Probabilistic models of DNA sequence evolution with context dependent rates of substitution

Jens Ledet Jensen
Department of Theoretical Statistics, Institute of Mathematics,
Ny Munkegade, DK-8000 Aarhus C, Denmark

Anne-Mette Krabbe Pedersen
Department of Genetics and Ecology, Institute of Biological Sciences,
Ny Munkegade, DK-8000 Aarhus C, Denmark

Abstract

We consider Markov processes of DNA sequence evolution in which the instantaneous rates of substitution at a site are allowed to depend upon the states at the sites in a neighbourhood of the site at the instant of the substitution. We characterize the class of Markov process models of DNA sequence evolution for which the stationary distribution is a Gibbs measure, and give a procedure for calculating the normalizing constant of the measure. We develop an MCMC method for estimating the transition probability between sequences under models of this type. Finally, we analyze an alignment of two HIV–1 gene sequences using the developed theory and methodology.

Key words: dependent substitution rates; DNA sequences; Gibbs sampler; Markov dynamics; MCMC; stationary distribution;

AMS 1991 subject classification: primary 60J27; secondary 92D20, 62E25.

1 Introduction

In light of the tremendous effort currently invested in the extraction of molecular sequence data, e.g. in the numerous genome projects, the development of statistical tools for the analysis of this type of data is of great importance. Although within the last 20 years major advances have been made in providing sound statistical footing to the analysis (e.g. Felsenstein 1981 and Goldman 1993) the field is still relatively unexplored.

DNA sequences are strings of consecutive nucleotides of which there are four types: Adenosine (A), Guanine (G), Cytosine (C) and Thymine (T). Certain regions of the sequences - the genes - encode proteins, that is, strings of amino acids. In these regions triplets of nucleotides, called codons, are translated into amino acids via the genetic code. As there are 64 codons, with three being stop codons that signal end of translation, and only 20 amino acids, the genetic code is degenerate. Some amino acids are coded for by many codons (as many as six) whereas others are coded for by fewer (as few as one).

The statistical analysis of DNA sequences faces a number of difficulties due to the nature of the data and the complexity of the evolutionary processes shaping the data. Although some advances have been made in simultaneously dealing with the processes of insertion of new nucleotides and deletion of existing nucleotides and the process of substitution of existing nucleotides by new ones, the general approach taken is to separate the two. One aligns the sequences, that is, one arranges the nucleotides in columns so that nucleotides in the same column are believed to have descended from some common ancestral nucleotide through an evolutionary process that involves substitutions only. Having made this assumption one has reduced the size of the state space of the sequences considerably, now being of the order 4^n, where n is the length of the alignment. Sequences of interest, however, are generally from a few hundred to several thousands or ten thousands nucleotides long, leaving one with a state space that is still of considerable size.

The classical statistical approach taken when analyzing aligned DNA sequences is to assume that the evolutionary processes in the nucleotide sites are independent identical reversible Markov processes. The Markov process operating in a site is described by a rate matrix defining the rates of the different types of nucleotide substitutions (Felsenstein 1981). More recently

codon-based models designed to describe the evolution of protein coding sequences have been developed (Muse and Gaut 1994, Goldman and Yang 1994). These models allow the instantaneous rates of substitution at a site in a codon to depend upon the nucleotides occupying the other sites of the codon at the instant of the substitution. The evolutionary processes in the codons are assumed to be independent identical Markov processes with rates described by a matrix with 61×61 entries. The assumption of identical processes in different sites has been relaxed to the extent that the overall rates in the sites have been allowed to differ. One approach has been to draw a rate factor for each site independently from a Gamma distribution (Yang 1993), another to let the rates in sites be assigned by a Hidden Markov model (Felsenstein and Churchill 1996). The independence assumption, however, of the processes in non-overlapping entities along the sequences (nucleotide sites, codon sites or other short subsequences) is characteristic of models of DNA sequence evolution in general (e.g. Felsenstein 1981, Muse and Gaut 1994, Goldman and Yang 1994, Haeseler and Schöniger 1998). Having made this assumption the calculation of the likelihood is a matter of obtaining equilibrium frequencies and transition probabilities for the rate matrix assumed, and multiplying the appropriate products of the two along the alignment.

In this study we consider Markov processes of nucleotide substitution in which the independence assumption has been relaxed. The instantaneous rate of substitution at any site is allowed to depend upon the states of the sites in the neighbourhood of the site at the instant of the substitution. In Section 2 we present a model for the substitution process in Lentiviral genes which serves as a motivation for the study. We show that under this model the stationary measure for the codon sequence has a Gibbs form. The Gibbs form allows the measure to be written as a Markov chain along the sequence of codons, so that analysis of the stationary measure can be performed in a simple manner. In Section 3 we study models with context dependent rates of substitution in general. We arrive at a characterization of the class of intensities (substitution rates) for which the stationary distribution of the Markov process is a Gibbs measure, and describe how the Gibbs measure may be identified from the intensities. In Section 4 we utilize the Markovian nature of the Gibbs measure to derive a procedure for calculating the normalizing constant of the measure.

In Section 5 we define a codon-based Markov process of nucleotide substitution in which the only non-zero rates of substitution are those in which

a codon is changed at one position only. In this model the rates of substitution in the codon positions are allowed to depend upon the states at the other codon positions as well as on those at the nucleotide sites at either side of the codon. We show that the stationary distribution of this process is a Gibbs distribution, and give a simple procedure for calculating the normalizing constant, using the results derived earlier. In order to make a likelihood analysis of two sequences we develop an MCMC algorithm in Section 6 for calculating the transition probability from one sequence to another under a model of the type discussed in Section 5.

We finally apply the developed theory and methodology in an analysis of an alignment of two HIV–1 gene sequences (Section 7).

2 Motivating example

In this section we motivate our study through an example which is an extension of a model for the substitution process in Lentiviral genes considered by Pedersen et al. (1998). We consider a model where a change in a codon sequence consists in a change (substitution) of one nucleotide only. We write a codon sequence as $z_1, \ldots, z_n$, with $z_i = (z_i^1, z_i^2, z_i^3)$, where the upper index j in z_i^j indicates the position within the codon and $z_i^j \in \mathcal{T} = \{A, C, G, T\}$. Furthermore, we let $z_i(j, b)$ denote the new codon which is identical to z_i except at codon position j where z_i^j has been replaced by b. We allow the intensity γ for such a change to depend upon z_i as well as the neighbours z_{i-1}^3 and z_{i+1}^1:

$$\begin{aligned}\gamma(z_i(j,b); z_{i-1}^3, z_i, z_{i+1}^1) &= M(z_i, z_i(j,b))\pi_b^j \\ &\times \lambda_{31}^{1_{CG}(z_{i-1}^3, z_i(j,b)^1) - 1_{CG}(z_{i-1}^3, z_i^1)} \lambda_{12}^{1_{CG}(z_i(j,b)^1, z_i(j,b)^2) - 1_{CG}(z_i^1, z_i^2)} \\ &\times \lambda_{23}^{1_{CG}(z_i(j,b)^2, z_i(j,b)^3) - 1_{CG}(z_i^2, z_i^3)} \lambda_{31}^{1_{CG}(z_i(j,b)^3, z_{i+1}^1) - 1_{CG}(z_i^3, z_{i+1}^1)}, \end{aligned} \tag{1}$$

where the function M is given by

$$M(z_i, z_i(j,b)) = K^{1_{\mathrm{TS}}(z_i^j, b)} f^{1_{\mathrm{NON-SYN}}(z_i, z_i(j,b))},$$

with 1_{TS} an indicator function for a transition, that is the substitution of a purine (A or G) for a purine or a pyrimidine (C or T) for a pyrimidine, and with $1_{\mathrm{NON-SYN}}$ an indicator function for a change in the amino acid. We

restrict the π_b^j-parameters to sum to 1 for each j, and in the definition of the intensities we only consider those j, b for which $z_i(j,b)$ is not a stop codon. The interpretation of the λ parameters is given below.

