. It is often discussed whether or not to use time reversible models in an evolutionary context. Following the review section a characterization of time reversibility is given, with special emphasis on the relation to the properties of the stationary measure along the DNA sequence. A discussion of time reversibility in relation to one of the papers being reviewed is given in Appendix C. Turning to the main subject of this paper, a description of the continuous time model of Jensen and Pedersen (2000) for a coding

sequence is given in Section 4. A time discretized version of this model is introduced in the following section. A new estimation procedure is given based on an analogue of the EM algorithm, and where conditional mean values are calculated via a Markov chain Monte Carlo simulation. The Gibbs update within the MCMC is described in detail. In Section 6 we specialize to a specific model. The discussion in Sections 4 and 5 is based on a modelling of two aligned sequences. In Section 7 the general case of multiple sequences connected in a known phylogeny is given. The concluding Section 8 is fairly theoretical with proofs of asymptotic properties of the estimates obtained from the MCMC simulation. The asymptotic considerations can be compared with the asymptotics of maximum likelihood estimates in hidden Markov models, a subject where new results have appeared within the last few years.

2 Review

2.1 Continuous time Markov processes

Stochastic models for the evolution of DNA are usually continuous time Markov processes defined through their infinitesimal rates. Let $z(t)$ be a homogeneous continuous time irreducible Markov process with a finite state space. The (infinitesimal) rate q_{ij} for a change from i to j is defined as $q_{ij} = \lim_{t\to 0} P(z(t) = j|z(0) = i)/t$. The matrix Q with entries q_{ij} off the diagonal, and with $q_{ii} = -\sum_{j\neq i} q_{ij}$, is called the rate matrix. Let $P(t) = \{p_{ij}(t)\}$ be the transition matrix, that is, $p_{ij}(t) = P(z(t) = j|z(0) = i)$. It is a standard result (Karlin and Taylor (1975)) that $P(t)$ is given by

$$P(t) = \exp(tQ) = \sum_{n=0}^{\infty} \frac{t^n Q^n}{n!}.$$

The matrix Q is called diagonalizable if the set of eigenvectors span the whole space. In this case one can write $Q = SDS^{-1}$, where D is the diagonal matrix of eigenvalues, some of which can be complex and some of which can be identical, and S is the matrix of columnwise eigenvectors. It is easy to see that this leads to the important formula

$$P(t) = S\exp(tD)S^{-1}, \tag{1}$$

where $\exp(tD)$ is a diagonal matrix with entries $\exp(td_{ii})$. In Appendix A a simple formula for $\frac{\partial P(t)}{\partial \theta}$ is given for the case where the rates q_{ij} are functions of a parameter θ.

2.2 Review of independent sites models

In this section we describe models where the sites (either nucleotides or codons) along the DNA sequence evolve independently of one another. The models, therefore, reduce to models for the evolution of a single site.

We first consider models at the nucleotide level. The *HKY* model (Hasegawa et al. (1985)) has a parameter α for a *transition* (a change within $\{A, G\}$ or within $\{C, T\}$), a parameter β for a *transversion* (a change from one of the groups $\{A, G\}$

and $\{C, T\}$ to the other), and allows for a general stationary distribution $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$ of the Markov process. The rate matrix is

	A	G	C	T
A	$-(\alpha\pi_G + \beta\pi_{CT})$	$\alpha\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha\pi_A$	$-(\alpha\pi_A + \beta\pi_{CT})$	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	$-(\alpha\pi_T + \beta\pi_{AG})$	$\alpha\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha\pi_C$	$-(\alpha\pi_C + \beta\pi_{AG})$

(2)

where $\pi_{AG} = \pi_A + \pi_G$ and $\pi_{CT} = \pi_C + \pi_T$. That π indeed is the stationary distribution is shown in Section 3. This model contains two special cases. If $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ Kimura's two parameter model (Kimura (1980)) is obtained. If, furthermore, $\alpha = \beta$ the model of Jukes and Cantor (1969) appears. The HKY model is *reversible* as described in Section 3. From that section one can also see the most general reversible rate matrix. The eigenvalues and eigenvectors of the rate matrix (2) are given in Appendix B.

The next model to be considered is the general *strand symmetric* model. In the double stranded DNA molecule nucleotide A sits opposite to T and nucleotide G sits opposite to C. In the strand symmetric model a rate q_{ab} equals the rate where a and b are replaced by the nucleotides that they form a pair with. The general rate matrix is

	A	G	C	T
A	$-(\alpha + \beta + \delta)$	α	β	δ
G	γ	$-(\gamma + \omega + \kappa)$	κ	ω
C	ω	κ	$-(\gamma + \omega + \kappa)$	γ
T	δ	β	α	$-(\alpha + \beta + \delta)$

(3)

The stationary distribution is $\pi = (\theta, 1-\theta, 1-\theta, \theta)/2$ with $\theta = (\gamma+\omega)/(\gamma+\omega+\alpha+\beta)$. The eigenvalues and eigenvectors of the rate matrix are given in Appendix B. If one requires strand symmetry as well as reversibility, γ and ω must satisfy

$$\gamma = \alpha\frac{\theta}{1-\theta} \quad \text{and} \quad \omega = \beta\frac{\theta}{1-\theta},$$

where $0 < \theta < 1$ is now a free parameter.

Turning to models for independent codons the situation is much more complex. On one hand we should keep some of the modelling ideas from the nucleotide models and, on the other hand, we should take into considerations the translation from the codon to an amino acid. A change of the codon is called a *synonymous* change if it does not change the amino acid and a *nonsynonymous* change if the amino acid is being changed. It is generally believed that nonsynonymous changes occur relatively more rarely than synonymous changes, since the former may change the function of the protein. A change that produces a stop codon has zero probability, as this will destroy the protein, and the model, therefore, effectively has 61 states. Also, the only changes allowed are those where a single nucleotide is changed at each time point.

Muse and Gaut (1994) seem to be the first to make a model at the codon level that differentiates between synonymous and nonsynonymous changes. The rate $\lambda_j(z|x)$, for a change of nucleotide x^j within the codon $x = (x^1, x^2, x^3)$, is

$$\lambda_j(z|x) = \mu\beta^{1_{NS}}\pi_z,$$

where π is a set of "equilibrium frequencies" for the nucleotides, and 1_{NS} is one if the change is nonsynonymous and zero if the change is synonymous. Had there been no stop codons, the stationary distribution for this Markov process would be $\pi_{x^1}\pi_{x^2}\pi_{x^3}$, and the stationary probability for a nucleotide z would be π_z. When there are stop codons the probabilities $\pi_{x^1}\pi_{x^2}\pi_{x^3}$ must be normalized to sum to one over the 61 nonstop codons, and the stationary probability of a nucleotide z is only approximatively equal to π_z.

Goldman and Yang (1994) consider a more complex model incorporating a transition/transversion parameter as well as a differentiation between the different nonsynonymous changes. Letting $x(j,z)$ be the codon where x^j is replaced by z, and letting π_x be frequencies summing to one over the set of nonstop codons, the rates are

$$\lambda_j(z|x) = \mu\kappa^{1_{tv}} \exp\bigl(-d_{x,x(j,z)} \,/\, V\bigr)\pi_{x(j,z)}.$$

Here 1_{tv} is one for a transversion and zero for a transition, and $d_{x,y}$ is a distance between the amino acids encoded by the codons x and y. Goldman and Yang (1994) use distances between the amino acids given by Grantham (1974) and based on molecular properties. From the theory in Section 3 it follows easily that this model is reversible and that the stationary probabilities are proportional to π_x.

The models by Muse and Gaut (1994) and by Goldman and Yang (1994) represent two extremes, the former having only 5 parameters and the latter having 63 parameters. Both Pedersen et al. (1998) and Schadt and Lange (2002) consider models in between. Pedersen et al. (1998) consider the rates

$$\lambda_j(z|x) = \mu\kappa^{1_{tv}}\beta^{1_{NS}}\gamma^{1_{CG}(x)-1_{CG}(x(j,z))}\pi_z^j,$$

where $(\pi_A^j, \pi_G^j, \pi_C^j, \pi_T^j)$ is a set of nucleotide frequencies for each $j = 1, 2, 3$, and $1_{CG}(x)$ is one if the codon x contains a nucleotide C followed by a nucleotide G and is zero otherwise. This model has 12 parameters. Schadt and Lange (2002) consider rates of the form

$$\lambda_j(z|x) = \mu_1^{1_{ts}^{AG}}\mu_2^{1_{ts}^{CT}}\kappa^{1_{tv}}k(x, x(j,z); \rho)\pi_z,$$

where 1_{ts}^{AG} is one for a transition within $\{A, G\}$ and zero otherwise, and 1_{ts}^{CT} is one for a transition within $\{C, T\}$ and zero otherwise. The function $k(x, x(j,z); \rho)$ depends on the amino acids only. The 20 amino acids are divided into 4 groups, and k can take on 4 different values $\rho_0, \ldots, \rho_3$ depending on the change being within a group or between the groups. The model has 10 parameters. Using again the theory in Section 3 it is seen that this model is reversible.

The independent sites models are often extended by allowing *rate variation* along the sequence. In these models the rate matrix is multiplied by an individual rate factor at each site. Often the factor is gamma distributed or can take on a small number of values only. In the simplest case the site independence is kept by having

the rate factors being independent. Alternative models let the rate factors constitute a Markov chain along the sequence, leading to an analysis as for hidden Markov models. Rate variation is not included in the models presented in this paper, although the feature can easily be incorporated in the MCMC analysis performed.

2.3 Review of context dependent models

Let $x = (x_1, x_2, \ldots, x_n)$ be a DNA sequence, where x_i is either a single nucleotide or a single codon, and let $x(t)$ be the process at time t. In this section we review papers where the rate for a change of x_i depends on the two flanking values x_{i-1} and x_{i+1}. Such models are called *context dependent* models.

Jensen and Pedersen (2000); Pedersen and Jensen (2001)

These two papers form the background for the present paper. A context dependent model at the codon level is presented including the *CG-depression* effect. The latter refers to the observation that in some parts of the genome one sees less Cs followed by a G than the nucleotide frequencies would suggest. A discussion of the relation between reversibility and the Markov property of the stationary measure is given. A Markov chain Monte Carlo method is suggested for evaluating likelihood ratios in the case of two sequences. This is a fairly slow procedure making it less feasible for multiple sequences. The approach suggested in this paper makes the model useful for multiple sequences also.

Arndt et al. (2003) [ABH]

In this paper the authors consider a context dependent model at the nucleotide level suitable for the description of the noncoding parts of the genome. An approximation to the stationary distribution is derived, and this is the only part of the model used in the data analysis.

If $\lambda(y_i|x_{i-1}, x_i, x_{i+1})$ is the rate for a change of x_i to y_i, when the two neighbouring nucleotides are x_{i-1} and x_{i+1}, the context dependent model in [ABH] is of the form

$$\lambda(y_i|x_{i-1}, x_i, x_{i+1}) = \lambda_0(y_i|x_i) + \lambda_l(y_i|x_{i-1}, x_i) + \lambda_r(y_i|x_i, x_{i+1}). \tag{4}$$

Here λ_0 is a rate not depending on the context, λ_l is a rate depending on the left neighbour, and λ_r is a rate depending on the right neighbour. Actually, the possibility of a simultaneous change of both nucleotides that are neighbours is also allowed, but this model is not used in the data analysis. Imagine now that the model is for a double infinitely long sequence and that we want to find the stationary probabilities $\pi_{ab} = P(x_1 = a, x_2 = b)$. [ABH] use the Kolmogorov forward differential equations to establish a set of equations for π_{ab}. These equations involve the triplet probabilities $\pi_{abc} = P(x_1 = a, x_2 = b, x_3 = c)$, and [ABH] introduce the approximation

$$\pi_{abc} \approx \frac{\pi_{ab}\pi_{bc}}{\pi_b} \tag{5}$$

in order to solve the equations. The differential equations are explained in Appendix C below.

The authors do not discuss which parts of the rates (4) that are identifiable from the stationary measure, nor do they delineate the cases for which the Markov

approximation (5) is exact. From Proposition 4 in Section 3 below, one sees that the stationary measure is a Markov chain if the rates are time reversible. However, the additive structure in the rates (4) does not fit well with the multiplicative structure in the reversible rates as given in Proposition 4. In Appendix C of this paper we give a complete characterization for the simplified case of a two letter alphabet.

Arndt et al. (2003)

These authors use the model (4) from [ABH] with four parameters in λ_0, one nonzero term in λ_l ($CG \to CA$), and one nonzero term in λ_r ($CG \to TG$). A star phylogeny is considered with the ancestor known. Instead of calculating the true likelihood under the model, a "pseudo likelihood" is used. The latter involves two approximations. The likelihood is approximated by a product of marginal likelihoods of the form $P(x_i(T)|x_{i-1}(0), x_i(0), x_{i+1}(0))$, and to calculate the latter another approximation is needed. This last step is not spelled out in full detail, but presumably an approximate model is used where the λ_l term is left out for position $i-1$ and the λ_r term is left out for position $i+1$. The authors write that the probability is calculated by iterating 64 differential equations. Alternatively, a 64×64 transition matrix $\exp(Qt)$ can be calculated using an eigenvalue decomposition.

Hwang and Green (2004)

Hwang and Green (2004) consider a general nonreversible context dependent nucleotide model. Time is dicretized so that the average substitution rate in each time step is less than 0.005. In each time step the nucleotides evolve independently given the present sequence (this is in contrast to the model presented in this paper where multiple reading frames enters). The distribution of the root sequence is modelled by a second order Markov chain. The context dependent rates are modelled completely freely with a total of 192 parameters. In one of the models considered, the tree is divided into 12 parts so that 12×192 parameters are used. A Bayesian MCMC approach is used to obtain samples from the posterior distribution of the parameters. In the MCMC updating step either a single parameter is updated or the path of a single nucleotide is updated. A dataset with 19 mammalian species and spanning approximately 1.7 Mb is used. With the large number of parameters one should probably be cautious in the interpretations of the results. As an example the authors point to a difference in the substitution rates for different groups (clades) of species, but it is unclear if this difference can perhaps be caused by the use of a clade specific normalization of the rates.

Siepel and Haussler (2004) [SH]

These authors consider N-tuples of nucleotides, where N is either 1, 2, or 3. The marginal distribution of the process of an N-tuple over the phylogenetic tree is modelled by a homogenous continuous time Markov process, where the rates allow single nucleotide changes only. For each choice of N four models for the rate matrix are considered: an unrestricted model, a reversible model, a strand symmetric model, and a strand symmetric reversible model. The parameters are estimated using a pseudo likelihood consisting of the product of the likelihoods from the marginal distribution of nonoverlapping N-tuples. The actual maximization is done using an EM algorithm where the full likelihood is based on the process at all the nodes of the tree.

[SH] do not discuss the quality of their model when viewed as an approximation to a full context dependent model. Thus, a model formulated in terms of rates, that depend on the neighbouring nucleotides, presumably do not lead to a homogenous Markov process for an N-tuple of nucleotides. The quality of this approximation is, therefore, of interest, as well as a translation of the rates from the full context dependent model to the rates in the approximating marginal model of an N-tuple. Unfortunately, there is no comparison in the paper with the data analysis of Hwang and Green (2001), where the full context dependent model is used.

[SH] also consider the use of the N-tuple model to define a Markov model. Thus, from the marginal model for an N-tuple over the phylogenetic tree a transition matrix is obtained as the conditional distribution of the Nth term given the $N-1$ first terms. The estimation within this model becomes complicated, and the authors use the parameters obtained from the marginal N-tuple model. It is noted that the log likelihood for the data becomes much larger for the Markov model as compared to the model with independent N-tuples. A simple entropy inequality shows that this increase is to be expected if the Markov model is a better fit to the data. Actually, it would be interesting to see if the increase in the log likelihood is as one would expect if the Markov model is the correct description of the data.

Christensen et al. (2004)

In this paper a codon model is considered, mainly for the analysis of two species. The context dependency is through a CG depression across codon boundaries. A pseudo likelihood is used, where the contribution from the ith codon is calculated as though the evolutionary history of the two flanking nucleotides is known. The true evolutionary history for a flanking nucleotide is approximated by either a history with no changes if the nucleotides in the two sequences are identical, and a history with one change in the middle of the time interval if the nucleotides in the two sequences are different. A comparison with the full analysis described in this paper shows that the estimates obtained from the pseudo likelihood are very close to the maximum likelihood estimates.

3 Time reversibility

In this section we discuss time reversibility of different DNA evolutionary models, ending up with a discussion of the relationship between time reversibility and the Markov property of the stationary measure for context dependent models. The setup is that of a finite state irreducible continuous time Markov process. The rate matrix is Λ with entries λ_{ij}. The transition probability $p_{ij}(t) = P(x(t) = j|x(0) = i)$ is given by the (i, j)th entry of $\exp(t\Lambda)$, and the stationary probabilities $\{\pi_i\}$ are all positive. Considering the process in reverse time the transition probabilities are given by

$$q_{ij}(t) = P(x(0) = j|x(t) = i) = \frac{P(x(0) = j, x(t) = i)}{P(x(t) = i)} = \frac{\pi_j p_{ji}(t)}{\pi_i}.$$

Time reversibility means that the transition probability in reverse time equals the transition probability in forward time. From the above formula for $q_{ij}(t)$ the time

reversibility requirement can be written as $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$ for all i, j. Letting $t \to 0$ one finds that time reversibility implies the *detailed balance condition* $\pi_i \lambda_{ij} = \pi_j \lambda_{ji}$, for all $i \neq j$.

Lemma 1. *The Markov process is time reversible if and only if there exists a symmetric function $h(i, j)$ and a positive function $g(i)$, such that the rates are given by $\lambda_{ij} = g(i)h(i, j)$ for all i, j. In the latter case the stationary probabilities are proportional to $g(i)^{-1}$.*

Proof. If $\lambda_{ij} = g(i)h(i, j)$ for all i, j, we define $\pi_i = g(i)^{-1} \left(\sum_j g(j) \right)$. The symmetry of h implies detailed balance, $\pi_i \lambda_{ij} = \pi_j \lambda_{ji}$, and this in turn gives that π is the stationary distribution: $\sum_i \pi_i \lambda_{ij} = \sum_i \pi_j \lambda_{ji} = 0$. Detailed balance also gives

$$\pi_i p_{ij}(t) = \sum_n \pi_i \frac{t^n (\Lambda^n)_{ij}}{n!} = \sum_n \pi_j \frac{t^n (\Lambda^n)_{ji}}{n!} = \pi_j p_{ji}(t),$$

so that the process is time reversible.

If, on the other hand, the process is time reversible, so that $\pi_i \lambda_{ij} = \pi_j \lambda_{ji}$ for all i, j, we define $h(i, j) = \pi_i \lambda_{ij}$ and $g(i) = 1/\pi_i$. Then, clearly, h is symmetric and $\lambda_{ij} = g(i)h(i, j)$. □

A general discussion of reversibility, including the above lemma, can be found in Kelly (1979). If the Markov process has k states, the *general reversible process* has $\frac{1}{2}k(k-1) + (k-1)$ free parameters. One possibility is to choose the parameters as $\{\lambda_{i,j}, j > i\}$ together with a set of positive values $\{\theta_j, j = 2, \ldots, k\}$. The lower triangular part of the rate matrix is then given by $\lambda_{ji} = \frac{\theta_i}{\theta_j} \lambda_{ij}$, $j > i$, where $\theta_1 = 1$. In this case the stationary probabilities are $\pi_i = c\theta_i$, where c is a normalizing constant.

Example 2. We consider the HKY model with rate matrix (2). To prove reversibility define $g(i) = 1/\pi_i$, and let $h(i, j)$ be $\pi_i \pi_j$ times α for a transition and $\pi_i \pi_j$ times β for a transversion. The rates are then given by $g(i)h(i, h)$, and since h is clearly symmetric, the process is reversible according to Lemma 1. The stationary distribution is $g(i)^{-1} = \pi_i$ which enters directly in the rates (2).

Next, we consider the general strand symmetric matrix (3). To see that this is not a time reversible rate matrix look at the cycle $A \to G \to C \to A$. Had the process been time reversible, the equality

$$(\pi_A \alpha)(\pi_G \epsilon)(\pi_C \gamma) = (\pi_G \phi)(\pi_C \epsilon)(\pi_A \beta)$$

holds. This reduces to $\alpha\gamma = \phi\beta$, which is clearly not satisfied in general, and the process is not time reversible. Using a cycle to show the nonreversibility is known as Kolmogorovs condition (Kolmogorov (1936); Kelly (1979)).

Example 3. Let the state space be the set of all nonstop codons. Thus, an element of this space is of the form $x = (x^1, x^2, x^3)$ with $x^i \in \{A, G, C, T\}$, and $x \notin \{TAA, TAG, TGA\}$. All models to be considered allow for a change of one nucleotide only, at each time point. The rate for a substitution of x^j by z is denoted $\lambda_j(z; x)$. We first describe a general class of reversible models. Let $x(j, z)$ be the

codon x with x^j replaced by z, let $1_{CG}(x)$ be one if x contains a C followed by G, and let $(\pi_A^j, \pi_G^j, \pi_C^j, \pi_T^j)$ be a set of frequencies for each $j = 1, 2, 3$. We consider rates of the form

$$\lambda_j(z|x) = v_j(x^j, z)\pi_z^j \gamma_1^{1_{CG}(x(j,z))} \gamma_2^{1_{CG}(x)} w_{\text{am}}(x, x(j, z)),$$

where the function v_j is symmetric, and the function $w_{\text{am}}(x, x(j, z))$ depends on the amino acids only, and is symmetric in the two amino acids encoded by x and $x(j, z)$.

To show reversibility of this model, define $g(x) = \gamma_2^{1_{CG}(x)} \ / \ \left[\gamma_1^{1_{CG}(x)} \pi_{x^1}^1 \pi_{x^2}^2 \pi_{x^3}^3\right]$ and define the function $h(x, y)$ to be zero if the two nonstop codons x and y differ at more than one nucleotide position and define h to be

$$h(x, y) = v_j(x^j, y^j) w_{\text{am}}(x, y) \gamma_1^{1_{CG}(x)+1_{CG}(y)} \pi_{x^1}^1 \pi_{x^2}^2 \pi_{x^3}^3 \pi_{y^j}^j$$

when x and y differ at position j only. Clearly $h(x, y)$ is symmetric in x and y, and the rates are given by $g(x)h(x, y)$. According to Lemma 1 the model is reversible and the stationary measure is

$$\pi(x) = \frac{\gamma_1^{1_{CG}(x)} \pi_{x^1}^1 \pi_{x^2}^2 \pi_{x^3}^3}{C \gamma_2^{1_{CG}(x)}},$$

where C is a norming constant.

To illustrate nonreversibility by a simple example we consider the rates

$$\lambda_j(z; x) = \begin{cases} \gamma^{1_{AA}(x^1, x^2)} & j = 1, \\ \gamma^{1_{AA}(x^2, x^3)} & j = 3, \\ \gamma^{1_{AAA}(x^1, x^2, x^3)} & j = 2, \end{cases}$$

and look at the cycle $AAG \to GAG \to GGG \to AGG \to AAG$. Had the process been reversible, the following should be true

$$(\pi_{AAG}\gamma)(\pi_{GAG}1)(\pi_{GGG}1)(\pi_{AGG}1) = (\pi_{GAG}1)(\pi_{GGG}1)(\pi_{AGG}1)(\pi_{AAG}1).$$

This reduces to $\gamma = 1$ and, so, the process is not reversible.

We next consider a process with state space $\Omega = \{x = (x_1, \ldots, x_n) : x_i \in S_i^1, i = 1, \ldots, n, (x_{i-1}, x_i) \in S_i^2, i = 1, \ldots, n+1\}$, where S_i^1 is a finite set. The set S_i^2 allows for the possibility of neighbour restrictions. No restrictions correspond to taking $S_i^2 = S_{i-1}^1 \times S_i^1$. For a DNA string, where each $x_i = (x_i^1, x_i^2, x_i^3)$, S_i^1 is usually taken to be the set of nonstop codons. If the sequence contains a double reading frame, a typical neighbour restriction is $S_i^2 = \{(x_{i-1}, x_i) | (x_{i-1}^2, x_{i-1}^3, x_i^1) \in S^1\}$. We consider models where the only changes $x \to y$ that are allowed are those where $y_i \neq x_i$ for some i and $y_j = x_j$ for $j \neq i$. The rate for a change of x_i depends on x through x_i and its two neighbours x_{i-1} and x_{i+1} only. Such models are called *context dependent models* as opposed to independent sites models where the rate depends on x_i only. The rate for a change of x_i to y_i is written as

$$\lambda_i(y_i | x_{i-1}, x_i, x_{i+1}). \tag{6}$$

For $i = 1$ and $i = n$ the rates are defined through given fixed values of x_0 and x_{n+1}. We always assume that the rates are such that the Markov process is irreducible. The stationary distribution is denoted φ and when the process is irreducible the stationary probabilities $\varphi(x)$ are strictly positive for all $x \in \Omega$. We say that the process satisfies the *neighbour support condition* if there exists $z^* \in \Omega$ such that

$$(a, z_i^*) \in S_i^2 \text{ and } (z_i^*, b) \in S_{i+1}^2 \quad \forall\, i,\ \forall\, a \in S_{i-1}^1,\ \forall\, b \in S_{i+1}^1. \tag{7}$$

Proposition 4. *Assume that the neighbour support condition is satisfied. Then, the context dependent model is time reversible if and only if the rates λ_i can be written as*

$$\lambda_i(y_i|x_{i-1}, x_i, x_{i+1}) = g_i(x_i; x_{i-1}, x_{i+1}) h_i(x_i, y_i; x_{i-1}, x_{i+1}), \tag{8}$$

where h_i is symmetric in its first two arguments and where g_i can be written as

$$g_i(x_i; x_{i-1}, x_{i+1}) = q_i(x_i, x_{i-1}) q_{i+1}(x_{i+1}, x_i)$$

for some positive functions q_i, $i = 1, \ldots, n+1$. In the latter case the stationary probability $\varphi(x)$ is proportional to $\prod_{i=1}^{n+1} q_i(x_i, x_{i-1})^{-1}$.

Proof. If the rates are on the form given in the proposition, we define

$$g(x) = \prod_{i=1}^{n+1} q_i(x_i, x_{i-1})$$

and

$$h(x, y) = \begin{cases} \left(\prod_{j=1}^{i-1} q_j(x_j, x_{j-1}) \prod_{j=i+2}^{n+1} q_j(x_j, x_{j-1})\right)^{-1} h_i(x_i, y_i; x_{i-1}, x_{i+1}) \\ \qquad\qquad \text{when } y_i \neq x_i,\ y_j = x_j, j \neq i, \\ 0 \qquad\qquad \text{otherwise.} \end{cases}$$

Then, h is symmetric in (x, y), and the rates (8) are on the form $g(x)h(x, y)$. From Lemma 1 this implies time reversibility, and also gives the form of the stationary distribution.

Assume now that the process is time reversible and let $\varphi(x)$ be the stationary distribution. The reversibility implies that

$$\begin{aligned} &\varphi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \lambda_i(y_i|x_{i-1}, x_i, x_{i+1}) \\ &\quad = \varphi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n) \lambda_i(x_i|x_{i-1}, y_i, x_{i+1}). \end{aligned} \tag{9}$$

Looking at this equation for fixed i and fixed $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, one sees that the rates (6) for fixed i and fixed (x_{i-1}, x_{i+1}) are reversible. Using Lemma 1 we, therefore, obtain that the rates are of the form given in (8) for a general function g_i. Using this, (9) is rewritten as

$$\frac{\varphi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{\varphi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)} = \frac{g_i(y_i; x_{i-1}, x_{i+1})}{g_i(x_i; x_{i-1}, x_{i+1})}. \tag{10}$$

This equation shows that in the stationary distribution the conditional distribution of x_i, given all the other variables, depends on (x_{i-1}, x_{i+1}) only. Below it is argued

that the Hammersley Clifford Theorem is valid in our case. When using this the above Markov property (10) implies that φ can be written as

$$\varphi(x) = \prod_{i=1}^{n+1} \varphi_i(x_i, x_{i-1})$$

for some functions φ_i, $i = 1, \ldots, n+1$. Inserting this back into (10) we obtain the form of g_i specified in the proposition.

The original version of the Hammersley Clifford Theorem (see Besag (1974)) assumed a positivity condition, which corresponds to having the state space Ω being a product space $\Omega = \prod_i S_i^1$. To handle the neighbour restrictions that are present in our general model, we need the version of the Hammersley Clifford Theorem given in Kaiser and Cressie (2000). There the state space can be more general under an assumption called the *MRF support condition*. The latter requires the existence of $z^* \in \Omega$ such that for any i and any point $(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, belonging to the marginal support of $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ under the stationary measure, it holds that

$$(z_1, \ldots, z_{i-1}, z_i^*, z_{i+1}, \ldots, z_n) \in \Omega. \tag{11}$$

In our case the marginal support of $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ cannot be larger than

$$\Omega_{-i} = \{z = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n) : \ z_j \in S_j^1, \ j = 1, \ldots, n, \ j \neq i, \\ (z_{j-1}, z_j) \in S_j^2, \ j = 1, \ldots, n+1, \ j \notin \{i, i+1\}\}.$$

From the neighbour support condition (7) we, therefore, see that (11) is satisfied. The neighbour support condition actually also implies that the marginal support is equal to Ω_{-i}. □

In a phylogenetic context the mathematical importance of time reversibility (and stationarity) is that the likelihood of a set of sequences connected through a tree can be calculated with the root of the tree positioned anywhere on the tree. The simple basic step in this argument runs as follows. Consider a root with sequence x and two branches descending from the root of lengths t_1 and t_2 and with the sequences y and z at the end of the branches. The likelihood of (y, z) is then

$$\begin{aligned} \sum_x \phi(x) p(y|x; t_1) p(z|x; t_2) &= \sum_x \phi(y) p(x|y; t_1) p(z|x; t_2) \\ &= \phi(y) p(z|y; t_1 + t_2), \end{aligned}$$

where $p(y|x; t) = P(x(t) = y | x(0) = x)$ is the transition probability. Using this argument iteratively the root can be moved to any position on the tree.

4 Continuous time model

In this section we describe the class of context dependent codon models introduced in Jensen and Pedersen (2000) and Pedersen and Jensen (2001). These models are

defined in continuous time and this section sets the stage for the time discretization of the next section.