The gene sequences of Lentiviruses, of which the HIV virus causing AIDS is an example, share a number of characteristic features. Their genomes have high contents of the A nucleotide, especially so at third codon positions, and extremely low contents of the CG dinucleotide. In a study of HIV–1 gene sequences Pedersen et al. (1998) considered the above model with $\lambda_{31} = 1$ and $\lambda_{12} = \lambda_{23}$. When $\lambda_{31} = 1$ the codons evolve independently and the stationary frequency for a codon (s_1, s_2, s_3) is

$$\begin{cases} \kappa\lambda^2\pi^1_{s_1}\pi^2_{s_2}\pi^3_{s_3} & \text{if } (s_1,s_2) = (CG) \text{ or } (s_2,s_3) = (CG) \\ \kappa\pi^1_{s_1}\pi^2_{s_2}\pi^3_{s_3} & \text{otherwise,} \end{cases} \tag{2}$$

where $\lambda = \lambda_{12} = \lambda_{23}$ is the common value and κ is a normalizing constant. As can be seen from the stationary frequencies this model can take into account different nucleotide compositions in the three codon positions and low frequencies of CG pairs at codon positions (1,2) and (2,3). However, in Lentiviral gene sequences low frequencies of CGs are observed *across* codon boundaries, that is, at codon positions (3,1), as well as *within* codons. In one of the HIV–1 genes examined by Pedersen et al. (1998) one has the following counts

		position one	
		G	non-G
position three	C	3	75
	non-C	155	267

These numbers clearly invalidate the hypothesis of independence among the codons. In Pedersen et al. (1998) a first attempt to take this into account is made. They picture a scenario where pentets evolve independently according to the intensities given in (1). In such a model the stationary frequencies for pentets $(z^3_{i-1}, z^1_i, z^2_i, z^3_i, z^1_{i+1})$ are on the form

$$\begin{cases} \kappa\lambda^4\pi^3_{s_0}\pi^1_{s_1}\pi^2_{s_2}\pi^3_{s_3}\pi^1_{s_4} & \text{two } CG\text{s in } \tilde{s} \\ \kappa\lambda^2\pi^3_{s_0}\pi^1_{s_1}\pi^2_{s_2}\pi^3_{s_3}\pi^1_{s_4} & \text{one } CG \text{ in } \tilde{s} \\ \kappa\pi^3_{s_0}\pi^1_{s_1}\pi^2_{s_2}\pi^3_{s_3}\pi^1_{s_4} & \text{no } CG \text{ in } \tilde{s}, \end{cases} \tag{3}$$

where $\tilde{s} = (s_0, s_1, s_2, s_3, s_4)$ is the pentet. They then perform an analysis in which the observed overlapping pentets in single sequences are treated as

independent observations and pentet counts are compared to the expected counts under different versions of the pentet based model (3). The results obtained through this analysis establish the importance of depression of CGs between codons. In recognition of the fact that overlapping pentets do not evolve independently the authors refrain from performing any evolutionary analysis using the pentet model.

The equilibrium frequencies (3) raise two immediate questions. The first question is what kind of rates $\gamma(\cdot,\cdot)$ give stationary frequencies of the form (3)? Secondly, what is the true dependency among the pentets? More generally, we want to study the relation between a specification of rates and the stationary measure for the codon sequence. For several reasons we also want the stationary measure to have a simple form. As argued in Pedersen et al. (1998) since two aligned sequences differ in a few positions only much of the information in the data concerns the stationary distribution. We may therefore want to start the modelling process with the stationary measure and to make simple tests here. Also, when a full model is analyzed it is useful to get initial values of the parameters from the stationary distribution (see Section 7).

For the model defined in (1) the stationary measure for the codon sequence is

$$\pi(z) = \frac{1}{Z}\left(\prod_{i=1}^{n}(\pi^1_{z_i^1}\pi^2_{z_i^2}\pi^3_{z_i^3})\right)\lambda_{31}^{2\sum_{i=1}^{n+1}1_{CG}(z_{i-1}^3,z_i^1)}\lambda_{12}^{2\sum_{i=1}^{n}1_{CG}(z_i^1,z_i^2)}\lambda_{23}^{2\sum_{i=1}^{n}1_{CG}(z_i^2,z_i^3)}, \tag{4}$$

where Z is a normalizing constant. This can be proved from Proposition 4 below, but can also be seen directly by showing that

$$\begin{aligned}&\pi(z)\,\gamma(z_i(j,b);z_{i-1}^3,z_i,z_{i+1}^1)\\ &\quad= \pi(z_1,\ldots,z_{i-1},z_i(j,b),z_{i+1},\ldots,z_n)\,\gamma(z_i,z_{i-1}^3,z_i(j,b),z_{i+1}^1).\end{aligned}$$

The stationary measure (4) is of the Gibbs form in (20) below, that is, it is specified through an interaction function $\exp(\phi(\cdot,\cdot))$ between z_{i-1}^3 and z_i^1:

$$\exp(\phi(a,b)) = (\lambda_{31}^2)^{1_{CG}(a,b)} \tag{5}$$

that measures the presence of the pair CG across a codon boundary, and a potential $\exp(\phi_0(\cdot))$ for the codon $s=(s_1,s_2,s_3)$ itself:

$$\exp(\phi_0(s)) = \pi^1_{s_1}\pi^2_{s_2}\pi^3_{s_3}(\lambda_{12}^2)^{1_{CG}(s_1,s_2)}(\lambda_{23}^2)^{1_{CG}(s_2,s_3)}. \tag{6}$$

The importance of the Gibbs structure is that the stationary measure becomes a Markov chain along the codon sequence. This makes it possible to analyze the stationary measure in a simple way. Since the interaction in (5) is between z_{i-1}^3 and z_i^1 only we have that the conditional distribution of z_i given z_{i-1} depends on z_{i-1}^3 only. Note that the Gibbs structure gives immediately from (5) and (6), in the case with $\lambda_{12} = \lambda_{23} = \lambda_{31}$, that *conditionally* on z_{i-1}^3 and z_{i+1}^1 the stationary frequencies for the codons are given by (3).

The Markov structure along the codon sequence embodied in (4) can be written in the form

$$P(z_i = (s_1, s_2, s_3) | z_{i-1}^3 = a) = \frac{\exp\{\phi(a, s_1) + \phi_0(s)\}}{q h(a)} h(s_3) \qquad (7)$$

for some number q and some function $h(\cdot)$. However, from (5) it follows that we can take $h(A) = h(G) = h(T) = 1$ and $h(C) = \tau$, say. Then τ and q are determined from the two equations

$$q = \sum_{s \in \tilde{S}} \exp(\phi_0(s)) h(s_3) \quad \text{and} \quad q\tau = \sum_{s \in \tilde{S}} (\lambda_{31}^2)^{1_G(s_1)} \exp(\phi_0(s)) h(s_3),$$

where $\tilde{S}$ is the set of non-stop codons. The important aspect of (7) is that this formula is explicit and that simple tests can be designed to test the adequacy of the transition probabilities in (7).

If we want to calculate the stationary probabilities of pentets we also need the invariant distribution of z_i^3 from (7). To this end, define the matrix $V(\cdot, \cdot)$ by

$$V(a, s_3) = \sum_{s_1, s_2 : s \in \tilde{S}} \exp(\phi(a, s_1) + \phi_0(s)), \qquad s = (s_1, s_2, s_3).$$

Then h is actually a right eigenvector of V with eigenvalue q. Since $V(A, \cdot) = V(G, \cdot) = V(T, \cdot)$ it is easy to see that a left eigenvector $l(\cdot)$, normalized such that $l(C) = 1$, has to be of the form

$$q l(b) = V(C, b) + V(A, b)\xi,$$

where $\xi = l(A) + l(G) + l(T)$ is determined by

$$q = V(C, C) + V(A, C)\xi.$$

The invariant distribution for z_i^3 in the Markov chain from (7) can now be written as

$$P(z_i^3 = a) = \frac{l(a)h(a)}{\tau + \xi} = \frac{1}{\tau + \xi}(V(C, a) + V(A, a)\xi)\tau^{1(a=C)}.$$

3 Relation between dynamics and stationary distribution: general case

We will consider a continuous time Markov process on the space of codon sequences of length n. Typically, we denote the sequence by $z = (z_1, \ldots, z_n)$ with $z_i \in \tilde{S}$, the set of non-stop codons. Usually we will imagine that we have two boundary codons z_0 and z_{n+1} which are fixed. These could be initiation or stop codons. When the process makes a jump there will only be a change in one codon. In this section a codon can be changed from any value to any other value in $\tilde{S}$, corresponding to all intensities being positive. In Section 5 we consider the case where a codon may be changed at one of its three positions only. The intensity for a change from z_i to y_i is allowed to depend on the two neighbouring codons, that is, the intensity is of the form

$$\gamma(y_i; z_{i-1}, z_i, z_{i+1}). \tag{8}$$

Note that when $1 < i < n$ all the codon variables will belong to $\tilde{S}$, whereas when $i = 1$ we allow z_0 to be a stop codon and similarly with z_{n+1}.

As mentioned in Sections 1 and 2 we want the stationary distribution to be of a form that allows explicit calculations. This gives us the possibility of designing tests based on one sequence only and simple estimates can be obtained for some of the parameters of the model. We have chosen a Gibbs measure for the stationary distribution, i.e. a measure defined through an interaction function $\phi : \tilde{S} \times \tilde{S} \to \mathbf{R}$ and given by

$$P(z) = \frac{1}{Z} \exp\left\{\phi_1(z_1) + \sum_{i=2}^{n} \phi(z_{i-1}, z_i) + \phi_n(z_n)\right\}, \tag{9}$$

where Z is a normalizing constant. The functions ϕ_1 and ϕ_n are defined on $\tilde{S}$. We show in the next section that the Gibbs form allows us to rewrite this measure as a Markov chain along the codon sequence, thereby making the analysis of the measure simple. Here our first result describes a sufficient

condition on γ in order that the stationary distribution is given by (9). Actually, the sufficient condition imposed will be derived from a requirement of time reversibility of the Markov process.

Proposition 1 *If*

$$\frac{\gamma(t; z_0, s, r_2)}{\gamma(s; z_0, t, r_2)} = \frac{\exp\{\psi_1(t; r_2)\}}{\exp\{\psi_1(s; r_2)\}}, \quad \frac{\gamma(t; r_1, s, z_{n+1})}{\gamma(s; r_1, t, z_{n+1})} = \frac{\exp\{\psi_n(t; r_1)\}}{\exp\{\psi_n(s; r_1)\}}, \tag{10}$$

$$\frac{\gamma(t; r_1, s, r_2)}{\gamma(s; r_1, t, r_2)} = \frac{\exp\{\psi(t; r_1, r_2)\}}{\exp\{\psi(s; r_1, r_2)\}}, \tag{11}$$

for all s, t, r_1, r_2 *in* $\tilde{S}$, *where*

$$\psi_1(s; r_2) = \phi_1(s) + \phi(s, r_2), \quad \psi_n(s; r_1) = \phi(r_1, s) + \phi_n(s),$$
$$\psi(s; r_1, r_2) = \phi(r_1, s) + \phi(s, r_2),$$

then (9) is the stationary distribution for the Markov process with intensities given in (8).

Proof. For any continuous time Markov process with intensity $\lambda(a, b)$ of going from state a to state b we have that π is the stationary distribution if

$$\pi(a)\lambda(a, b) = \pi(b)\lambda(b, a) \quad \forall\ a \neq b.$$

In our case the intensity is zero unless the codon sequence is changed in one codon only. Let us consider the case where there is a change at codon j with $1 < j < n$. The change is from z_j to y_j. Let $\tilde{z}$ be equal to z except at position j where $\tilde{z}_j = y_j$. Then the above equation becomes

$$\begin{aligned} &\frac{1}{Z} \exp\left\{\phi_1(z_1) + \sum_{i=2}^{n} \phi(z_{i-1}, z_i) + \phi_n(z_n)\right\} \gamma(y_j; z_{j-1}, z_j, z_{j+1}) \\ &= \frac{1}{Z} \exp\left\{\phi_1(\tilde{z}_1) + \sum_{i=2}^{n} \phi(\tilde{z}_{i-1}, \tilde{z}_i) + \phi_n(\tilde{z}_n)\right\} \gamma(z_j; z_{j-1}, y_j, z_{j+1}). \end{aligned}$$

Taking the γ-terms on one side and the ϕ-terms on the other side we get

$$\frac{\gamma(y_j; z_{j-1}, z_j, z_{j+1})}{\gamma(z_j; z_{j-1}, y_j, z_{j+1})} = \frac{\exp\{\phi(z_{j-1}, y_j) + \phi(y_j, z_{j+1})\}}{\exp\{\phi(z_{j-1}, z_j) + \phi(z_j, z_{j+1})\}},$$

which is of the form stated in the proposition. The cases with $j = 1$ and $j = n$ are treated similarly. $\square$

Next, we consider what class of intensities satisfies the restrictions in Proposition 1.

Proposition 2 *The intensities (8) satisfy the relation (11) if and only if we can write*

$$\log(\gamma(t; r_1, s, r_2)) = -\psi(s; r_1, r_2) + l(s, t; r_1, r_2) \tag{12}$$

for all s, t, r_1, r_2 in $\tilde{S}$, where l is symmetric in (s, t). Similar statements can be made for the restrictions in (10) where ψ is replaced by ψ_1 and ψ_n, respectively.

Proof. The 'if' statement is trivial. For the 'only if' statement we write $f(s, t; r_1, r_2) = \log(\gamma(t; r_1, s, r_2))$. Then the relation in Proposition 1 can be written as

$$f(s, t; r_1, r_2) - f(t, s; r_1, r_2) = \psi(t; r_1, r_2) - \psi(s; r_1, r_2). \tag{13}$$

If f_1 and f_2 are solutions to (13) then we find that $f_1 - f_2$ is a symmetric function in (s, t). Since also $f_0(s, t; r_1, r_2) = -\psi(s; r_1, r_2)$ is a solution we have that all solutions to (13) are on the form

$$f(s, t; r_1, r_2) = -\psi(s; r_1, r_2) + l(s, t; r_1, r_2),$$

with l an arbitrary function symmetric in (s, t). $\square$

The relation (12) can be formulated more symmetrically in s and t. Let $k(s; r_1, r_2)$ be an arbitrary function and let $\tilde{l}(s, t; r_1, r_2)$ be an arbitrary function symmetric in (s, t). Then $l(s, t; r_1, r_2) = \tilde{l}(s, t; r_1, r_2) + k(s; r_1, r_2) + k(t; r_1, r_2)$ is also symmetric in (s, t), and (12) becomes

$$\begin{aligned} f(s, t; r_1, r_2) &= (-\psi(s; r_1, r_2) + k(s; r_1, r_2)) + k(t; r_1, r_2) + \tilde{l}(s, t; r_1, r_2) \\ &= h(s; r_1, r_2) + k(t; r_1, r_2) + \tilde{l}(s, t; r_1, r_2). \end{aligned} \tag{14}$$

Thus, the only restriction in (14) is that the functions h and k must satisfy the relation $k(s; r_1, r_2) - h(s; r_1, r_2) = \psi(s; r_1, r_2)$.