Let $x(t)$ be a codon sequence of length n at time t, $x(t) = (x_1(t), \ldots, x_n(t))$. The jth nucleotide of codon i is $x_i^j(t)$, $j = 1, 2, 3$. The models studied in the above two mentioned papers are defined through rates λ of the form

$$\lambda\big[j, \nu\big|x_i^{(j-2):(j+2)}\big], \tag{12}$$

where $x_i^{(j-2):(j+2)} = (x_i^{j-2}, x_i^{j-1}, x_i^j, x_i^{j+1}, x_i^{j+2},)$ and $x_i^{-1} = x_{i-1}^2$, $x_i^0 = x_{i-1}^3$, $x_i^4 = x_{i+1}^1$, $x_i^5 = x_{i+1}^2$. Here λ is the rate for a replacement of the nucleotide x_i^j by ν. The dependency on the two flanking nucleotides reflects the possibility of more than one reading frame as well as the possibility of dinucleotide interactions. The full likelihood, conditionally on the initial sequence $x(0)$, from observing $x(t)$ from time zero to time one is a product

$$L(\cdot|x(0)) = \prod_{i=1}^{n}\prod_{j=1}^{3} L_i^j, \tag{13}$$

where each term is the contribution from the events at the corresponding site. In this likelihood enters x_0 and x_{n+1}, and here these are taken to be known and fixed. Alternatively, one can consider the conditional likelihood given x_0 and x_{n+1}, in which case (13) includes terms with $i = 0$ and $i = n + 1$ as well as a norming constant. The total rate for a change of nucleotide x_i^j is

$$\lambda\big[j\big|x_i^{(j-2):(j+2)}\big] = \sum_{\nu \neq x_i^j} \lambda\big[j, \nu\big|x_i^{(j-2):(j+2)}\big],$$

and using this the individual terms L_i^j can be written as

$$L_i^j = \exp\Big\{-\int_0^1 \lambda\big[j\big|x_i^{(j-2):(j+2)}(s)\big]\,ds\Big\} \prod_{m=1}^{K} \lambda\big[j, x_i^j(t_m)\big|x_i^{(j-2):(j+2)}(t_m-)\big], \tag{14}$$

where K is the number of jumps for nucleotide j in codon i, and t_m is the mth jump time.

We now consider a class of models for the rates (12), for which the estimation of the parameters based on the full likelihood (13) becomes simple. The substitutions are divided into a number of different *types*. Possible types of jumps are *synonymous* versus *nonsynonymous* and *transitions* versus *transversions*. Each type r has a parameter θ_r attached to it. The rates (12) can be written as $\theta_R\gamma\big[j, \nu\big|x_i^{(j-2):(j+2)}\big]$, where R is the type of jump depending on all the arguments of γ. The term γ is a product of codon or nucleotide frequencies and a function related to dinucleotide interactions. The stationary distribution of the sequence depends on the γ part of the rates only, and we choose to estimate the parameters of γ from the stationary distribution of one of the observed sequences, $x(0)$ say. For an explicit example see Section 6 below. Having estimated the parameters of γ, the parameter θ is estimated from the conditional distribution given $x(0)$. Excluding a multiplicative term, depending on γ only, (14) can be written as

$$L_i^j = \exp\Big\{-\sum_r \theta_r W_r^{ij}\Big\} \prod_r \theta_r^{N_r^{ij}}, \tag{15}$$

where N_r^{ij} is the number of jumps of type r experienced by nucleotide j within codon i, and $W_r^{ij} = \int_0^1 \gamma_r\left[j\middle|x_i^{(j-2):(j+2)}(s)\right]$ with γ_r the sum of $\gamma\left[j,\nu\middle|x_i^{(j-2):(j+2)}(s)\right]$ over those ν for which the jump is of type r, $R = r$. The likelihood equations for θ, therefore, become

$$\theta_r W_r = N_r, \quad j = 1, \ldots, p, \tag{16}$$

where $W_r = \sum_{i,j} W_r^{ij}$ and $N_r = \sum_{i,j} N_r^{ij}$ is the number of jumps of a particular type r. From the definition of W_r^{ij} one sees that equation (16) says in words that the expected number of jumps of a particular type should equal the observed number. In case θ is a function of a parameter ξ, the likelihood equation for ξ becomes

$$\sum_r \{N_r - \theta_r W_r\} \frac{1}{\theta_r} \frac{\partial \theta_r}{\partial \xi} = 0. \tag{17}$$

Having observed $x(0)$ and $x(1)$ only, define $L_m(\theta)$ to be the marginal likelihood. A classical formula says that

$$\frac{L_m(\theta)}{L_m(\theta_0)} = E_{\theta_0}\left\{\frac{L(\theta)}{L(\theta_0)}\middle| x(0), x(1)\right\}, \tag{18}$$

where $L(\theta)$ is the full likelihood from observing all of $x(t)$. Jensen and Pedersen (2000) evaluated the mean value in (18) by a Gibbs sampler. The path of a single codon were updated conditioned on the paths of all the other codons. The distribution of the latter conditional path depends on the paths of the two neighbouring codons only, making this a feasible approach. It was not possible to simulate directly from the conditional path and, instead, a path was suggested which were then accepted with a suitable probability (Metropolis-Hasting step). When considering more than two sequences, connected in a known phylogeny, the suggestion of a conditional path becomes more complicated. This is the main reason for introducing a time discretized version of the model where it is possible to simulate directly from the conditional distribution of a codon path. Furthermore, the estimation approach via a maximization of (18) can be quite demanding in terms of computer time. This is because the variance of the simulated mean value grows exponentially with the length of the DNA sequence. For this reason it is much better to use an estimating function as for example

$$\frac{\partial l_m}{\partial \theta}(\theta) = E_\theta\left[\frac{\partial l}{\partial \theta}(\theta)\middle| x(0), x(1)\right], \tag{19}$$

where l and l_m are log likelihoods. Using the particular function in (19) can be combined with an EM algorithm for finding the estimates, but more general estimating functions can often be combined with an EM type algorithm. We consider such a possibility below using a simple estimating function.

When the continuous time process is time discretized, let the probability of a change be τ times the above rate (12), where τ reflects the length of the time step. Furthermore, to make sure that at each time step a stop codon (one of TAA, TAG, and TGA) is not introduced within a reading frame, nucleotides at positions 1 are updated before nucleotides at positions 2 and 3, and nucleotides at positions 2 are

updated before nucleotides at positions 3. Another way of thinking about this is to imagine that each time step is subdivided into three steps and changes at the jth codon position are only allowed at the jth of the three substeps. Taking into account the restrictions from the reading frames in this way has no influence on the fact that as the time step tends to zero the continuous time model is retrieved. Instead of using the full likelihood from the discrete time model we use estimating equations of the same form as (16) to estimate the parameters, that is, we equate the expected number of jumps to the observed number of jumps. Of course, in the limit where the time step tends to zero, the estimates from the continuous time full likelihood are recovered.

The discrete time model is an approximation to the continuous time model. However, when the evolutionary distance between the two sequences is small we think of the discrete time model as representing reality equally well as the continuous time model even though the number of discrete time steps is taken to be a small number.

5 Time discretized model: two sequences

5.1 Set-up

Let $x(m)$ be the codon sequence at the discrete time points $m = 1, \ldots, M$. The observed sequences are $y = x(0)$ and $z = x(M)$. Substitutions that produce a stop codon are not allowed and, therefore, we let the state space of $x(m)$ be all sequences with no stop codons along the sequence within the relevant reading frames. Using this state space one avoids writing the prohibition of stop codons explicitly in the instantaneous substitution rates.

To take into account multiple reading frames we let in our most general model the probability of a change in codon position j depend on the two previous nucleotides, the two following nucleotides, as well as the nucleotide being changed. Also, as mentioned above, changes at position one within a time step takes place before changes at position two that in turn takes place before changes at position three. Formally, the transition probability for a change of $x_i^j(m)$ to the new nucleotide ν can functionally be written as

$$\begin{aligned} &p_1\left(\nu \middle| x_{i-1}^2(m), x_{i-1}^3(m), x_i^1(m), x_i^2(m), x_i^3(m)\right), \\ &p_2\left(\nu \middle| x_{i-1}^3(m), x_i^1(m+1), x_i^2(m), x_i^3(m), x_{i+1}^1(m+1)\right), \\ &p_3\left(\nu \middle| x_i^1(m+1), x_i^2(m+1), x_i^3(m), x_{i+1}^1(m+1), x_{i+1}^2(m+1)\right), \end{aligned} \tag{20}$$

where p_j is the probability for a change in position j. Note, how p_2 and p_3 incorporate the rule that all nucleotides at codon positions one are updated before nucleotides at positions two, that in turn are updated before nucleotides at positions three. When calculating the likelihood function for the path the two boundary codons $x_0(m)$ and $x_{n+1}(m)$ are considered nonrandom (typically these will be a start codon and a stop

codon). The likelihood function, conditionally on $x(0)$, is then

$$\begin{aligned} L = \prod_{m=1}^{M}\prod_{i=1}^{n} & p_1\left(x_i^1(m)\middle|x_{i-1}^2(m-1), x_{i-1}^3(m-1), x_i(m-1)\right) \\ & \times p_2\left(x_i^2(m)\middle|x_{i-1}^3(m-1), x_i^1(m), x_i^2(m-1), x_i^3(m-1), x_{i+1}^1(m)\right) \\ & \times p_3\left(x_i^3(m)\middle|x_i^1(m), x_i^2(m), x_i^3(m-1), x_{i+1}^1(m), x_{i+1}^2(m)\right). \end{aligned} \tag{21}$$

For the estimation to be described below we want to simulate the process $x(t)$, $t = 1, \ldots, M-1$, conditionally on the values of $x(0)$ and $x(M)$. This is done via a Markov chain Monte Carlo method using a Gibbs update, that is, we update the path $x_i^j(m)$, $m = 1, \ldots, M$, conditionally on the paths of all the other nucleotides.

5.2 Gibbs update

In order to perform the Gibbs update, we need the conditional distribution of the nucleotide path $x_i^j(m)$, $m = 1, \ldots, M$, given the paths of all other nucleotides. To make the formulae below more transparent, the path of interest is denoted by $\nu(m)$, $m = 1, \ldots, M$. Collecting all the terms in (21) that contain the relevant path, it is seen that the conditional density is proportional to:

case $j = 1$:

$$\begin{aligned} \prod_{m=1}^{M} & p_2\left(x_{i-1}(m)^2\middle|x_{i-2}^3(m-1), x_{i-1}^1(m), x_{i-1}^2(m-1), x_{i-1}^3(m-1), \nu(m)\right) \\ & \times p_3\left(x_{i-1}^3(m)\middle|x_{i-1}^1(m), x_{i-1}^2(m), x_{i-1}^3(m-1), \nu(m), x_i^2(m)\right) \\ & \times p_1\left(\nu(m)\middle|x_{i-1}^2(m-1), x_{i-1}^3(m-1), \nu(m-1), x_i^2(m-1), x_i^3(m-1)\right) \\ & \times p_2\left(x_i^2(m)\middle|x_{i-1}^3(m-1), \nu(m), x_i^2(m-1), x_i^3(m-1), x_{i+1}^1(m)\right) \\ & \times p_3\left(x_i^3(m)\middle|\nu(m), x_i^2(m), x_i^3(m-1), x_{i+1}^1(m), x_{i+1}^2(m)\right), \end{aligned} \tag{22}$$

case $j = 2$:

$$\begin{aligned} \prod_{m=1}^{M} & p_3\left(x_{i-1}^3(m)\middle|x_{i-1}^1(m), x_{i-1}^2(m), x_{i-1}^3(m-1), x_i^1(m), \nu(m)\right) \\ & \times p_1\left(x_i^1(m+1)\middle|x_{i-1}^2(m), x_{i-1}^3(m), x_i^1(m), \nu(m), x_i^3(m)\right) \\ & \times p_2\left(\nu(m)\middle|x_{i-1}^3(m-1), x_i^1(m), \nu(m-1), x_i^3(m-1), x_{i+1}^1(m)\right) \\ & \times p_3\left(x_i^3(m)\middle|x_i^1(m), \nu(m), x_i^3(m-1), x_{i+1}^1(m), x_{i+1}^2(m)\right) \\ & \times p_1\left(x_{i+1}^1(m+1)\middle|\nu(m), x_i^3(m), x_{i+1}^1(m), x_{i+1}^2(m), x_{i+1}^3(m)\right), \end{aligned} \tag{23}$$

case $j = 3$:

$$\prod_{m=1}^{M} p_1\left(x_i^1(m+1)\middle|x_{i-1}^2(m), x_{i-1}^3(m), x_i^1(m), x_i^2(m), \nu(m)\right) \tag{24}$$
$$\times\, p_2\left(x_i(m+1)^2\middle|x_{i-1}^3(m), x_i^1(m+1), x_i^2(m), \nu(m), x_{i+1}^1(m+1)\right)$$
$$\times\, p_3\left(\nu(m)\middle|x_i^1(m), x_i^2(m), \nu(m-1), x_{i+1}^1(m), x_{i+1}^2(m)\right)$$
$$\times\, p_1\left(x_{i+1}^1(m+1)\middle|x_i^2(m), \nu(m), x_i^1(m), x_{i+1}^2(m), x_{i+1}^3(m)\right)$$
$$\times\, p_2\left(x_{i+1}(m+1)^2\middle|\nu(m), x_{i+1}^1(m+1), x_{i+1}^2(m), x_{i+1}^3(m), x_{i+2}^1(m+1)\right),$$

where for $m = M$ the terms in the product with $m+1$ are not present. Since each term in the product $\prod_{i=1}^{M}$ depends on $\nu(m)$ and $\nu(m-1)$ only, we can rewrite these conditional densities as an inhomogeneous Markov chain for $m = 1, \ldots, M$. This is also true when $\nu(M)$ is fixed at the value given by the z-sequence. The Markov structure makes it easy to simulate a new path $\nu(m)$, $m = 1, \ldots, M$.

Let us formally write one of the products in (22), (23), and (24) as

$$\prod_{i=1}^{M} g_m(\nu(m); \nu(m-1)). \tag{25}$$

Then the inhomogeneous Markov chain is given by the transition probabilities

$$q_m(\nu(m)|\nu(m-1)) = \frac{g_m(\nu(m); \nu(m-1))h_m(\nu(m))}{h_{m-1}(\nu(m-1))},$$

where the functions h_m, $m = 0, \ldots, M$, are defined recursively by

$$h_M(\nu) = 1(\nu = z_i^j),$$

and for $m = M, \ldots, 1$:

$$h_{m-1}(\xi) = \sum_{\nu} g_m(\nu; \xi)h_m(\nu). \tag{26}$$

5.3 Estimation

As mentioned in Section 4, the model contains a parameter θ that we want to estimate based on the transition probabilities. In this section we describe an EEE algorithm for finding the estimates. *EEE* is an acronym for Expectation-Estimating-Equation. The probabilities and expectations below are for the conditional measure given $x(0)$. Let

$$\Psi(\theta, x(\cdot)) = 0 \tag{27}$$

be an estimating equation for the parameter θ based on observing $x(m)$ at all time points $m = 1, \ldots, M$. Observe $x(M)$ (and $x(0)$) only, this equation is not useful. Instead, we obtain an estimate $\hat{\theta}$ by solving the estimating equation

$$E_\theta[\Psi(\theta, x(\cdot))|x(M)] = 0. \tag{28}$$

To solve this equation, an iterative procedure is used, where θ_{k+1} is found from θ_k by solving

$$E_{\theta_k}[\Psi(\theta, x(\cdot))|x(M)] = 0.$$

The expectation is calculated using MCMC and the Gibbs update described in the previous subsection. For articles related to the EEE algorithm see Heyde and Morton (1996), Rosen et al. (2000), and Elashoff and Ryan (2004).