Proposition 2 shows that the log intensity splits into a non-symmetrical part, $-\psi(s; r_1, r_2)$, that can be determined from the stationary distribution, and a symmetrical part that does not influence the stationary distribution. The latter must be determined from the dynamics of the process.

We next consider the possibility of determining ψ and ϕ from the log intensities $f(s, t; r_1, r_2) = \log(\gamma(t; r_1, s, r_2))$.

Proposition 3 *Assume that the log intensities can be written as*

$$f(s, t; r_1, r_2) = -g(s; r_1, r_2) + l(s, t; r_1, r_2), \quad s, t, r_1, r_2 \in \tilde{S}, \tag{15}$$

where l is symmetric in (s, t). Further assume that there exists a function $q(r_1, r_2)$ such that

$$g(s; r_1, r_2) = g(s; r_1, *) - g(*; r_1, *) + g(r_2; s, *) - g(*; s, *) + q(r_1, r_2) \tag{16}$$

for all $s, r_1, r_2 \in \tilde{S}$, where an asterisk means that we have taken the average over $\tilde{S}$ with respect to that coordinate. Then the stationary distribution is of the form (9) with

$$\phi(r, s) = g(s; r, *) - g(*; r, *). \tag{17}$$

Proof. Using (17) the formula (16) states that

$$g(s, r_1, r_2) = \phi(r_1, s) + \phi(s, r_2) + q(r_1, r_2).$$

Defining $\psi(s; r_1, r_2) = \phi(r_1, s) + \phi(s, r_2)$ we then have

$$g(t, r_1, r_2) - g(s, r_1, r_2) = \psi(t, r_1, r_2) - \psi(s, r_1, r_2),$$

which proves (11), and from Proposition 1 we obtain that the stationary distribution is given by (9). □

The assumption (16) is necessary if the intensities are of the form (15) and we want (11) to hold as well. This is because we from Proposition 2 have $f = -\psi + \tilde{l}$, and combining this with (15) we find

$$\begin{aligned} \psi(s; r_1, r_2) - g(s; r_1, r_2) &= \tilde{l}(s, t; r_1, r_2) - l(s, t; r_1, r_2) \\ &= \tilde{l}(t, s; r_1, r_2) - l(t, s; r_1, r_2) \\ &= \psi(t; r_1, r_2) - g(t; r_1, r_2), \end{aligned}$$

which shows that

$$g(s; r_1, r_2) = \psi(s; r_1, r_2) + \tilde{q}(r_1, r_2),$$

for some function $\tilde{q}$. Using this one finds that (16) holds.

The importance of Proposition 3 is that it gives a way of verifying (11) and at the same time identifying ϕ in the stationary distribution (9).

4 Stationary Gibbs measure as a Markov chain

In this section we consider the Gibbs measure (9) for a codon sequence. The normalizing constant Z in (9) is not easy to find directly. To find Z and properties of the Gibbs measure we will rewrite the measure as a Markov chain along the sequence of codons. Define

$$Q(a, b) = \exp\{\phi(a, b)\}, \quad a, b \in \tilde{S},$$

where $\tilde{S}$ is the set of 61 non-stop codons. Let q be the largest eigenvalue of Q and let r be a right eigenvector corresponding to q,

$$\sum_b Q(a, b) r(b) = q r(a), \quad \forall\ a \in \tilde{S}.$$

Define next

$$T(a, b) = \frac{Q(a, b) r(b)}{q r(a)}, \quad a, b \in \tilde{S}.$$

This is a transition matrix and clearly

$$\prod_{i=2}^{n} T(z_{i-1}, z_i) = \frac{r(z_n)}{q^{n-1} r(z_1)} \exp\left\{ \sum_{i=2}^{n} \phi(z_{i-1}, z_i) \right\},$$

and $P(z)$ from (9) becomes

$$P(z) = \frac{1}{Z} \frac{q^{n-1} r(z_1)}{r(z_n)} \exp\{\phi_1(z_1) + \phi_n(z_n)\} \prod_{i=2}^{n} T(z_{i-1}, z_i).$$

This finally shows that

$$Z = q^{n-1} \sum_{a, b \in \tilde{S}} \frac{r(a)}{r(b)} \exp\{\phi_1(a) + \phi_n(b)\} T^{n-1}(a, b), \tag{18}$$

which can be evaluated numerically, allowing $P(z)$ to be calculated. The calculation of T^{n-1} requires of the order $61^2 n$ calculations. An approximation to Z based on the stationary distribution for T is given below.

Frequencies in the limit $n \to \infty$ can be evaluated with the help of the stationary distribution for the transition matrix T. Let l be a left eigenvector for Q with eigenvalue q,

$$\sum_a l(a)Q(a,b) = ql(b), \quad \forall\ b \in \tilde{S},$$

and normalize l so that

$$\sum_a l(a)r(a) = 1.$$

Then $\{l(a)r(a),\ a \in S\}$ is the stationary distribution for T. Thus the frequency of a particular codon c will converge to $l(c)r(c)$ as $n \to \infty$, and the frequency of a particular neighbouring pair of codons (c_1, c_2) will converge to $l(c_1)r(c_1)T(c_1,c_2) = q^{-1}l(c_1)\exp(\phi(c_1,c_2))r(c_2)$.

Returning to (18) we approximate $T^{n-1}(a,b)$ by the stationary probability $l(b)r(b)$. This gives the approximation

$$Z \approx q^{n-1} \sum_{a,b \in \tilde{S}} r(a)\exp\{\phi_1(a) + \phi_n(b)\}l(b)r(b). \tag{19}$$

5 Restricted dependency

In this section we consider the case where the only allowed jumps in the Markov process are those where one codon is changed in one position only. For two codons $s, t \in \tilde{S}$, $s \neq t$, we write $s \sim t$ if they differ at one position only, and generally we write $s = (s^1, s^2, s^3)$, $s^j \in \mathcal{T}$, with $\mathcal{T}$ being the set of nucleotides, for the values at the three positions. Also, we will make the restriction that the intensity for a change at position i depends on (z_{i-1}, z_{i+1}) through (z^3_{i-1}, z^1_{i+1}) only. We write the intensity for a change from z_i to y_i as

$$\gamma(y_i; z^3_{i-1}, z_i, z^1_{i+1}), \quad y_i \sim z_i,\ z^3_{i-1}, z^1_{i+1} \in \mathcal{T}.$$

In analogy with (9) we want the stationary distribution of the codon sequence to be of the following Gibbs form

$$P(z) = \frac{1}{Z}\exp\left\{\sum_{i=1}^{n+1}\phi(z^3_{i-1}, z^1_i) + \sum_{i=1}^{n}\phi_0(z_i)\right\}, \tag{20}$$

where again Z is a normalizing constant, and z_0^3, z_{n+1}^1 are given. The function $\phi(\cdot,\cdot)$ is now defined on $\mathcal{T}\times\mathcal{T}$ and ϕ_0 is defined on $\tilde{S}$. The $\phi(z_{i-1}, z_i)$ function in (9) corresponds here to $\phi(z_{i-1}^3, z_i^1) + \phi_0(z_i)$. The next proposition gives a sufficient condition on the intensities in order that the stationary distribution is given by (20).