Let $L(\theta, x(\cdot))$ be the full likelihood from observing all the evolutionary events and let $L(\theta, x(M))$ be the likelihood from observing $x(M)$ only. We write the latter formally as $\int L(\theta, x(\cdot))d\mu[x(\cdot)|x(M)]$ and, similarly, we write

$$E_\theta[\Psi(\theta, x(\cdot))|x(M)] = \frac{\int \Psi(\theta, x(\cdot))L(\theta, x(\cdot))d\mu[x(\cdot)|x(M)]}{\int L(\theta, x(\cdot))d\mu[x(\cdot)|x(M)]}.$$

To calculate the "observed information", the derivative of the left hand side of $E_\theta[\Psi(\theta, x(\cdot))|x(M)]$ is written as

$$\begin{aligned}
J(\theta) = & -\frac{\partial}{\partial\theta}E_\theta\left[\Psi(\theta, x(\cdot))|x(M)\right] \\
= & \frac{\int \left[-\frac{\partial\Psi}{\partial\theta}(\theta, x(\cdot))L(\theta, x(\cdot)) - \Psi(\theta, x(\cdot))\frac{\partial L}{\partial\theta}(\theta, x(\cdot))\right] d\mu[x(\cdot)|x(M)]}{L(\theta, x(M))} \\
& + \frac{\int \Psi(\theta, x(\cdot))L(\theta, x(\cdot))d\mu[x(\cdot)|x(M)]}{L(\theta, x(M))^2}\int \frac{\partial L}{\partial\theta}(\theta, x(\cdot))d\mu[x(\cdot)|x(M)] \\
= & E_\theta\left[-\frac{\partial\Psi}{\partial\theta}(\theta, x(\cdot))\middle|x(M)\right] - V_\theta\left[\Psi(\theta, x(\cdot)), \frac{\partial l}{\partial\theta}(\theta, x(\cdot))\right], \qquad (29)
\end{aligned}$$

where l is the log likelihood corresponding to the likelihood L. When inserting $\theta = \hat{\theta}$ the covariance term reduces to the conditional mean of the product of the two terms, and thus

$$J(\hat{\theta}) = \left\{E_\theta\left[-\frac{\partial\Psi}{\partial\theta}(\theta, x(\cdot))\middle|x(M)\right] - E_\theta\left[\Psi(\theta, x(\cdot))\frac{\partial l}{\partial\theta}(\theta, x(\cdot))\middle|x(M)\right]\right\}\Bigg|_{\theta=\hat{\theta}}. \quad (30)$$

For the usual EM algorithm, where $\Psi(\theta, x(\cdot)) = \frac{\partial l}{\partial\theta}(\theta, x(\cdot))$, the corresponding formula is given in Louis (1982). For the maximum likelihood estiamte $J(\hat{\theta})$ is used as the asymptotic variance, however, in the general case we also need to calculate the variance $\Sigma(\theta)$ of $E_\theta\left[\Psi(\theta, x(\cdot))|x(M)\right]$ to obtain the asymptotic variance

$$J(\hat{\theta})^{-1}\Sigma(\hat{\theta})J(\hat{\theta})^{-1}$$

of $\hat{\theta}$.

Let us for a moment consider the continuous time model and let us take Ψ to be the score function as given through (16) and (17). To calculate the observed information, the conditional means of the terms W_r and the conditional means of the terms $(N_r - \theta_r W_r)(N_s - \theta_s W_s)$ are needed. For the time discretized model we take the estimating function Ψ to resemble the score function from the continous time model, and when the number of time steps M is large, this analogy can be used to calculate the observed information.

6 Specific model

We now consider a specific model where the transition probabilities $p_j(\cdot|\cdot)$ from (20) are a product of three terms. First, there is a term A that is symmetric in the old and new nucleotide. Typically, this term depends on whether the change is a transition or a transversion and whether the amino acid is changed in one of the reading frames in use. Next, there is a term caused by dinucleotide interactions and, finally, a term D dependent on the new nucleotide. Often, D represents nucleotide frequencies. Let $v = (v^1, v^2, v^3)$ be a generic codon and let v^{-1}, v^0 be the nucleotides at positions 2 and 3 in the left flanking codon and let v^4, v^5 be the nucleotides at positions 1 and 2 in the right flanking codon. The transition probability for a change of v^j to ν, $j = 1, 2, 3$, is

$$\begin{aligned} &p_j(\nu\big|v^{j-2}, v^{j-1}, v^j, v^{j+1}, v^{j+2}) \\ &\quad = A_j(\nu, v^j; v^{j-2}, v^{j-1}, v^{j+1}, v^{j+2}) \frac{\phi_j(v^{j-1}, \nu)\phi_{j+1}(\nu, v^{j+1})}{\phi_j(v^{j-1}, v^j)\phi_{j+1}(v^j, v^{j+1})} D_j(\nu), \end{aligned} \tag{31}$$

where $\phi_4 = \phi_1$. We also consider the model with $D_j(\nu)$ replaced by $D(v(j, \nu))$, where $v(j, \nu)$ is the codon obtained from v by replacing v^j by ν. In this way codon frequencies enter instead of nucleotide frequencies. The ϕ terms in this expression represent dinucleotide interaction. The nominator, where ν enters, can be seen as a selection mechanism, whereas the denominator, that depends on the present nucleotides only, can be seen as a change in the mutation rate. The model with the ϕ part of (31) replaced by the more general term

$$\frac{\phi_j(v^{j-1}, \nu)\phi_{j+1}(\nu, v^{j+1})}{\psi_j(v^{j-1}, v^j)\psi_{j+1}(v^j, v^{j+1})}$$

can be treated in exactly the same way as the model (31).

6.1 Stationary distribution

The model with transition probabilities (31) is time reversible and the stationary distribution for a sequence $x = (x_1, \ldots, x_n)$ is given by

$$\pi(x) = \frac{1}{C}\phi_1(x_n^3, x_{n+1}^1)^2 \prod_{i=1}^{n} \phi_1(x_{i-1}^3, x_i^1)^2 \phi_2(x_i^1, x_i^2)^2 \phi_3(x_i^2, x_i^3)^2 D_1(x_i^1) D_2(x_i^2) D_3(x_i^3), \tag{32}$$

where C is a norming constant and x_0 and x_{n+1} are fixed. For the model where $D_j(\nu)$ is replaced by $D(v(j, \nu))$, the term $D_1(x_i^1)D_2(x_i^2)D_3(x_i^3)$ in the stationary density is replaced by $D(x_i^1, x_i^2, x_i^3)$.

To prove the above statement, we show directly that $\pi(x)p(y|x) = \pi(y)p(y|x)$ for a one step transition probability $p(\cdot|\cdot)$. Since, in our model, nucleotides at codon positions 1 are updated before nucleotides at positions 2 and 3, one can look at these

transition probabilities seperately. Let us consider updatings at codon positions 1 and let the present sequence be x and the new sequence be y, where

$$y_i^2 = x_i^2 \quad \text{and} \quad y_i^3 = x_i^3, \quad i = 1, \ldots, n.$$

The aim is to prove that

$$\begin{aligned}\pi(x) \prod_{i=1}^{n} p_1(y_i^1 \big| x_{i-1}^2(m), x_{i-1}^3(m), x_i^1(m), x_i^2(m), x_i^3(m)) \\ = \pi(y) \prod_{i=1}^{n} p_1(x_i^1 \big| y_{i-1}^2(m), y_{i-1}^3(m), y_i^1(m), y_i^2(m), y_i^3(m)). \end{aligned} \tag{33}$$

Terms in the products where $y_i^1 = x_i^1$ can be removed. Similarly, when inserting (31), the A_1 term can be removed due to the symmetri in the first two arguments. The equation (33) then reduces to

$$\begin{aligned}\prod_{i: y_i^1 \neq x_i^1} \phi_1(x_{i-1}^3, x_i^1)\phi_2(x_i^1, x_i^2)\phi_1(x_{i-1}^3, y_i^1)\phi_2(y_i^1, x_i^2)D_1(x_i^1)D_1(y_i^1) \\ = \prod_{i: x_i^1 \neq y_i^1} \phi_1(x_{i-1}^3, y_i^1)\phi_2(y_i^1, x_i^2)\phi_1(x_{i-1}^3, x_i^1)\phi_2(x_i^1, x_i^2)D_1(y_i^1)D_1(x_i^1),\end{aligned}$$

which is clearly true. Updatings at positions 2 and 3 are treated in the same way. For the model where $D_j(\nu)$ is replaced by $D(v(j, \nu))$ the term $D_1(x_i^1)D_1(y_i^1)$ in the above argument is replaced by $D(x_i)D(y_i)$.

When looking at (32), one should keep in mind that the state space is all sequences with no stop codons along the sequence in the relevant reading frames.

6.2 Estimation

We split the estimation into two steps. For parameters entering ϕ_j and D_j, $j = 1, 2, 3$, we use the stationary density (32) to find estimates. Details on how to solve this estimation problem are given in Appendix D. For parameters entering A_j we use the transition probability when going from the y-sequence to the z-sequence, and the optimization is done via the stochastic EEE algorithm mentioned in subsection 5.3.

As an example, consider the Goldman and Yang model (2), with one reading frame only, defined by

$$A_j = \begin{cases} \tau & \text{if synonymous transition,} \\ \tau\beta & \text{if synonymous transversion,} \\ \tau\xi & \text{if nonsynonymous transition,} \\ \tau\eta & \text{if nonsynonymous transversion.} \end{cases} \tag{34}$$

Here τ has the interpretation of a time distance between the two sequences. The full likelihood (21) is in this case

$$\tau^{N_1}(\tau\beta)^{N_2}(\tau\xi)^{N_3}(\tau\eta)^{N_4} \prod_{\substack{m,i,j: \\ x_i^j(m) = x_i^j(m-1)}} p_j, \tag{35}$$

where N_i, $i = 1, \ldots, 4$, counts the number of jumps of the particular types given in (34). In (35) the argument of p_j has been left out. Each of these p_j terms consists of 1 minus a linear combinations of τ, $\tau\beta$, $\tau\xi$, and $\tau\eta$, and there is a large number of different linear combinations. Maximization of this function cannot be done analytically and, instead, we suggest to base the estimation on a set of simple estimating equations. The suggestion is to equate the expected number of changes of a particular type to the actual number in the full likelihood. We then take the conditional mean (via simulations), given the observed sequences, on both sides of the equation and solve the resulting equation. In explicit terms, this gives the equation

$$\theta_r \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{j=1}^{3} \sum_{(\nu, x_i^j(m-1)) \in T_r} \omega(\nu; i, j, m) D_j(\nu) = \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{j=1}^{3} 1\left[(x_i^j(m), x_i^j(m-1)) \in T_r\right], \tag{36}$$

where

$$\omega(\nu; i, j, m) = \frac{\phi_j(\tilde{x}_b, \nu)\phi_{j+1}(\nu, \tilde{x}_a)}{\phi_j(\tilde{x}_b, x_i^j(m-1))\phi_{j+1}(x_i^j(m-1), \tilde{x}_a)}, \qquad \phi_4 = \phi_1,$$

with

$$\tilde{x}_b = \begin{cases} x_{i-1}^3(m-1) & j = 1, \\ x_i^{j-1}(m) & j = 2, 3, \end{cases} \quad \text{and} \quad \tilde{x}_a = \begin{cases} x_i^{j+1}(m-1) & j = 1, 2, \\ x_{i+1}^1(m) & j = 3. \end{cases}$$

In this equation $\theta_1 = \tau$, $\theta_2 = \tau\beta$, $\theta_3 = \tau\xi$, and $\theta_4 = \tau\eta$, and T_r is the corresponding set of changes according to (34).

When running the EEE algorithm, we simulate the conditional mean of both sides of (36) (not including θ_r), using the present values of the parameters, and find a new value of θ_r as the ratio of the two conditional means. To find the conditional mean, given the observed sequences y and z, the Gibbs sampler from subsection 5.2 is used.

Let us write the conditional mean of (36) in the form

$$\theta_r g_r = h_r, \quad r = 1, 2, 3, 4. \tag{37}$$

When we consider a model with restrictions on the parameters the analogy to (17) is used to obtain a set of estimating equations. As an example, when we consider the restriction $\eta = \beta\xi$ the analogy to (17) gives the equations

$$\begin{aligned} h_1 + h_2 + h_3 + h_4 &= \tau(g_1 + \beta g_2 + \xi g_3 + \beta\xi g_4), \\ h_2 + h_4 &= \tau\beta(g_2 + \xi g_4), \\ h_3 + h_4 &= \tau\xi(g_3 + \beta g_4), \end{aligned}$$

that are easily solved.

6.3 Comparison with pseudolikelihood estimates

In Christensen et al. (2004) a pseudo likelihood is suggested for obtaining simple estimates in the type of models considered in this paper. The psudo likelihood consists of a product over codon sites of the probabilities related to the evolutionary events at the site, and can be maximized using an EM algorithm. A comparison with the estimates obtained using the method of this paper was given. Using simulated data with various degrees of dinucleotide interactions, in the form of CG-depression, it was found that the two sets of estimates were almost identical for small to moderate evolutionary distances between the two sequences. The method in Christensen et al. (2004) has as yet been developed for the analysis of two sequences only.

7 Three (or more) sequences

We now extend the model to the case of several sequences connected in a phylogenetic tree. To make the notation as simple as possible, we consider the case of three sequences, where all the features of the multiple sequence case are present.

The three sequences are connected in a 3-star tree. The likelihood is calculated as though the observed sequence y is the ancestor that develops along branch 1 into a sequence a at the inner node. The sequence a next develops into the observed sequences z and u along branches 2 and 3. Along each branch there is a process as above with M discrete time steps. The branch length appears through a separate value of τ in (34) for each branch. Below we start by generalizing the Gibbs update from Subsection 5.2 to the case here and next generalize the estimation procedure of Subsection 6.2. We conclude with the analysis of a small data set.

7.1 Gibbs update

In (25) the terms entering the conditional distribution of a nucleotide path along a branch are given. For a 3-star tree there are three sets of products as in (25), one for each branch. For a branch ending at an inner node, in our case branch 1, the Mth term in (23) and (24) is slightly changed. In (23) the two terms with p_1 are doubled for $m = M$ since there is a term for each of the two branches descending from the inner node. Similarly, for (24) the four terms with p_1 and p_2 are doubled for $m = M$. We use an upper index k to indicate the branch on g_m from (25) and on the path $\nu(m)$. Thus $\nu^1(0)$ is the observed nucleotide y_i^j from the y-sequence, $\nu^2(M)$ is the observed nucleotide z_i^j from the z-sequence, and $\nu^3(M)$ is the observed nucleotide w_i^j from the w-sequence. Furthermore, $\nu^2(0) = \nu^3(0) = \nu^1(M)$. Now define the functions $h_m^1(\nu)$, $h_m^2(\nu)$, and $h_m^3(\nu)$ by the backward recursion in (26) with

$$h_M^2(\nu) = 1(\nu = z_i^j), \quad h_M^3(\nu) = 1(\nu = w_i^j),$$

and

$$h_M^1(\nu) = h_0^2(\nu)h_0^3(\nu).$$

Then the conditional path is given as an inhomogenous Markov chain with transition probabilities

$$q_m^k(\nu^k(m)|\nu^k(m-1)) = \frac{g_m^k(\nu^k(m);\nu^k(m-1))h_m^k(\nu^k(m))}{h_{m-1}^k(\nu^k(m-1))},$$

where one first simulates $\nu^1(1),\ldots,\nu^1(M)$ and next simulate $\nu^2(1),\ldots,\nu^2(M)$ and $\nu^3(1),\ldots,\nu^3(M)$.

For a general tree, we use the same method, where $h_m^k(\nu)$ is calculated backward in time (with respect to m and k), and a new path is simulated forward in time.

7.2 Estimation

We consider the model (34) with the same transition probabilities along the three branches except for a different time scaling τ_k, $k = 1, 2, 3$.

Let us write equation (37) in the form

$$\theta_r g_r^k = h_r^k, \tag{38}$$

where θ_r for $r = 1, 2, 3, 4$ is τ_k, $\tau_k\beta$, $\tau_k\xi$, and $\tau_k\eta$, respectively, and where $k = 1, 2, 3$ is the branch number. In (38) g_r^k and h_r^k are the conditional means of the terms in (36), given the observed sequences, which are calculated by simulations using the Gibbs update. The 12 equations of the form (38) are reduced to 6 equations by summing some of them. The resulting equations are

$$\sum_r h_r^k = \tau_k\left(g_1^k + \beta g_2^k + \xi g_3^k + \eta g_4^k\right), \quad k = 1, 2, 3, \tag{39}$$

and

$$\sum_k h_2^k = \beta \sum_k \tau_k g_2^k, \tag{40}$$

$$\sum_k h_3^k = \xi \sum_k \tau_k g_3^k, \tag{41}$$

$$\sum_k h_4^k = \eta \sum_k \tau_k g_4^k. \tag{42}$$

We have used an iterative procedure to solve the above equations. For fixed values of β, ξ, and η, we find τ_1, τ_2, and τ_3 from (39), and then use the new values of τ_j to obtain new values of β, ξ, and η from (40-42). Since the above equations are analogous to the likelihood equations for the continuous time model, the iterative procedure is an iterative partial maximization algorithm (sometimes called Zellner's twostage procedure). Properties of this kind of iterative maximization are given in (Lauritzen, 1996, Appendix A.4) and in (Drton, 2004, Appendix A).

7.3 Data example

We consider three short aligned sequences from sus (domestigated pig), man, and mouse consisting of 337 codons. The accession number of the human gene is

NM_012111 (the gene name is AHA1), and the three sequences used here are part of the investigation in Jørgensen et al. (2005). The alignment of the three sequences is gap free. We use the model with one reading frame given in (31) and (34) with

$$\phi_1(a,b) = \phi_2(a,b) = \phi_3(a,b) = \lambda^{1(a=C,b=G)}. \tag{43}$$

When $\lambda < 1$ this is the socalled CG-depression where CG pairs are seen more rarely than predicted from the nucleotide frequencies. Furthermore, for each j we let $D_j(\nu)$ be a probability distribution on the four nucleotides. The stationary distribution (32), on the set of sequences with no stop codons, can then be written

$$\frac{1}{C}\lambda^{2N_{CG}} \prod_{j=1}^{3} \prod_{\nu=A}^{T} D_j(\nu)^{N(j,\nu)}, \tag{44}$$

where N_{CG} is the number of CG pairs along the sequence and $N_j(\nu)$ is the number of times the nucleotide ν appears at codon position j. We estimate λ and $D_j(\nu)$ from this marginal distribution. Details of the estimation are given in Appendix D.

The estimates of the CG-depression and the nucleotide probabilities obtained from the marginal distribution of the sequences are

$\lambda = 0.50 \pm 0.054,$

	A	G	C	T
position 1	0.253	0.407	0.158	0.183
position 2	0.332	0.178	0.267	0.224
position 3	0.180	0.321	0.335	0.164

With $\lambda \approx 0.50$ these data shows a clear CG-depression. The algorithm described in this paper, for estimation based on the transition probabilities, was run with $M = 40$ time steps in each branch of the tree. We obtained estimates both using all three sequences connected in a 3-star tree and using two sequences only: sus-man, sus-mouse, and man-mouse. The estimated Goldman-Yang parameters are

	β	ξ	η
3-star	0.37 ± 0.082	0.080 ± 0.021	0.019 ± 0.0064
sus-man	0.25	0.061	0.015
sus-mouse	0.35	0.076	0.027
man-mouse	0.38	0.070	0.026

Note that these values indicate that the nonsynonymous transversion rate η is roughly the product of β and ξ. The estimate of the branch length τ, together with the standardized version T, calculated as the expected number of nucleotide changes per codon, are

	τ, 3-star	T, 3-star	T, 2-seq
sus-star	0.012 ± 0.0022	0.17	0.17
star-man	0.0067 ± 0.0017	0.10	0.10
star-mouse	0.018 ± 0.0029	0.25	0.24

The last column contains the expected number of nucleotide changes per codon obtained by performing three seperate estimations for a pair of sequences.

The standard deviations were calculated from the observed information which were found as described in section 5.3. We used the approximation to the observed information in the continuous time model as obtained from the discrete time approximation.

8 Asymptotics

As will be explained below the process x_i, $i = 1, \ldots, n$, can be viewed as a Markov chain. When observing $x_i(M)$ only, we are, therefore, in the situation of a hidden Markov chain. Asymptotic results as $n \to \infty$ for the maximum likelihood estimator in a hidden Markov chain can be found in Douc et al. (2004), Jensen and Petersen (1999), Baum and Petrie (1966). The situation in this paper differs from the above papers by considering an estimate obtained from an estimating equation and by having boundary conditions at each end of the Markov chain.

We start by establishing exponential mixing of the evolutionary process along the sequence. Using this, and a central limit theorem based on the work of Götze and Hipp (1983), we prove asymptotic normality of the estimating function in (28). Convergence of the observed information is established from the mixing properties and from ergodicity. Using these results, established in Subsections 8.1-8.4, standard asymptotic theory gives the asymptotic normality of the estimate of θ obtained from solving (28)

8.1 Mixing

For simplicity we consider the model with one reading frame only, given through (34) with $\beta \leq 1$, $\xi \leq 1$, $\zeta \leq 1$, and with $3\tau < 1$. In this case the full likelihood (21) becomes

$$w_1(x_1) \prod_{i=2}^{n} w(x_i, x_{i-1}) w_n(x_n), \tag{45}$$

where, leaving out some of the functional arguments, w is given as

$$w(x_i, x_{i-1}) = \prod_{m=1}^{M} p_3\left(x_{i-1}^3(m)|\cdot\right) p_1\left(x_i^1(m)|\cdot\right) p_2\left(x_i^2(m)|\cdot\right), \tag{46}$$

and w_1 contains p_1 and p_2 only, and w_n contains p_3 only. The structure in (45) implies that, when conditioning on x_i, the past $(x_1, \ldots, x_{i-1})$ and the future $(x_{i+1}, \ldots, x_n)$ are independent, which means that x_i, $i = 1, \ldots, n$, is an inhomogeneous Markov chain. The inhomogeneity is because $x_i(0)$ is fixed and acts as a parameter in w. Below we consider in some of the arguments the process with no conditioning on $x(0)$, in which case w includes the terms

$$\prod_{j=1}^{3} \phi_j(x_i^{j-1}(0), x_i^j(0)) \, D_j(x_i^j(0)) \tag{47}$$

from the stationary density (32), and now the Markov chain is homogeneous. Depending on whether one conditions on $x(0)$ and $x(M)$ the Markov chain has 61^K states with $K = M-1$, M, or $M+1$. For a finite state Markov chain it is fairly easy to obtain mixing results and we describe this below. This is relevant when one thinks of the number of time steps M as small and fixed. However, when looking at the time discretized model as an approximation to the continuous time model, the limit $M \to \infty$ is of interest and it becomes relevant to make mixing estimates independent of M.

Let $N_s = \{x_s | x_s(m) = x_s(m-1), m = 1, \ldots, M\}$ be the event of no jumps in codon s. Also, we use the notation $p(A_s|u_r, v_t) = P(X_s \in A | x_r = u, x_t = v)$ and the corresponding notation when conditioning on one variable only. For the case of M fixed there exists a $\rho < 1$ such that, for all s and all values of u, v, we have

$$P(N_s|u_{s-1}, v_{s+1}) \geq 1 - \rho. \tag{48}$$

Here the conditioning is on $x_{s-1} = u$ and $x_{s+1} = v$ and the same bound trivially applies when conditioning on $x_{s-1} = u$ only. The proof of (48) runs as follows. We write formally the conditional probability of x_s given x_{s-1} and x_{s+1} as

$$\frac{1}{Z} \prod_{m=1}^{M} p_3(x_{s-1}^3(m)|\cdot) p_1(x_s^1(m)|\cdot) p_2(x_s^2(m)|\cdot) p_3(x_s^3(m)|\cdot) p_1(x_{s+1}^1(m)|\cdot), \tag{49}$$

where Z is a normalizing constant. Let $\gamma = \min\{\beta, \xi, \zeta, 1-3\tau\} \leq 1$ (remember that $\beta \leq 1$, $\xi \leq 1$, $\zeta \leq 1$, and $3\tau < 1$). In the case of no jumps in x_s, we bound each term in the product from below by $\gamma(1-3\tau)(1-3\tau)(1-3\tau)\gamma$ leading to the lower bound

$$\frac{1}{Z} \gamma^M (1-3\tau)^{3M} \gamma^M. \tag{50}$$

Similarly, to get an upper bound for (49), we use the bound 1 for the first p_3 term and the last p_1 term as well as the remaining terms when there are no jumps, and the bound τ when there is a jump. This gives the bound τ^K, where K is the number of jumps in x_s, and summing over K lead to the bound $Z \leq (1+\tau)^{3M}$ (here we have counted all jump sequences although some of these contain a stop codon). Thus $1-\rho$ in (48) can be taken as $[(1-3\tau)/(1+\tau)]^{3M} \gamma^{2M}$. When there is no conditioning on $x(0)$, (47) is included in (49) and the sum runs from $m = 0$. However, this does not change the above estimation of $1-\rho$. Note, also, that for the parameters in a compact set we can choose ρ independent of the parameters so that (48) is valid for all values of the parameters.

We next study the mixing properties of x_i conditioned on $x_i(M)$, $i = 1, \ldots, n$. This process is again an inhomogenuous Markov chain. To obtain a result as in (48), consider (49), now with $x_s(M)$ fixed. If $x_s(M) = x_s(0)$, we use as above the event N_s of no jumps in x_s. If $x_s(M) \neq x_s(0)$, we use another event where $x_s(0)$ is changed to $x_s(M)$ in the smallest number of jumps (at most 3 jumps) using the first few time points. Instead of (50), this gives the lower bound

$$\frac{1}{Z} \gamma^M (\gamma\tau)^q (1-3\tau)^{3M-q} \gamma^M,$$

where q is the number of nucleotide changes between $x_s(0)$ and $x_s(M)$. For the norming constant Z in (49) the upper bound from before applies, the difference being that more jump sequences that are not allowed are included in the upper bound. Thus, we again obtain a lower bound as in (48) for a suitable value of ρ.

Let $r < s$ and let A be a fixed set. Define $D(r) = \max_u P(x_s \in A | x_r = u)$, $d(r) = \min_u P(x_s \in A | x_r = u)$, and $S_r = \{z : P(x_r = z | x_{r-1} = u) > P(x_r = z | x_{r-1} = v)\}$. Proceeding as in Doob (1953, page 198), it is seen that

$$\begin{aligned} D(r-1) - d(r-1) &= \max_{u,v} [P(A_s|u_{r-1}) - P(A_s|v_{r-1})] \\ &= \max_{u,v} \sum_z P(A_s|z_r) [P(z_r|u_{r-1}) - P(z_r|v_{r-1})] \\ &= \max_{u,v} \sum_z (D(r) - d(r)) [P(z_r|u_{r-1}) - P(z_r|v_{r-1})] \\ &\leq (D(r) - d(r)) \max_{u,v} [P(S_r|u_{r-1}) - P(S_r|v_{r-1})] \\ &\leq (D(r) - d(r))\rho. \end{aligned} \tag{51}$$

Iterating, we obtain

$$\max_{u,v} |P(A_s|x_r = u_r) - P(A_s| = v_r)| \leq \rho^{s-r}, \tag{52}$$

which shows that the process x_i, $i = 1, \ldots, n$, is mixing exponentially fast.

Let $r < s < t$ and define $D(r)$, $d(r)$, and S_r as above, except that the conditioning is on $x_r = u$ as well as $x_t = w$. The steps in (51) can now be repeated by adding $x_t = w$ to the conditioning event. The same bound as in (51) is obtained because

$$\begin{aligned} P(N_r|u_{r-1}, w_t) &= \sum_v P(N_r|u_{r-1}, v_{r+1}) P(v_{r+1}|u_{r-1}, w_t) \\ &\geq \sum_v (1-\rho) P(v_{r+1}|u_{r-1}, w_t) = 1 - \rho. \end{aligned}$$

As in (52) we obtain

$$\max_{u,v} |P(A_s|u_r, w_t) - P(A_s|v_r, w_t)| \leq \rho^{s-r}.$$

A similar argument gives

$$\max_{u,v} |P(A_s|w_r, u_t) - P(A_s|w_r, v_t)| \leq \rho^{t-s}.$$

Combining the two latter bounds lead to

$$\begin{aligned} &\max_{a,b,u,v} |P(A_s|a_r, b_t) - P(A_s|u_r, v_t)| \\ &\quad \leq |P(A_s|a_r, b_t) - P(A_s|u_r, b_t)| + |P(A_s|u_r, b_t) - P(A_s|u_r, v_t)| \\ &\quad \leq \rho^{s-r} + \rho^{t-s}. \end{aligned} \tag{53}$$

It is clear, also, that a similar argument can be used when considering the joint probability of (x_{s-1}, x_s, x_{s+1}), reducing the power of ρ by 1.

To summarize, the mixing result (52) and (53) can be used for the homogeneous process, where both $x(0)$ and $x(M)$ are stochastic, for the process where we condition on $x(0)$ or $x(M)$, and for the process where we condition on both $x(0)$ and $x(M)$.

8.2 Central limit theorem

The estimating function Ψ in (27), for the case when all of x is observed, takes in our case the form

$$\Psi(\theta, x) = \sum_{i=1}^{n} \psi_i(\theta) \quad \text{with} \quad \psi_i(\theta) = \psi(\theta, x_{i-1}, x_i, x_{i+1}), \tag{54}$$

where the ith term in the sum relates to the evolutionary events in codon i. We use the notation $E(\cdot|(i_1, i_2))$ for the conditional mean given $x_{i_1:i_2}(M) = (x_{i_1}(M), x_{i_1+1}(M), \ldots, x_{i_2}(M))$, and $E(\cdot|[i_1, i_2])$ for the conditional mean given $x_{i_1:i_2}(M)$ as well as x_{i_1} and x_{i_2}. The estimating funtion (28), based on observing $x(M)$ only, is then

$$E_\theta\left(\Psi(\theta, x)\middle|(1, n)\right) = \sum_{i=1}^{n} E_\theta\left(\psi_i(\theta)\middle|(1, n)\right), \tag{55}$$

where the expectation E_θ is for the measure conditioned on the value of $x(0)$. We want to show a central limit theorem for this sum.

The central limit theorem given in Jensen (2005), based on the work of Götze and Hipp (1983), is taylored to a situation as here. There are two requirements: an exponentially fast mixing of a set of sigma algebras $\{\mathcal{D}_j\}$, and an approximation with an exponentially small error in k to the individual terms in the sum by a variable measurable with respect to $\{\mathcal{D}_{i-k:i+k}\}$. For the case here let $\{\mathcal{D}_i\}$ be the σ-algebras generated by $x_i(M)$, which and exponentially mixing according to (52). Note that this is true both for the conditional process given $x(0)$ as well as the stationary process where the distribution of $x(0)$ is included. Furthermore, using (53), the conditional mean $E_\theta\left(\psi_i(\theta)\middle|(1, n)\right)$ can for each k be approximated by a function of $x_{i-k:i+k}(M)$ with an error that is exponentially small in k. The precise argument runs as follows. Since the state space is finite, there exists a constant c_1 such that

$$|\psi(\theta, \bar{x}_i)| \leq c_1 \quad \text{for all } \bar{x}_i = (x_{i-1}, x_i, x_{i+1}). \tag{56}$$

Then, for the case $i - k \geq 1$ and $i + k \leq n$, one finds that

$$\begin{aligned}
&\left|E_\theta\left(\psi_i(\theta)\middle|(1, n)\right) - E_\theta\left(\psi_i(\theta)\middle|(i-k, i+k)\right)\right| \\
&\quad = \left|\int E_\theta\left(\psi_i(\theta)\middle|[i-k, i+k]\right)\left\{P\left(d(x_{i-k}, x_{i+k})\middle|(1, n)\right)\right.\right. \\
&\qquad\qquad \left.\left. - P\left(d(x_{i-k}, x_{i+k})\middle|(i-k, i+k)\right)\right\}\right| \\
&\quad \leq 2c_1 \max_{A,a,b,u,v} |P^c(\bar{x}_i \in A|a_{i-k}, b_{i+k}) - P^c(\bar{x}_i \in A|u_{i-k}, v_{i+k})| \\
&\quad \leq 2c_1\left(\rho^{k-1} + \rho^{k-1}\right),
\end{aligned} \tag{57}$$

according to (53) for the measure P^c conditioned on $x(0)$ and $x(M)$. When $i-k < 1$ the approximation $E_\theta\left(\psi_i(\theta)\middle|(1, i+k)\right)$ is used instead together with (52). This gives the bound $2c_1\rho^{k-1}$ instead of (57). Similarly, when $i + k > n$, only one of the terms in (57) is used. The requirement in Jensen (2005) is that the mean of (57) is exponentially small, but since (57) is not stochastic this is of course trivially true,

both for the conditional process given $x(0)$ as well as the stationary process where the distribution of $x(0)$ is included.