Proposition 4 *If*

$$\frac{\gamma(t;a,s,b)}{\gamma(s;a,t,b)} = \frac{\exp\{\psi(t;a,b)\}}{\exp\{\psi(s;a,b)\}} \quad \forall a,b\in\mathcal{T}, s,t\in\tilde{S}, \tag{21}$$

where $\psi(s;a,b) = \phi(a,s^1) + \phi_0(s) + \phi(s^3,b)$, *then the stationary distribution is given by (20). Furthermore, the intensity satisfies (21) if and only if we can write*

$$\log(\gamma(t;a,s,b)) = -\psi(s;a,b) + l(s,t;a,b), \quad \forall a,b\in\mathcal{T}, s,t\in\tilde{S}, s\sim t, \tag{22}$$

where l *is symmetric in* (s,t), $s\sim t$.

Proof. The proof is as for Proposition 1 and Proposition 2. □

We next turn to a discussion of the possibility of determining the ϕ and ϕ_0 functions in (20) from the intensities. From (22) we see that, intuitively, the question is if we can determine ϕ and ϕ_0 from knowing the function ψ.

Proposition 5 *Assume that the log intensities can be written as*

$$\log(\gamma(t;a,s,b)) = -g(s;a,b) + l(s,t;a,b), \quad s,t\in\tilde{S}, t\sim s, a,b\in\mathcal{T}, \tag{23}$$

where l *is symmetric in* (s,t). *Define*

$$\tilde{g}(s^1,s^3;a,b) = g(s^1,*,s^3;a,b),$$

where an asterisk denotes that the average has been taken with respect to s^2 *over the set with* $(s^1,s^2,s^3)\in\tilde{S}$. *Assume further that*

(i) There exists a function $q(a,b)$ *such that*

$$\begin{aligned}\tilde{g}(s^1,s^3;a,b) - \tilde{g}(s^1,*;a,b) - \tilde{g}(*,s^3;a,b)\\ +\tilde{g}(s^1,*;*,*) + \tilde{g}(*,s^3;*,*) - \tilde{g}(s^1,s^3;*,*) = q(a,b)\end{aligned}$$

for all $a,b,s^1,s^3\in\mathcal{T}$;

(ii) The function

$$g(s;a,b) - \tilde{g}(s^1, s^3; a, b) + \tilde{g}(s^1, s^3; *, *)$$

does not depend on (a,b);

(iii) The function

$$\tilde{g}(s^1, s^3; a, b) - \tilde{g}(s^1, *; a, b) - \tilde{g}(s^1, s^3; *, *) + \tilde{g}(s^1, *; *, *)$$

does not depend on (a, s^1)*, the function*

$$\tilde{g}(s^1, s^3; a, b) - \tilde{g}(*, s^3; a, b) - \tilde{g}(s^1, s^3; *, *) + \tilde{g}(*, s^3; *, *)$$

does not depend on (s^3, b)*, and the two functions are identical.*

We denote the function in (ii) by $\phi_0(s)$ *and the common function in (iii) by* $\phi(\cdot, \cdot)$*, that is,* $\phi(s^3, b)$ *in the first case and* $\phi(a, s^1)$ *in the second case. Then the stationary distribution is given by (20).*

Proof. With the definitions from (ii) and (iii) we can write (i) as

$$g(s;a,b) = \phi(a, s^1) + \phi_0(s) + \phi(s^3, b) + q(a, b).$$

The assumption (21) is then satisfied and Proposition 4 shows that (20) is the stationary distribution. □

As in the case with Proposition 3 the assumptions (i), (ii) and (iii) are necessary if we want the stationary distribution to be given by (20) and we want (21) to be satisfied.

We next turn to a discussion of the stationary distribution given by (20). Define

$$V(a, b) = \sum_{s^1, s^2 : (s^1, s^2, b) \in \tilde{S}} \exp\{\phi(a, s^1) + \phi_0(s^1, s^2, b)\}, \tag{24}$$

where b here is the value at the third position of the codon, and let q be the largest eigenvalue with corresponding right eigenvector h and left eigenvector l,

$$\sum_{b \in \mathcal{T}} V(a, b)h(b) = qh(a), \quad \sum_{a \in \mathcal{T}} V(a, b)l(a) = ql(b).$$

We standardize so that $h(A) = 1$ and $\sum_{a\in\mathcal{T}} l(a)h(a) = 1$. Finally, we define

$$k(a) = \sum_{s^2,s^3:(a,s^2,s^3)\in\tilde{S}} \exp\{\phi_0(a, s^2, s^3)\}h(s^3).$$

We then have

$$\begin{aligned}
&\frac{1}{Z}\exp\left\{\sum_{i=1}^{n+1}\phi(z_{i-1}^3, z_i^1) + \sum_{i=1}^{n}\phi_0(z_i)\right\} \\
&= \frac{q^{n+1}h(z_0^3)}{Zk(z_{n+1}^1)}\prod_{i=1}^{n}\left\{p_1(z_i^1|z_{i-1}^3)p_2(z_i^2, z_i^3|z_i^1)\right\}p_1(z_{n+1}^1|z_n^3),
\end{aligned}$$

where

$$p_1(b|a) = \frac{\exp\{\phi(a,b)\}k(b)}{qh(a)} \tag{25}$$

is the conditional distribution of z_i^1 given z_{i-1}^3, and

$$p_2(s^2, s^3|s^1) = \frac{\exp\{\phi_0(s^1, s^2, s^3)\}h(s^3)}{k(s^1)} \tag{26}$$

is the conditional distribution of (z_i^2, z_i^3) given (z_{i-1}^3, z_i^1) which depends on z_i^1 only. Finally, $\{z_i^3\}$ constitute a Markov chain with transition probabilities

$$p_3(b|a) = \frac{V(a,b)h(b)}{qh(a)}, \tag{27}$$

and with stationary distribution

$$P(z_i^3 = a) = l(a)h(a), \quad a \in \mathcal{T}. \tag{28}$$

The normalizing constant Z is then

$$\begin{aligned}
Z &= q^n h(z_0^3)\sum_{z_1^3,\ldots,z_n^3}\left\{\left(\prod_i p_3(z_i^3|z_{i-1}^3)\right)\frac{\exp(\phi(z_n^3, z_{n+1}^1))}{h(z_n^3)}\right\} \\
&= q^n h(z_0^3)\sum_b\left\{P(Z_n^3 = b|z_0^3)\frac{\exp(\phi(b, z_{n+1}^1)}{h(b)}\right\}.
\end{aligned} \tag{29}$$

In the particular case of (20) where $\{z_i^3\}$ constitute a sequence of independent variables some of the formulas above are simplified. Since $\{z_i^3\}$ is a

Markov chain with transition probabilities given in (27) we see that we have independence if and only if there exist functions r and t such that

$$V(a,b) = r(a)t(b), \quad a,b \in \mathcal{T}. \tag{30}$$

If we standardize so that $r(A) = 1$ we find that

$$h = r, \quad q = \sum_b t(b)r(b), \quad \text{and} \quad l = t/q.$$

The formulas (25)–(28) can be used to test the adequacy of the model. They were also used for the formulas given in Section 2 above.