In conclusion, a central limit theorem for (55) holds, both in the stationary process with $x(0)$ stochastic and in the conditional process given $x(0)$.

8.3 Uniform convergence of "observed information"

We now study $J(\theta)$ from (29). In (29), as well as in all of this section, the expectations, variances, and mixing bounds are from the conditional process given $x(0)$. Recall that $\psi_u(\theta) = \psi(\theta, \bar{x}_u)$ given in (54) and define $\psi_u^s(\theta)$ to be the sth coordinate of ψ_u and define $\psi_u^{rs}(\theta) = -\frac{\partial}{\partial\theta_r}\psi_u^s(\theta)$. The full likelihood (21) can be written as $\prod_{u=1}^n \omega_u(\theta)$, where $\omega_u(\theta)$ depends on $\bar{x}_u$ and is given by

$$\omega_u(\theta) = \prod_{m=1}^{M} p_1(x_u^1(m)|\cdot)p_2(x_u^2(m)|\cdot)p_3(x_u^3(m)|\cdot).$$

Define $\omega_u^r(\theta) = \frac{\partial}{\partial\theta_r}\omega_u(\theta)$. The (r,s) entry of the $\nu \times \nu$ matrix $\frac{1}{n}J(\theta)$ is now

$$\frac{1}{n}\sum_{u=1}^{n} E_\theta\left(\psi_u^{rs}(\theta)\middle|(1,n)\right) - \frac{1}{n}\sum_{u,v=1}^{n} V_\theta\left(\psi_u^s(\theta), \omega_v^r(\theta)\middle|(1,n)\right).$$

We first show uniform convergence with respect to θ.

As above, let $P_\theta(\cdot|(m_1,m_2))$ be the conditional distribution given $x_{i_1:i_2}(M)$ and let $P_\theta(\cdot|[i_1,i_2])$ be the conditional distribution given $x_{i_1:i_2}(M)$ as well as x_{i_1} and x_{i_2}. If $i_1 < 1$ the conditioning is with respect to $(x_{1:i_2}(M), x_{i_2})$ only, and similarly If $i_2 > n$ the conditioning is with respect to $(x_{i_1:n}(M), x_{i_1})$ only. The corresponding changes in the derivations below are not spelled out. Since ω and its derivatives are continuous and the state space is finite, one trivially has the existence of a constant c_2 such that for all i

$$|\omega_i^r| \le c_2 \quad \text{for } |\theta - \theta_0| \le \delta_0. \tag{58}$$

Finally, let ν be the dimension of θ.

Lemma 5. *Let h^u be a function of $\bar{x}_u$ with $|h^u| \le 1$. For $|\theta - \theta_0| \le \delta_0$ we have*

$$|E_\theta(h^u|[i_1,i_2]) - E_{\theta_0}(h^u|[i_1,i_2])| \le 2c_2(i_2 - i_1 + 1)\nu|\theta - \theta_0|.$$

Proof. Let p_θ^u be the density of $P_\theta(\bar{x}_u \in \cdot|[i_1,i_2])$. We first obtain a bound on the derivative of p_θ^u. To this end write

$$p_\theta^u = \frac{\sum^{(1)}\prod_{i=i_1}^{i_2}\omega_i}{\sum^{(2)}\prod_{i=i_1}^{i_2}\omega_i},$$

where $\sum^{(1)}$ is the sum over the possible values of $(x_{(i_1+1):(u-2)}, x_{(u+2):(i_2-1)})$ and $\sum^{(2)}$ is the sum over the possible values of $(x_{(i_1+1):(i_2-1)})$. The derivative is then

$$\begin{aligned}\left|\frac{\partial p_\theta^u}{\partial\theta_r}\right| &= \left|\frac{\sum^{(1)}\sum_{j=i_1}^{i_2}\omega_j^r\prod_{i=i_1}^{i_2}\omega_i}{\sum^{(2)}\prod_{i=i_1}^{i_2}\omega_i} - \frac{\sum^{(1)}\prod_{i=i_1}^{i_2}\omega_i}{\left(\sum^{(2)}\prod_{i=i_1}^{i_2}\omega_i\right)^2}\left(\sum^{(2)}\sum_{j=i_1}^{i_2}\omega_j^r\prod_{i=i_1}^{i_2}\omega_i\right)\right| \\ &\le 2c_2(i_2 - i_1 + 1)p_\theta^u.\end{aligned}$$

From this bound one finds

$$\begin{aligned}
|E_\theta(h^u|[i_1,i_2]) - E_{\theta_0}(h^u|[i_1,i_2])| &= \left|\int_0^1 \sum_{\bar{x}_u} h^u \frac{d}{dt} p^u_{\theta_0+t(\theta-\theta_0)} \mathrm{dt}\right| \\
&= \left|\int_0^1 \sum_{\bar{x}_u} h^u \sum_{r=1}^{d} (\theta-\theta_0)_r \frac{\partial}{\partial\theta_r} p^u_{\theta_0+t(\theta-\theta_0)} \mathrm{dt}\right| \\
&\le \nu|\theta-\theta_0| 2c_2(i_2-i_1+1) \int_0^1 \sum_{\bar{x}_u} p^u_{\theta_0+t(\theta-\theta_0)} \mathrm{dt} \\
&= \nu|\theta-\theta_0| 2c_2(i_2-i_1+1),
\end{aligned}$$

which proves the lemma. □

Lemma 6. *Let h^u be a function of $\bar{x}_u$ with $|h^u| \le 1$. For $|\theta-\theta_0| \le \delta_0$ and for any integer l we have*

$$|E_\theta(h^u|(1,n)) - E_{\theta_0}(h^u|(1,n))| \le 8\rho^{l-1} + 2c_2(2l+1)\nu|\theta-\theta_0|.$$

Proof. From the mixing (53) (conditioning on $x(0)$ and $x(M)$) and arguing as in (57) the bound

$$\begin{aligned}
&|E_\theta(h^u|(1,n)) - E_\theta(h^u|[u-l,u+l])| \\
&\quad = \left|\int E_\theta(h^u|[u-l,u+l]))P_\theta(d(x_{u-l},x_{u+l})|(1,n)) - E_\theta(h^u|[u-l,u+l])\right| \\
&\quad \le 2(\rho^{l-1}+\rho^{l-1}),
\end{aligned}$$

is obtained,which is valid for all $|\theta-\theta_0| \le \delta_0$. Using this and Lemma 5 we obtain the bound

$$\begin{aligned}
|E_\theta(h^u|(1,n)) - E_{\theta_0}(h^u|(1,n))| \le\ & |E_\theta(h^u|(1,n)) - E_\theta(h^u|([u-l,u+l]))| \\
&+ |E_\theta(h^u|[u-l,u+l]) - E_{\theta_0}(h^u|[u-l,u+l])| \\
&+ |E_{\theta_0}(h^u|[u-l,u+l]) - E_{\theta_0}(h^u|(1,n))| \\
&\le 4\rho^{l-1} + 2c_2(2l+1)\nu|\theta-\theta_0| + 4\rho^{l-1},
\end{aligned}$$

which gives the result of the lemma. □

From the continuity of ψ and its derivatives, and the finiteness of the state space, there exist, trivially, bounds c_3, c_4 such that for all i and all $|\theta-\theta_0| \le \delta_0$

$$|\psi^{rs}_u(\theta)| \le c_3 \quad \text{and} \quad |\psi^{rs}_u(\theta) - \psi^{rs}_u(\theta_0)| \le c_4|\theta-\theta_0|. \tag{59}$$

Lemma 7. *Uniform convergence of conditional average:*

$$\lim_{\delta\to 0} \sup_n \sup_{|\theta-\theta_0|\le\delta} \left|\frac{1}{n}\sum_{u=1}^{n} E_\theta\left(\psi^{rs}_u(\theta)\middle|(1,n)\right) - \frac{1}{n}\sum_{u=1}^{n} E_{\theta_0}\left(\psi^{rs}_u(\theta_0)\middle|(1,n)\right)\right| = 0.$$

Proof. From (59)) it follows that

$$\left|\frac{1}{n}\sum_{u=1}^{n} E_\theta\left([\psi_u^{rs}(\theta)-\psi_u^{rs}(\theta_0)]\big|(1,n)\right)\right| \le c_4|\theta-\theta_0|, \tag{60}$$

and from Lemma 6 one sees that

$$\left|\frac{1}{n}\sum_{u=1}^{n}\left[E_\theta\left(\psi_u^{rs}(\theta_0)\big|(1,n)\right)-E_{\theta_0}\left(\psi_u^{rs}(\theta_0)\big|(1,n)\right)\right]\right| \le 8\rho^{l-1}+2c_2(2l+1)\nu|\theta-\theta_0|. \tag{61}$$

If for example we take $l=\delta^{-1/2}$, the sum of the two terms (60) and (61) is of order $\delta^{1/2}$ when $|\theta-\theta_0|\le\delta$ and, thus, the result of the lemma follows. □

Above we introduced c_1 and c_2 such that for all u and all $|\theta-\theta_0|\le\delta_0$

$$|\psi_u^s(\theta)|\le c_1 \qquad \text{and} \qquad |\omega_u^r(\theta)|\le c_2.$$

Similarly, from continuity and the finiteness of the state space there exist constants c_5 and c_6 such that

$$|\psi_u^s(\theta)-\psi_u^s(\theta_0)|\le c_5|\theta-\theta_0| \ \text{ and } \ |\omega_u^r(\theta)-\omega_u^r(\theta_0)|\le c_6|\theta-\theta_0|. \tag{62}$$

Lemma 8. *There exists a constant c_7 such that for $|\theta-\theta_0|\le\delta_0$ and for any integer l we have*

$$\begin{aligned}&\left|V_\theta\big(\psi_u^s(\theta),l_v^r(\theta)\big|(1,n)\big)-V_{\theta_0}\big(\psi_u^s(\theta_0),l_v^r(\theta_0)\big|(1,n)\big)\right|\\ &\qquad\le c_7\left\{\rho^{l-1}+(1+l+|u-v|)|\theta-\theta_0|\right\}.\end{aligned}$$

Proof. From (56), (58), and (62) one finds

$$\begin{aligned}&\left|V_\theta\big(\psi_u^s(\theta),\omega_v^r(\theta)\big|(1,n)\big)-V_\theta\big(\psi_u^s(\theta_0),\omega_v^r(\theta_0)\big|(1,n)\big)\right|\\ &\quad=\left|V_\theta\big(\psi_u^s(\theta)-\psi_u^s(\theta_0),\omega_v^r(\theta)\big|(1,n)\big)+V_\theta\big(\psi_u^s(\theta_0),\omega_v^r(\theta)-\omega_v^r(\theta_0)\big|(1,n)\big)\right|\\ &\quad\le 2(c_5c_2+c_1c_6)|\theta-\theta_0|.\end{aligned}$$

Next, from Lemma 6 it follows that

$$|E_\theta(\psi_u^s(\theta_0)|(1,n))-E_{\theta_0}(\psi_u^s(\theta_0)|(1,n))|\le c_1(8\rho^{l-1}+2c_2(2l+1)\nu|\theta-\theta_0|),$$

and

$$|E_\theta(\omega_v^r(\theta_0)|(1,n))-E_{\theta_0}(\omega_v^r(\theta_0)|(1,n))|\le c_2(8\rho^{l-1}+2c_2(2l+1)\nu|\theta-\theta_0|),$$

which together give

$$\begin{aligned}&|E_\theta(\psi_u^s(\theta_0)|(1,n))E_\theta(\omega_v^r(\theta_0)|(1,n))-E_{\theta_0}(\psi_u^s(\theta_0)|(1,n))E_{\theta_0}(\omega_v^r(\theta_0)|(1,n))|\\ &\quad\le 2c_1c_2(8\rho^{l-1}+2c_2(2l+1)\nu|\theta-\theta_0|).\end{aligned}$$

Proceeding as in the proof of Lemma 6, conditioning on $x_{(u-l,v+l)}$ (for the case $v > u$) instead of $x_{(u-l,u+l)}$, we find

$$\begin{aligned}|E_\theta(\psi_u^s(\theta_0)\omega_v^r(\theta_0)|(1,n)) - E_{\theta_0}(\psi_u^s(\theta_0)\omega_v^r(\theta_0)|(1,n))| \\ \leq c_1c_2\left\{8\rho^{l-1} + 2c_2(2l + |u-v| + 1)\nu|\theta - \theta_0|\right\}.\end{aligned}$$

The bound stated in the lemma now follows on combining the above error terms. □

Lemma 9. *Uniform convergence of conditional covariance:*

$$\lim_{\delta\to 0}\sup_n \sup_{|\theta-\theta_0|\leq\delta}\left|\frac{1}{n}\sum_{u,v=1}^n V_\theta\big(\psi_u^s(\theta),\omega_v^r(\theta)\big|(1,n)\big) - \frac{1}{n}\sum_{u,v=1}^n V_{\theta_0}\big(\psi_u^s(\theta_0)\omega_v^r(\theta_0)\big|(1,n)\big)\right| = 0.$$

Proof. In the proof we skip the indices r and s in ψ_u^s and l_u^r. The mixing (53) (conditioning on $x(0)$ and $x(M)$) gives that

$$\big|V_\theta\big(\psi_u^s(\theta),\omega_v^r(\theta)\big|(1,n)\big)\big| \leq 4c_1c_2\rho^{|u-v|-2}, \tag{63}$$

(Ibragimov and Linnik (1971)). Together with Lemma 8 this gives

$$\begin{aligned}&\left|\sum_{v=1}^n V_\theta\big(\psi_u(\theta),\omega_v^r(\theta)\big|(1,n)\big) - V_{\theta_0}\big(\psi_u^s(\theta_0)\omega_v^r(\theta_0)\big|(1,n)\big)\right| \\ &\quad\leq \sum_{v:|v-u|>l} 4c_1c_2\rho^{|u-v|-2} + \sum_{v:|v-u|\leq l} c_7\left\{\rho^{l-1} + (1+l+|u-v|)|\theta-\theta_0|\right\} \\ &\quad\leq c_{11}l\left\{\rho^{l-1} + l|\theta-\theta_0|\right\}.\end{aligned}$$

Taking $l = \delta^{-1/4}$, the latter bound tends to zero and the proof of the lemma is completed.

□

8.4 Convergence of the "observed information"

When considering the limit as $n \to \infty$ of the "observed information"

$$-\partial E_\theta(\Psi\big(\theta)|x(M))/\partial\theta$$

we no longer condition on $x(0)$, that is, the stationary density (32) is included in the full likelihood (45). In this case the underlying Markov chain is homogeneous. This homogeneous Markov chain has positive probability for all the possible transitions and, so, has a stationary initial distribution. We can, therefore, extend the Markov chain into a stationary process from $-\infty$ to $+\infty$. The probability measure for $x_1, \ldots, x_n$ then corresponds to a segment of the stationary process, but where the joint distribution of (x_1, x_n) is different from that derived from the stationary process. Probabilities from the stationary process are indicated by a bar above the appropiate quantities.

For convenience the argument θ_0 is not displayed in the following. For the stationary process let $\bar{P}(\cdot|(-\infty,\infty))$ be the measure conditioned on $(x_i(0), x_i(M))$, with $-\infty < i < \infty$.