6 Simulation of the transition probability

Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two aligned codon sequences with gaps removed. Assume that aligned codons in the two sequences have evolved from some common ancestor codon through an evolutionary process that involves substitution events only. We want to evaluate the probability for the transition from x to y from time 0 to time t in a model specified by a set of intensities $\gamma(z_i(j,b); (z^3_{i-1}, z_i, z^1_{i+1}))$, where an upper index 1,2 or 3 indicates the position within the codon. This is the intensity for a change in codon i from the present value z_i to the codon $z_i(j,b)$ obtained by replacing z^j_i by b, and z^3_{i-1} and z^1_{i+1} are the present values at positions immediately to the left and right, respectively, of the codon.

Since the state space of x is enormous there seems to be no possibility of evaluating the transition probability analytically nor is it possible to evaluate the probability by a forward simulation of the codon sequence. Instead we here propose an MCMC simulation technique for evaluating the ratio of two transition probabilities. The idea of considering a ratio is similar to the method discussed in Geyer (1996).

A path L starting at the codon sequence x and ending at the codon sequence y can be written as $\{r; d_1, \ldots, d_r; a_1, \ldots, a_r; u_1, \ldots, u_r\}$, where r is the number of jumps, the d_i's are the positions of the jumps, the a_i's are the new nucleotides (the jump type) and the u_i's are the jump times. In order that the path ends at y and in order that a stop codon is not produced, the positions and the jump types belong to a subset K_r of $(\{1, \ldots, n\} \times \{1,2,3\} \times$

$\mathcal{T})^r$. The set of paths can then be described as

$$\mathcal{X}_t = \cup_{r=r_0}^{\infty} K_r \times [0,t]^r,$$

where $r_0 = \sum_{i,j} 1(x_i^j \neq y_i^j)$, that is, the minimal number of jumps required for one sequence to evolve into the other. Let $\mathcal{X}$ be a 'standard' jump space corresponding to $t = 1$. We equip both $\mathcal{X}$ and the original jump space $\mathcal{X}_t$ with the measure which is Lebesgue measure for the jump times and counting measure on the other coordinates. We denote the two measures by μ and μ_t respectively. When we write tL we mean the jump path where all the jump times have been multiplied by t. Let θ parametrize the intensities. For a path L in $\mathcal{X}_t$ we denote the weight of this path with respect to μ_t by $q_\theta(t; L)$. We can then write the transition probability as

$$\begin{aligned} P_{\theta,t}(y|x) &= \int_{\mathcal{X}_t} q_\theta(t; L) d\mu_t \\ &= \int_{\mathcal{X}_t} q_\theta(t; r; d_1, \ldots, d_r; a_1, \ldots, a_r; u_1, \ldots, u_r) d\mu_t \\ &= \int_{\mathcal{X}} t^r q_\theta(t; r; d_1, \ldots, d_r; a_1, \ldots, a_r; u_1 t, \ldots, u_r t) d\mu \\ &= \int_{\mathcal{X}} t^r q_\theta(t; tL) d\mu. \end{aligned}$$

Let now P be a measure living on $\mathcal{X}_{t_0}$ with density $q(\cdot)$ with respect to μ_{t_0}. Then

$$P_{\theta,t}(y|x) = \int_{\mathcal{X}} \frac{t^r q_\theta(t; tL)}{t_0^r q(t_0 L)} t_0^r q(t_0 L) d\mu = \int_{\mathcal{X}_{t_0}} \frac{t^r q_\theta(t; \frac{t}{t_0} L)}{t_0^r q(L)} dP,$$

and

$$\begin{aligned} \frac{P_{\theta_1,t}(y|x)}{P_{\theta_2,t_0}(y|x)} &= \frac{\int_{\mathcal{X}_{t_0}} \frac{t^r q_{\theta_1}(t; \frac{t}{t_0} L)}{t_0^r q(L)} dP}{\int_{\mathcal{X}_{t_0}} \frac{t_0^r q_{\theta_2}(t_0; L)}{t_0^r q(L)} dP} \\ &= E\left(\frac{t^r q_{\theta_1}(t; \frac{t}{t_0} L)}{t_0^r q_{\theta_2}(t_0; L)} \frac{\frac{q_{\theta_2}(t_0; L)}{q(L)}}{E\left(\frac{q_{\theta_2}(t_0; L)}{q(L)}\right)} \right) \\ &= \tilde{E}\left(\frac{t^r q_{\theta_1}(t; \frac{t}{t_0} L)}{t_0^r q_{\theta_2}(t_0; L)} \right), \end{aligned} \tag{31}$$

where $\tilde{E}$ is the mean under the measure $\tilde{P}$ having density

$$\frac{\frac{q_{\theta_2}(t_0;L)}{q(L)}}{E\left(\frac{q_{\theta_2}(t_0;L)}{q(L)}\right)}$$

with respect to P. We therefore want to be able to simulate observations from $\tilde{P}$.

We will use MCMC. The principle behind the Gibbs sampler seems appropiate here: that is, to update a codon path from the conditional distribution given the paths of all the other codons. So, we must investigate the conditional $\tilde{P}$ measure for a path for codon i given the paths of all the other codons. To make things easy we will assume that $P = \prod_i P_i$ is a product measure over the codons with corresponding density $\prod_i q_i(L_i)$, where L_i is the path for the i'th codon. The weight function $q_\theta(t;L)$ represents a product over all nucleotide sites of the contribution from the events along the path for this particular site. For the class of intensities we consider the contribution at site 1 of a codon depends on site three of the previous codon and sites 2 and 3 of the same codon. Similar statements can be made for sites 2 and 3 of a codon and we end up with

$$q_\theta(t;L) = \prod_i q_\theta(t;L_i^1|L_{i-1}^3,L_i^2,L_i^3)q_\theta(t;L_i^2|L_i^1,L_i^3)q_\theta(t;L_i^3|L_i^1,L_i^2,L_{i+1}^1), \quad (32)$$

where L_i^j is the path at position j of codon i, and each term represents the contribution from the path at the indicated position. The functions here are calculated explicitly below. The conditional density with respect to P_i for a path for codon i given the paths at all other codons is therefore

$$\frac{1}{Z_i}q_{\theta_2}(t_0;L_{i-1}^3|L_{i-1}^1,L_{i-1}^2,L_i^1)q_{\theta_2}(t_0;L_i^1|L_{i-1}^3,L_i^2,L_i^3)$$
$$\times q_{\theta_2}(t_0;L_i^2|L_i^1,L_i^3)q_{\theta_2}(t_0;L_i^3|L_i^1,L_i^2,L_{i+1}^1)q_{\theta_2}(t_0;L_{i+1}^1|L_i^3,L_{i+1}^2,L_{i+1}^3)/q_i(L_i),$$

where Z_i is a normalizing constant. We denote the density by $1/Z_i$ times the function $\pi(L_i|L_{i-1},L_{i+1})$. It is not obvious how to simulate directly from this conditional distribution. Instead we will take one step in a Markov chain that has the conditional distribution as its stationary distribution. We do this by proposing a move to $\tilde{L}_i$ drawn from P_i. The move is accepted with probability

$$\alpha(\tilde{L}_i) = \min\left\{\frac{\pi(\tilde{L}_i|L_{i-1},L_{i+1})}{\pi(L_i|L_{i-1},L_{i+1})},1\right\}.$$

The whole process is started by drawing L_i, $i = 1, \ldots, n$ from P_i. Next, we run through $i = 1$ to n a large number of times making draws from P_i and accepting these with the probability α.