Lemma 10. *As $n \to \infty$ we have the following convergence in probability*

$$\frac{1}{n}\sum_{u=1}^{n} E_{\theta_0}\left(\psi_u^{rs}(\theta_0)\middle|(1,n)\right) \to \bar{E}_{\theta_0}\left(\psi_0^{rs}(\theta_0)\right)$$

Proof. We first prove that the variance of the average in the lemma tends to zero. Using an argument similar to (57) one has

$$E(\psi_u^{rs}|(1,n)) = E(\psi_u^{rs}|(u-l,u+l)) + O(\rho^l),$$

and this gives

$$\begin{aligned}&\bar{V}\left(E(\psi_u^{rs}|(1,n)), E(\psi_v^{rs}|(1,n))\right)\\ &\quad = \bar{V}\left(E(\psi_u^{rs}|(u-l,u+l)), E(\psi_v^{rs}|(v-l,v+l))\right) + O(\rho^l).\end{aligned}$$

Also, the mixing of the stationary process implies a covariance bound as in (63), and this gives

$$\bar{V}\left(E(\psi_u^{rs}|(u-l,u+l)), E(\psi_v^{rs}|(v-l,v+l))\right) = O\left(\rho^{\max(0,|v-u|-2l)}\right).$$

Taking $l = |v-u|/4$ and combining the above two expressions give

$$\bar{V}\left(E(\psi_u^{rs}|(1,n)), E(\psi_v^{rs}|(1,n))\right) = O\left(\rho^{|u-v|/4}\right),$$

which implies that

$$\bar{V}\left(\frac{1}{n}\sum_{u=1}^{n} E(\psi_u^{rs}|(1,n))\right) = O\left(\frac{1}{n}\right).$$

Thus, it suffices to study the limiting behaviour of the average in the lemma. The latter mean is

$$\frac{1}{n}\sum_{u=1}^{n} \bar{E}\left(E(\psi_u^{rs}|(1,n))\right).$$

Using the mixing properties (53) (conditioning on $x(0)$ and $x(M)$), and an argument as in (57), it follows that

$$\begin{aligned}\left|E(\psi_u^{rs}|(1,n)) - \bar{E}(\psi_u^{rs}|(-\infty,\infty))\right| &= \left|\int E(\psi_u^{rs}|[1,n])P\left(d(x_1,x_n)|(1,n)\right)\right.\\ &\qquad \left. - \int E(\psi_u^{rs}|[1,n])\bar{P}\left(d(x_1,x_n)|(-\infty,\infty)\right)\right|\\ &\le 2c_3(\rho^{u-2} + \rho^{n-u-1}).\end{aligned}$$

Since $\bar{E}\left(\bar{E}(\psi_u^{rs}|(-\infty,\infty))\right) = \bar{E}(\psi_u^{rs})$ the latter bound implies

$$\begin{aligned}\left|\frac{1}{n}\sum_{u=1}^{n} \bar{E}\left(E(\psi_u^{rs}|(1,n))\right) - \bar{E}(\psi_0^{rs})\right| &\le \frac{1}{n}\sum_{u=1}^{n} 2c_3(\rho^{u-2} + \rho^{n-u-1})\\ &= \frac{2c_3}{n}\frac{2}{\rho(1-\rho)},\end{aligned}$$

which clearly tends to zero as $n \to \infty$. Thus the lemma has been proved. □

The limit of the covariance part of (29) is somewhat more difficult to obtain.

Lemma 11. *We have*

$$\left|V\left(\psi_u^s,\omega_v^r\middle|(1,n)\right)-\bar{V}\left(\psi_u^s,\omega_v^r\middle|(-\infty,\infty)\right)\right|\leq 3c_1c_2\left(\rho^{\min\{u,v\}-2}+\rho^{n-\max\{u,v\}-1}\right).$$

Proof. Write $E(\cdot|(1,n)) = \int E(\cdot|[1,n])P(d(x_1,x_n)|(1,n))$ and $\bar{E}(\cdot|(-\infty,\infty)) = \int E(\cdot|[1,n])\bar{P}(d(x_1,x_n)|(-\infty,\infty))$. From the mixing (53) (conditioning on $x(0)$ and $x(M)$) it is seen that

$$\begin{aligned}\left|E\left(\psi_u^s\omega_v^r\middle|(1,n)\right)-\bar{E}\left(\psi_u^s\omega_v^r\middle|(-\infty,\infty)\right)\right|&\leq c_1c_2\left(\rho^{\min\{u,v\}-2}+\rho^{n-\max\{u,v\}-1}\right),\\ \left|E\left(\psi_u^s\middle|(1,n)\right)-\bar{E}\left(\psi_u^s\middle|(-\infty,\infty)\right)\right|&\leq c_1\left(\rho^{u-2}+\rho^{n-u-1}\right),\\ \left|E\left(\omega_v^r\middle|(1,n)\right)-\bar{E}\left(\omega_v^r\middle|(-\infty,\infty)\right)\right|&\leq c_2\left(\rho^{v-2}+\rho^{n-v-1}\right)\end{aligned}$$

Combining these three bounds the result of the lemma follows. □

Lemma 12. *As $n\to\infty$ we have*

$$\left|\frac{1}{n}\sum_{u,v=1}^{n}V\left(\psi_u^s,\omega_v^r\middle|(1,n)\right)-\frac{1}{n}\sum_{u=1}^{n}\sum_{v=-\infty}^{\infty}\bar{V}\left(\psi_u^s,\omega_v^r\middle|(-\infty,\infty)\right)\right|\to 0.$$

Proof. We write V_{uv} for $V\left(\psi_u^s,\omega_v^r\middle|(1,n)\right)$ and $\bar{V}_{uv}$ for $\bar{V}\left(\psi_u^s,\omega_v^r\middle|(-\infty,\infty)\right)$ and define $J(u)=\{1\leq v\leq n : |u-v|\leq n^\alpha\}$. Using the mixing bound (63) and Lemma 11 it follows that

$$\begin{aligned}\frac{1}{n}\sum_{u,v=1}^{n}V_{uv}&=\frac{1}{n}\sum_{u=1}^{n}\sum_{v\in J(u)}V_{uv}+O\left(\sum_{j>n^\alpha}\rho^j\right)\\ &=\frac{1}{n}\sum_{u=1}^{n}\sum_{v\in J(u)}\bar{V}_{uv}+O\left(\frac{1}{n}\sum_{u=1}^{n}\sum_{v\in J(u)}\rho^{\min\{u,v\}}+\rho^{n-\max\{u,v\}}\right)+O\left(\rho^{n^\alpha}\right)\\ &=\frac{1}{n}\sum_{u=1}^{n}\sum_{v\in J(u)}\bar{V}_{uv}+O\left(\frac{n^\alpha}{n}\right)+O\left(\rho^{n^\alpha}\right)\\ &=\frac{1}{n}\sum_{u=1}^{n}\sum_{v=-\infty}^{\infty}\bar{V}_{uv}+O\left(\frac{1}{n}\sum_{u=1}^{n}\sum_{j=\min\{u,n^\alpha\}}^{\infty}\rho^j\right)+O\left(\frac{n^\alpha}{n}+\rho^{n^\alpha}\right)\\ &=\frac{1}{n}\sum_{u=1}^{n}\sum_{v=-\infty}^{\infty}\bar{V}_{uv}+O\left(\rho^{n^\alpha}\frac{n^\alpha}{n}+\rho^{n^\alpha}\right).\end{aligned}$$

Taking $\alpha=\frac{1}{2}$ the $O(\cdot)$ term tends to zero and the lemma has been proved. □

Lemma 13. *Under $\bar{P}$ we have*

$$\frac{1}{n}\sum_{u=1}^{n}\sum_{v=-\infty}^{\infty}\bar{V}\left(\psi_u^s,\omega_v^r\middle|(-\infty,\infty)\right)\to\sum_{v=-\infty}^{\infty}\bar{E}\left[\bar{V}\left(\psi_0^s,\omega_v^r\middle|(-\infty,\infty)\right)\right]$$

as $n\to\infty$.

Proof. Since the underlying Markov process x_i under $\bar{P}$ is ergodic, the sum $w_u=\sum_{v=-\infty}^{\infty}\bar{V}\left(\psi_u^s,\omega_v^r\middle|(-\infty,\infty)\right)$ is ergodic. The ergodic theorem, therefore, gives that $\frac{1}{n}\sum_{u=1}^{n}w_u\to\bar{E}(w_0)$. □

9 Concluding remarks

The time discretized model presented in this paper can be used for a moderately sized phylogenetic tree and for moderately sized sequence lengths. The mixing of the process along the sequence is fast making the Gibbs sampler a feasible tool for calculating mean values. This makes the EM algorithm (or the EEE algorithm) a very useful tool for obtaining parameter estimates. In this respect the method of this paper is an improvement over the estimation method in Jensen and Pedersen (2000). Furthermore, asymptotic normality of the estimates has been rigorously derived.

The actual model used in the data example is not crucial. Many other models can be used within the framework given here. Time reversibility is used in a crucial way for obtaining a simple stationary measure. However, the time reversibility is not used in the MCMC analysis, and the latter can, therefore, be used for any context dependent model. We must then have a rooted phylogenetic tree, and unless the root is observed we need a stochastic model for the root sequence. For simplicity a Markov chain along the root sequence seems appropriate, and with such a model the asymptotic results are still valid.

A Derivatives of $\exp(tQ)$

We consider the situation in Subsection 2.1 with a continuous time Markov process given through the rates q_{ij}. Let the rates $q_{ij} = q_{ij}(\theta)$ be functions of a scalar parameter θ and make the assumption that $Q = SDS^{-1}$, where D is the diagonal matrix of eigenvalues. Define $G = S^{-1}\frac{\partial Q}{\partial \theta}S$ and define a matrix F with entries F_{ij} equal to $t\exp(tD_{ii})$ if $D_{ii} = D_{jj}$, and equal to $(\exp(tD_{jj}) - \exp(tD_{ii}))/(D_{jj} - D_{ii})$ otherwise.

In Kalbfleisch and Lawless (1985) the following derivation is given:

$$\begin{aligned}
\frac{\partial P(t)}{\partial \theta} &= \sum_{n=1}^{\infty}\sum_{l=1}^{n}\frac{t^n}{n!}Q^{l-1}\frac{\partial Q}{\partial \theta}Q^{n-l} = S\left[\sum_{n=1}^{\infty}\sum_{l=1}^{n}\frac{t^n}{n!}D^{l-1}GD^{n-l}\right]S^{-1} \\
&= S\left[\sum_{ab}G_{ab}\sum_{n=1}^{\infty}\sum_{l=1}^{n}\frac{t^n}{n!}D_{aa}^{l-1}D_{bb}^{n-l}I^{ab}\right]S^{-1} \\
&= S\left[\sum_{ab}G_{ab}F_{ab}I^{ab}\right]S^{-1} \\
&= S[G \circ F]S^{-1},
\end{aligned}$$

where I_{ab} is the matrix with (a,b)th entry equal to one and all other entries being zero, and where $G \circ F$ is the matrix with entries $G_{ab}F_{ab}$.

B Eigenvalues and eigenvectors for nucleotide models

The *HKY* model (Hasegawa et al. (1985)) has rate matrix (2). Recall that $\pi_{AG} = \pi_A + \pi_G$ and $\pi_{CT} = \pi_C + \pi_T$. The eigenvalues and eigenvetors of this rate matrix are

	eigenvalues			
	0	$-\beta$	$-\beta$	$-(\alpha\pi_{AG} + \beta\pi_{CT})$
A	1	1	$-\pi_{CT}/\pi_{AG}$	1
G	1	1	$-\pi_{CT}/\pi_{AG}$	$-\pi_A/\pi_G$
C	1	$-\pi_{AG}/\pi_{CT}$	1	0
T	1	$-\pi_{AG}/\pi_{CT}$	1	0

where the eigenvectors are below the horizontal line.

For the general strand symmetric model with rate matrix (3) we divide the discussion of eigenvalues and eigenvectors into a number of different cases. Define

$$a = (\alpha + \beta + 2\delta) - (\gamma + \omega + 2\kappa) \quad \text{and} \quad b = a^2 + 4(\alpha - \beta)(\gamma - \omega).$$

When $\alpha \neq \beta$ and $b \neq 0$ the eigenvalues and eigenvectors are

	eigenvalues			
	0	$-(\gamma+\omega+\alpha+\beta)$	$\frac{a+\sqrt{b}}{2}-(\alpha+\beta+2\delta)$	$\frac{a-\sqrt{b}}{2}-(\alpha+\beta+2\delta)$
A	1	1	1	1
G	1	$-\frac{\gamma+\omega}{\alpha+\beta}$	$\frac{a+\sqrt{b}}{2(\alpha-\beta)}$	$\frac{a-\sqrt{b}}{2(\alpha-\beta)}$
C	1	$-\frac{\gamma+\omega}{\alpha+\beta}$	$-\frac{a+\sqrt{b}}{2(\alpha-\beta)}$	$-\frac{a-\sqrt{b}}{2(\alpha-\beta)}$
T	1	1	-1	-1

with the eigenvectors below the horizontal line. When $\alpha = \beta$ and $b > 0$ the two last eigenvalues become $-(\gamma+\omega+2\kappa)$ and $-(\alpha+\beta+2\delta)$, with corresponding eigenvectors $(0,1,-1,0)$ and $(1,c,-c,-1)$, where $c = (\omega-\gamma)/(\alpha+\beta+2\delta-(\gamma+\omega+2\kappa)$. Finally, when $b = 0$ the eigenvectors no longer span the 4 dimensional space and (1) cannot be used. Let $\lambda_1 = 0$, $\lambda_2 = -(\gamma+\omega+\alpha+\beta)$, and $\lambda_3 = \frac{a+\sqrt{b}}{2} - (\alpha+\beta+2\delta)$, where the last expression reduces to $\lambda_3 = -(\gamma+\omega+2\kappa)$ when $\alpha = \beta$. Instead of (1), one finds that

$$\exp(tQ) = S \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \mathrm{e}^{t\lambda_2} & 0 & 0 \\ 0 & 0 & \mathrm{e}^{t\lambda_3} & t\xi\mathrm{e}^{t\lambda_2} \\ 0 & 0 & 0 & \mathrm{e}^{t\lambda_3} \end{pmatrix},$$

where $\xi = \alpha - \beta$ when $\alpha \neq \beta$ and $\xi = \gamma - \omega$ when $\alpha = \beta$. The matrix S has its first three columns equal to the eigenvectors from before, and the fourth column is $(0,1,-1,0)$ when $\alpha \neq \beta$ and equal to $(1,1,-1,-1)$ when $\alpha = \beta$.

C Calculations in the model of Arndt, Burge, and Hwa (2003)

Consider a double infinitely long sequence evolving according to the rates (4). Let $f_{ab}(t) = P(x_1(t) = a, x_2(t) = b)$ be the probability at time t. The Kolmogorov forward differential equations can, in the notation used here, be written as

$$f'_{ab}(t) = \sum_{x_0,x_3} \Bigg\{ \sum_{c\neq a} \lambda(a|x_0,c,b) f_{x_0cbx_3}(t) + \sum_{d\neq b} \lambda(b|a,d,x_3) f_{x_0adx_3}(t) - \Bigg(\sum_{c\neq a} \lambda(c|x_0,a,b) + \sum_{d\neq b} \lambda(d|a,b,x_3) \Bigg) f_{x_0abx_3}(t) \Bigg\}$$

$$
\begin{aligned}
&= \sum_{c\neq a}\Big\{[\lambda_0(a|c)+\lambda_r(a|cb)]f_{cb}(t)+\sum_{x_0}\lambda_l(a|x_0,c)f_{x_0cb}(t)\Big\}\\
&\quad+\sum_{d\neq b}\Big\{[\lambda_0(b|d)+\lambda_l(b|ad)]f_{ad}(t)++\sum_{x_3}\lambda_r(b|dx_3)f_{adx_3}(t)\Big\}\\
&\quad-\Big[\lambda_0(|a)+\lambda_0(|b)+\lambda_r(|ab)+\lambda_l(|ab)\Big]f_{ab}(t)\\
&\quad-\Big\{\sum_{x_0}\lambda_l(|x_0,a)f_{x_0ab}(t)+\sum_{x_3}\lambda_r(t|bx_3)f_{abx_3}(t),\Big\},
\end{aligned}
\tag{64}
$$

where the notation with a λ without its first argument implies a summation over the rates, and f with three indices is the probability of three consequtive nucleotides $f_{abc}(t) = P(x_1(t) = a, x_2(t) = b, x_3(t) = c)$. Letting $t \to \infty$, the left hand side becomes zero and on the right hand side f can be replaced by the stationary probabilities π. To solve these equations Arndt et al. (2003) use the approximation in (5), which is exact if the stationary measure along the sequence is a Markov chain.