To impliment the above MCMC algorithm we need to make a choice of P_i. We will actually construct a measure, which we also denote by P_i, on an extended path space. The idea is that we first suggest a number of jumps, say k, and then the resulting path will have either $r = k$ or $r = k - 1$ jumps and this is described by a variable $J \in \{0, 1\}$. Below we only show the index i when needed. If we define $d_0 = \sum_j 1(x_i^j \neq y_i^j)$ we can describe the extended path space as

$$\begin{array}{ll} \{0\} \cup_{r=2}^{\infty} \{0,1\} \times K_r^i \times [0,t_0]^r & \text{if } x_i = y_i \\ \cup_{r=d_0}^{\infty} \{0,1\} \times K_r^i \times [0,t_0]^r & \text{if } x_i \neq y_i, \end{array} \tag{33}$$

where K_r^i is the subset of $(\{1,2,3\} \times \mathcal{T})^r$ corresponding to the path for codon i ending up at y_i. The proposed number of jumps is a random variable $K \geq 0$ with distribution

$$p(k) = P(K = k) = \begin{cases} \frac{\gamma^k}{k!} \mathrm{e}^{-\gamma} / (1 - \gamma \mathrm{e}^{-\gamma}) & x_i = y_i, \ k = 0, 2, 3, \ldots \\ \frac{\gamma^k}{k!} \mathrm{e}^{-\gamma} / (1 - \sum_{l=0}^{d_0} \frac{\gamma^l}{l!} \mathrm{e}^{-\gamma}) & x_i \neq y_i, \ k > d_0, \end{cases}$$

where

$$\gamma = \sum_{j,b} \gamma(x_i(j,b); x_{i-1}^3, x_i, x_{i+1}^1)$$

is the intensity of leaving x_i. Let $x(m) = (x^1(m), x^2(m), x^3(m))$ be the codon after the m'th jump, with $x(0) = x_i$, and let $d_m = \sum_j 1(x^j(m) \neq y_i^j)$ be the number of differences between $x(m)$ and y_i. If we want the codon x_i to evolve into y_i in k jumps, that is, if we want $x(k) = y_i$ we must have $k - m \geq d_m$ for all m. This then defines a set of allowable jumps $A(k, m)$ for the m'th jump, where this set is also restricted in order that $x(m) \neq x(m-1)$ and in order that $x(m)$ is not a stop codon. The total intensity among the allowable jumps is

$$\gamma(k, m) = \sum_j \sum_{b \,:\, x(m,j,b) \in A(k,m)} \gamma(x(m,j,b); x_{i-1}^3, x(m-1), x_{i+1}^1),$$

where $x(m, j, b)$ is the codon with value b at position j and equal to $x(m-1)$ at the two other positions. We then choose the m'th jump $x(m) = x(m, j, b)$

in $A(k,m)$ with probability $\gamma(x(m,j,b);x_{i-1}^3,x(m-1),x_{i+1}^1)/\gamma(k,m)$. If it happens that $d_{k-1}=0$ we set $J=1$ and $r=k-1$ and if $d_{k-1}=1$ we set $J=0$ and $r=k$. The probability of the chosen jump sequence is

$$\omega_r(J)=\begin{cases}\prod_{m=1}^{r-1}\frac{\gamma(x(m);x_{i-1}^3,x(m-1),x_{i+1}^1)}{\gamma(r,m)} & \text{if } J=0\\ \prod_{m=1}^{r}\frac{\gamma(x(m);x_{i-1}^3,x(m-1),x_{i+1}^1)}{\gamma(r+1,m)} & \text{if } J=1,\end{cases}.$$

Finally, conditioned on the jumps, we take the r jump times to be uniformly distributed on the interval $[0,t_0]$.

We have now constructed a measure on the extended path space (33). We now marginalize to get a mesure on the path space. The density of the marginal measure is

$$q_i(L_i)=\{p(r)\omega_r(0)+p(r+1)\omega_r(1)\}\frac{r!}{t_0^r},$$

where the path L_i has r jumps. Finally, we describe the calculation of the functions appearing in (32). Let the number of jumps in codon i together with the third position of codon $i-1$ and the first position of codon $i+1$ be s. Let $u_1,\ldots,u_s$ be the jump times and let $\tilde{z}(m)=(z^0(m),z(m),z^4(m))$ with $z(m)=(z^1(m),z^2(m),z^3(m))$ be the corresponding nucleotides occupying the third codon position of codon $i-1$, the three positions in codon i and the first position in codon $i+1$ after the m'th jump. Then

$$\begin{aligned}&q_\theta(t;L_i^j|\cdot)=\\ &\quad\left\{\prod_{m=1}^{s}(\gamma(z(m);\tilde{z}(m-1)))^{1(z^j(m)\neq z^j(m-1))}\exp\{-\gamma_j(*;\tilde{z}(m-1))(u_m-u_{m-1})\}\right\}\\ &\quad\times\exp\{-\gamma_j(*;\tilde{z}(s))(t-u_s)\},\end{aligned}$$

where $u_0=0$ and for $j=1,2,3$,

$$\gamma_j(*;\tilde{z}(m-1))=\sum_b\gamma(z(m,j,b);\tilde{z}(m-1)),$$

with an implicit dependence on i through $z(\cdot)$ and with $z(m,j,b)$ the codon $z(m-1)$ with $z^j(m-1)$ replaced by b.

7 Numerical example

In this section we analyze one of the pairwise alignments of HIV–1 sequences considered by Pedersen et al. (1998) using the model and MCMC estimation procedure described in Sections 2 and 6. We use the procedure described in Section 5 when calculating a normalizing constant of a stationary distribution. The alignment contains 430 consecutive codons from the single-coding region of the *gag* genes. In this region internal structural proteins of the virus particle are encoded. The aligned sequences differ at 80 nucleotide positions.

Good initial values of the parameters π_i^k, $i \in \{A, C, G, T\}$, $k = 1, 2, 3$, to be used in the MCMC procedure may be obtained by finding the maximum likelihood estimates under the stationary distribution for one of the sequences. As starting values for the remaining parameters we used the estimates obtained in Pedersen et al. (1998). We examined the burn-in of the Gibbs sampler for evaluating the transition probability using the parameter values mentioned above. Burn-in was detected by displaying the number of new paths accepted in one updating by the Gibbs sampler, as well as summary statistics regarding the total number of jumps and different types of jumps. For the parameter values used, the Gibbs sampler starts out accepting a high number of new paths, but stabilizes immediately hereafter. The summary statistics confirmed the almost immediate burn-in for the same parameter values. Only when extreme values of e.g. the λ_{31} and t-parameters were used did burn-in seem to be an issue.