We next consider a characterization of the cases where the stationary measure is a Markov chain along the sequence. Consider a proces $(x_1, \ldots, x_n)$, where the rates are given by (4), and where there are fixed values x_0 and x_{n+1} that define the rates at the two ends of the sequence. We look for cases where the stationary measure φ can be written in the form

$$
\varphi(x) = \phi_1(x_1)\phi_2(x_n)\prod_{i=2}^{n}\phi(x_i|x_{i-1}),
$$

where $\phi(a|b)$ is the transition probability of a Markov chain. Also, we have in mind $n \to \infty$ so that factors related to ϕ_1 and ϕ_2 are not of importance. The equation that φ has to satisfy, in order to be the stationary measure, is $\sum_x \varphi(x)\lambda_{xy} = 0$ for all sequences y, where λ_{xy} is the rate for a change of x to y. In our case there can be a change at one position in the sequence only, at each time point. The equation then becomes

$$
\begin{aligned}
0 &= \varphi(y)\sum_{i=1}^{n}\sum_{\nu\neq y_i}\lambda(y_i|y_{i-1},\nu,y_{i+1})\frac{\varphi(\nu|y_{i-1})\varphi(y_{i+1}|\nu)}{\varphi(y_i|y_{i-1})\varphi(y_{i+1}|y_i)} \qquad (65)\\
&\quad-\varphi(y)\sum_{i=1}^{n}\sum_{\nu\neq y_i}\lambda(\nu|y_{i-1},y_i,y_{i+1})\\
&= \varphi(y)\left\{h_1(y_0,y_1,y_2)+\sum_{i=2}^{n-1}h(y_{i-1},y_i,y_{i+1})+h_2(y_{n-1},y_n,y_{n+1})\right\}\\
&= \varphi(y)\left\{\sum_{a,b,c}N_y(a,b,c)h(a,b,c)+h_1(y_0,y_1,y_2)+h_2(y_{n-1},y_n,y_{n+1})\right\},
\end{aligned}
$$

where $\phi(y_1|y_0)$ and $\phi(y_{n+1}|y_n)$ should be replaced by $\phi_1(y_1)$ and $\phi_2(y_n)$, respectively. In the last equation $N_y(a, b, c)$ is the number of triplets (y_{i-1}, y_i, y_{i+1}), $i = 2, \ldots, n-1$, with the value (a, b, c), and

$$
h(a,b,c) \;=\; \sum_{\nu\neq b}\lambda(b|a,\nu,c)\frac{\phi(\nu|a)\phi(c|\nu)}{\phi(b|a)\phi(c|b)}-\sum_{\nu\neq b}\lambda(\nu|a,b,c).
$$

The terms h_1 and h_2 in (65) are not of importance when $n \to \infty$. It is tempting to think that having (65) equal to zero for all sequences y implies that $h(a,b,c) = 0$ for all combinations of (a,b,c). However, as the special case below shows, this need not be true.

To obtain explicit results we now specialize to the case of a two letter alphabet, that is, $x_i \in \{A, B\}$. Using the notation $\neg A = B$ and $\neg B = A$, the model (4) has a total of nine parameters:

$$\lambda_0(B|A) = 1, \quad \lambda_0(A|B) = \gamma, \quad \lambda_l(\neg b|ab) = \kappa_{ab}, \quad \lambda_r(\neg a|ab) = \omega_{ab},$$

where $a, b \in \{A, B\}$. Note that this is by itself an overparametrization: there are only eight different rates $\lambda(\neg b|abc)$. If $\kappa_{a_1a_2}$ is the smallest of the κ values and, similarly, $\omega_{b_1b_2}$ is the smallest of the ω values, then these two terms cannot be identified individually, only their sum $\kappa_{a_1a_2} + \omega_{b_1b_2}$ can be identified. Furthermore, we parametrize the Markov chain φ by

$$\phi(B|A) = \alpha, \quad \phi(A|B) = \beta. \tag{66}$$

This Markov chain has stationary probabilities $\pi_A = \beta/(\alpha+\beta)$ and $\pi_B = \alpha/(\alpha+\beta)$. Manipulating the equations (64), together with the approximation (5), give the following equations for α and β

$$(\xi_{AB} - \xi_{BB})\beta^2 + (\xi_{BA} - \xi_{AA})\alpha\beta + (\xi_{AA} + \xi_{BA} - \xi_{AB} + 3\xi_{BB})\beta = 2\xi_{BB}, \tag{67}$$

$$\xi_{AA}\beta - \xi_{BB}\alpha = (\xi_{AA} + \xi_{AB} - \xi_{BA} - \xi_{BB})\alpha\beta, \tag{68}$$

where

$$\xi_{AA} = 1 + \kappa_{AA} + \omega_{AA} = \lambda(B|AAA), \qquad \xi_{BB} = \gamma + \kappa_{BB} + \omega_{BB} = \lambda(A|BBB),$$
$$\xi_{AB} = \gamma + \kappa_{AB} + \omega_{BA} = \lambda(A|ABA), \qquad \xi_{BA} = 1 + \kappa_{BA} + \omega_{AB} = \lambda(B|BAB).$$

Equation (67) is derived from (64) with $(ab) = (BB)$, and equation (68) is the difference between (67) and the equation derived from (64) with $(ab) = (AA)$. The equations in (67) and (68) show how some of the rates do not enter the Markov chain approximation to the stationary measure.

To study when the stationary measure has the Markov structure, we look at the equations in (65) using the parametrization (66) of the Markov chain. Taking the y sequence to consist entirely of As or entirely of Bs, we find $h(AAA) = 0$ and $h(BBB) = 0$. Writing down the expressions for the h functions, one sees that $h(AAA) = 0$ implies $h(ABA) = 0$ and $h(BBB) = 0$ implies $h(BAB) = 0$. Next, we consider a y sequence where AAB is repeated, which gives $h(AAB) + h(BAA) = 0$ (using $h(ABA) = 0$). Again, writing down the equations one finds that $h(AAB) + h(BAA) = 0$ implies $h(ABB) + h(BBA) = 0$. The above 6 equations for the h functions suffice for (65) to be satisfied for all y in the limit $n \to \infty$. The argument for this runs as follows. Starting from one end of the y sequence, the h terms cancel with one another so that at most two terms are left. To understand this we consider a number of cases. Consider a subsequence starting with ab and consisting of $abb^k(\neg b)$ with $k \geq 0$. The sum of the h functions is $h(ab(\neg b)$ when $k = 0$ and $h(abb) + (k-1)h(bbb) + h(bb(\neg b)) = h(abb) + h(bb(\neg b))$ when $k > 0$. In both cases

there are at most one nonzero h term along the sequence. Furthemore, all these subsequences end with the letters $b(\neg b)$ and if $ab = AB$ or $ab = BA$ the sum of the h functions is zero. Thus, for the full sequence, irrespectively of the value of the two first letters, there is at most two nonzero terms in the cumulative sum of the h functions. Let us now consider the three equations $h(AAA) = 0$, $h(BBB) = 0$, and $h(AAB) + h(BAA) = 0$ in more detail. The three equations take the form

$$\xi_{11}(1-\alpha)^2 = \xi_{12}\alpha\beta, \qquad \xi_{22}(1-\beta)^2 = \xi_{21}\alpha\beta,$$
$$(\xi_{11}+\xi_{21})(1-\alpha) = (\xi_{22}+\xi_{12})(1-\beta).$$

Manipulations of these equations give

$$\alpha = \frac{\xi_{11}}{\xi_{11}+\xi_{12}}, \qquad \beta = \frac{\xi_{12}}{\xi_{11}+\xi_{12}}, \tag{69}$$
$$\xi_{11}\xi_{22} = \xi_{12}\xi_{21}. \tag{70}$$

With these values of α and β and using (70), equations (67) and (68) are both fulfilled. In summary, if and only if (70) is satisfied the stationary measure is a Markov chain with transition probabilities given by (69). The stationary process being a Markov chain does not automatically imply that the process is time reversible. We need one more restriction on the parameters apart from (70), namely

$$\xi_{AA}(\gamma + \kappa_{AB} + \omega_{BB}) = \xi_{AB}(1 + \kappa_{AA} + \omega_{AB}).$$

D Estimation based on stationary distribution

Estimation of the nucleotide frequencies $D_j(\nu)$ from (31) and the CG-depression λ from (43) through maximization of (44) is difficult due to the normalizing constant C. In Jensen and Pedersen (2000) a fairly accurate approximation to C is given. To make the maximization easier we estimate, for fixed λ, $D_j(\cdot)$ from the conditional distribution of the nucleotide at position j given the values of the flanking nucleotides. For position 1 there are a total of 6 different conditional distributions and the likelihood becomes

$$\begin{aligned}\left(\prod_{\nu=A}^{T} D_1(\nu)^{N(1,\nu)}\right)&[D_1(A)+\lambda^2 D_1(G)+\lambda^2 D_1(C)+D_1(T)]^{-K(1,1)} \\ &\times [D_1(A)+\lambda^2 D_1(G)+D_1(C)+D_1(T)]^{-K(1,2)} \\ &\times [D_1(A)+D_1(G)+\lambda^2 D_1(C)+D_1(T)]^{-K(1,3)} \\ &\times [D_1(A)+\lambda^2 D_1(G)+D_1(C)]^{-K(1,4)}[D_1(A)+D_1(G)+D_1(C)]^{-K(1,5)},\end{aligned} \tag{71}$$

where the last two terms reflect that stop codons are not allowed. Here $N(1,\cdot)$ and $K(1,\cdot)$ are all counts. Next, introduce a reparametrization by

$$(D_1(A), D_1(G), D_1(C), D_1(T)) = (\mathrm{e}^{\alpha_1}, \mathrm{e}^{\alpha_2}, \mathrm{e}^{\alpha_3}, 1)/(\mathrm{e}^{\alpha_1} + \mathrm{e}^{\alpha_2} + \mathrm{e}^{\alpha_3} + 1).$$

Differentiating the logarithm of (71) with respect to α_1, the likelihood equation can be written as

$$\begin{aligned} \mathrm{e}^{\alpha_1} = N(1,A)\Bigg[& \frac{N(1,T)}{\mathrm{e}^{\alpha_1}+\mathrm{e}^{\alpha_2}+\mathrm{e}^{\alpha_3}+1} + \frac{K(1,1)}{\mathrm{e}^{\alpha_1}+\lambda^2\mathrm{e}^{\alpha_2}+\lambda^2\mathrm{e}^{\alpha_3}+1} \\ & + \frac{K(1,2)}{\mathrm{e}^{\alpha_1}+\lambda^2\mathrm{e}^{\alpha_2}+\mathrm{e}^{\alpha_3}+1} + \frac{(K1,3)}{\mathrm{e}^{\alpha_1}+\mathrm{e}^{\alpha_2}+\lambda^2\mathrm{e}^{\alpha_3}+1} \\ & + \frac{K(1,4)}{\mathrm{e}^{\alpha_1}+\lambda^2\mathrm{e}^{\alpha_2}+\mathrm{e}^{\alpha_3}} + \frac{K(1,5)}{\mathrm{e}^{\alpha_1}+\mathrm{e}^{\alpha_2}+\mathrm{e}^{\alpha_3}}\Bigg]^{-1}, \end{aligned} \tag{72}$$

with similar expressions when differentiating with respect to α_2 and α_3. We solve these equations iteratively by inserting the present parameter value on the right hand side of (72) and solving for α on the left hand side.

Inserting the estimates for $D_j(\nu)$, as a function of λ, in (44) the estimate of λ is found by numerically maximizing this expression. To this end we use the approximation in Jensen and Pedersen (2000) for the normalizing constant C. The second derivative of this profile log likelihood function can also be found numerically.

Acknowledgements

I thank Ole Fredslund Christensen and Asger Hobolth for discussions related to the subject of this paper, and Breedette Hayes for assistance with the programming.

References

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

Arndt, P., D. Petrov, and T. Hwa (2003). Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol. 20*, 1887–1896.

Baum, L. and T. Petrie (1966). Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist. 37*, 1554–1563.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. B 36*, 192–236.

Christensen, O., A. Hobolth, and J. Jensen (2004). Pseudo-likelihood analysis of context-dependent codon substitution models. Resarch Report 29, MaPhySto, Aarhus University.

Douc, R., E. Moulines, and T. Rydén (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *Ann. Statist. 32*, 2254–2304.

Drton, M. (2004). *Maximum likelihood estimation in gaussian AMP chain graph models and gaussian ancestral graph models*. Ph. D. thesis, University of Washington.

Elashoff, M. and L. Ryan (2004). An em algorithm for estimating equations. *J. Comput. Graph. Statist. 13*, 485–65.

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

Götze, F. and C. Hipp (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrscheinlichkeitstheor. Verw. Geb. 64*, 211–240.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science 185*, 862–864.

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

Heyde, C. and R. Morton (1996). Quasi-likelihood and generalizing the em algorithm. *J. Roy. Statist. Soc. B 58*, 317–327.

Hwang, D. and P. Green (2004). Bayesian markov chain monte carlog sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA 101*, 13994–14001.

Ibragimov, I. and Y. Linnik (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

Jensen, J. (2005). Central limit theorem for functions of weakly dependent variables. Technical report, Department of Mathematical Sciences, University of Aarhus. In preparation.

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

Jensen, J. and N. Petersen (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist. 27*, 514–535.

Jørgensen, F., A. Hobolth, H. Hornshøj, C. Bendixen, M. Fredholm, and M. Schierup (2005). Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. *BMC Biology 3*, (http://www.biomedcentral.com/1741–7007/3/2).

Jukes, T. and C. Cantor (1969). Evolution of protein molecules. In H. Munro (Ed.), *Mammalian Protein Metabolism*, Volume 3, pp. 21–132. Academic Press, New York.

Kaiser, M. and N. Cressie (2000). The construction of multivariate distributions from marjov random fields. *J. Multivariate Anal. 73*, 199–220.

Kalbfleisch, J. and J. Lawless (1985). The analysis of panel data under a markov assumption. *J. Amer. Statist. Assoc. 80*, 863–871.

Karlin, S. and H. Taylor (1975). *A First Course in Stochastic Processes (Second edition)*. Academic Press, New York.

Kelly, F. (1979). *Reversibility and Stochastic Networks*. John Wiley & Sons, Chichester.

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

Kolmogorov, A. (1936). Zur theorie der markoffschen ketten. *Mathematische Annalen 112*, 155–160.

Lauritzen, S. (1996). *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, Oxford.

Louis, A. (1982). Finding the observed information matix when using the em algorithm. *J. Roy. Statist. Soc. B 44*, 226–233.

Muse, S. and B. Gaut (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substituion rates, with application to chloroplast genome. *Mol. Biol. Evol. 11*, 715–724.

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

Pedersen, A.-M., C. Wiuf, and F. Christiansen (1998). A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol. 15*, 1069–1081.

Rosen, O., W. Jiang, and M. Tanner (2000). Mixtures of marginal models. *Biometrika 87*, 391–404.

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

Siepel, A. and D. Haussler (2004). Phylogenetic estimation of a context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol. 21*, 469–488.