We performed a maximum likelihood analysis of the alignment under the model described in section 2. Since the model is reversible, the likelihood is the product of the equilibrium frequency of one of the sequences times the probability of transition from this sequence to the other sequence. A two-step maximization procedure was used to find maximum likelihood estimates: we started by finding preliminary maximum likelihood estimates while using the parameter values above in the simulation measure (31) and 100 Gibbs samples to estimate each of the transition probabilities. We replaced the parameter values initially used for the simulation measure by the preliminary maximum likelihood estimates and then continued to search for points of higher likelihood by lowering the threshold in the search procedure, now basing each of the calculations of the transition probability on 10000 Gibbs samples. To see the efficiency of the Gibbs sampler we looked at the means and variances of the logarithm of (31) as a function of the number of Gibbs

	ElogL	VarlogL
10	5.96654e-01	1.111134e-03
100	5.96055e-01	2.214717e-04
1000	5.965890e-01	2.000380e-05
10000	5.975188e-01	2.942038e-06

Table 1: The mean and variances of the estimate of the logarithm of (31) as function of the number of Gibbs samples.

samples used in a calculation. The values of the simulation parameters were as described above, and the parameters in the q_{θ_1} measure were the maximum likelihood estimates found. The means and variances are given in Table 1. The variance decreases by a factor of ten as the number of Gibbs samples used to estimate the logarithm of (31) is increased by a factor of ten. As can be seen from Table 1 a typical value for the variance is of the order 10^{-6} when the calculation of the transition probability is based on 10000 Gibbs samples.

We next investigated the Gibbs samplers dependency upon parameter values used in the simulations, and on the random seed. We performed a second run, identical to the first run described above, but using a different random seed. In a third run we used the maximum likelihood estimates of the parameters obtained in the first run as simulation parameters and initial values in the q_{θ_1} measure. Finally, we examined the performance of the procedure when 'naive' values were used in the simulations and as starting values in the q_{θ_1} measure. The simulation and starting values in this run were $\pi_i^k = 0.25$, $i \in \{A, C, G, T\}$, $k = 1, 2, 3$, $\lambda_{12} = \lambda_{23} = \lambda_{31} = 1.0$, $K = 1.0$, $f = 1.0$ and $t = 0.1$. In this last run the maximum likelihood estimates found after the initial maximization were used as simulation parameters in the more refined maximization. Maximum likelihood estimates obtained from the four runs are given in Table 2.

The outcome of the maximization procedure does not appear to depend upon the parameter values used in the simulation measure, nor on the random seed. The maximum likelihood estimates of the parameters found in the four runs differ by no more than a few percent. The time required for the maximization, however, was considerably longer in the fourth run – approximately 10 times that of the other runs. From a time requirement point of view it is clearly worthwhile to perform an analysis of the stationary distribution of

	π_A^1	π_C^1	π_G^1	π_T^1		
1:	.296	.200	.325	.179		
2:	.296	.199	.323	.181		
3:	.296	.199	.324	.181		
4:	.294	.196	.326	.184		
	π_A^2	π_C^2	π_G^2	π_T^2		
1:	.333	.238	.235	.194		
2:	.334	.238	.234	.194		
3:	.334	.238	.234	.194		
4:	.335	.242	.229	.194		
	π_A^3	π_C^3	π_G^3	π_T^3		
1:	.422	.184	.232	.162		
2:	.421	.184	.233	.163		
3:	.421	.184	.232	.163		
4:	.424	.184	.230	.162		
	λ_{12}	λ_{23}	λ_{31}	K	f	t
1:	.371	.293	.321	4.382	.295	.0987
2:	.374	.293	.322	4.506	.299	.0962
3:	.368	.294	.326	4.548	.299	.0960
4:	.377	.289	.321	4.374	.286	.1002

Table 2: Maximum likelihood estimates of the parameters under four different runs of the Gibbs sammpler.

the sequences considered prior to starting the full estimation scheme, and to use the maximum likelihood values of the parameters thus found, in the Gibbs sampler.

As the ratio of transition probabilities can be calculated using the MCMC procedure, likelihood ratio tests can be performed. We performed a test for the hypothesis that $\lambda_{12} = \lambda_{23}$, $\lambda_{31} = 1.0$, under the full model, by finding maximum likelihood estimates under each of the models, and then evaluating the log likelihood ratio test statistic

$$-2\log Q = \log(\pi_{\theta_1}(x)) - \log(\pi_{\theta_2}(x)) + \log(\frac{P_{\theta_{1,t}}(y|x)}{P_{\theta_{2,t_0}}(y|x)})$$

using the MCMC procedure for the last term. The -2logQ value for this test was 34.23 and the restricted model was thus clearly rejected. We then

performed a test for the hypothesis that $\lambda_{12} = \lambda_{23} = \lambda_{31}$, using a similar procedure. In this case the test statistic was 0.93 and the restricted model accepted. We thus conclude that in the HIV1 gene *gag* CG depression is a phenomenon that occurs at all dinucleotide positions, and that it does so with the same strength. We further illustrate the effect of extending the model to allow for the CG depression across codon boundaries by plotting a profile loglikelihood curve for the parameter λ_{31}, under the full model restricted to have $\lambda_{12} = \lambda_{23}$.

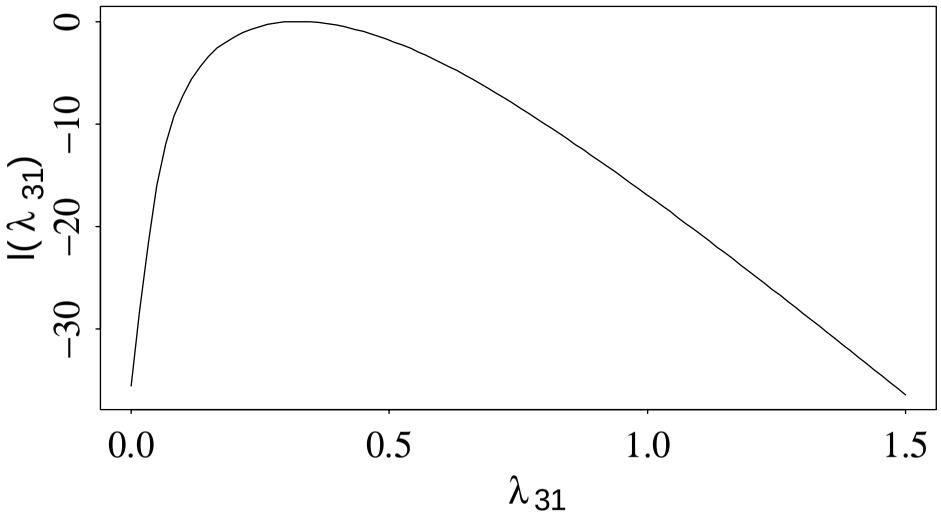

Figure 1: Plot of the profile log likelihood for the parameter λ_{31}.

8 Discussion

The simulation technique for estimating the transition probability between sequences developed in this paper provides a tool for analyzing DNA sequences under models with an additional level of complexity compared to those that are currently available. In particular we have focussed on a model in which instantaneous rates of substitutions at a site depend upon the states at sites in its neighbourhood. Another example where such a tool is called for is in the analysis of DNA sequences in which multiple genes are encoded in

different reading frames — a feature that is common for viral genomes. It is straight forward to extend the model and Gibbs sampler described in section 5 and 6 to allow for the incorporation of functional constraints induced by multiple overlapping reading frames. Moreover, the technique should prove useful in the analysis of sequences with (known) secondary and/or tertiary structures and structure requirements.

The relation between the stationary distribution of a sequence and the class of intensities that is consistent with it, is an important tool in the modelling process. Since it is much more easy to analyze a single sequence, we must be able to incorporate features extracted from single sequences in the Markov process of DNA sequence evolution. Or reversing the process: simple tests for the adequacy of some aspects of the model can be made on single sequences.

9 Acknowledgement

We thank an anonymous referee for helpful comments.

10 References

FELSENSTEIN, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

FELSENSTEIN, J., and CHURCHILL, G. (1996) A hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 93–104.

GEYER, C. J. (1996) Likelihood inference for spatial point processes. In *Proceedings Seminaire Europeen de Statistique, Stochastic Geometry, Likelihood and computation.* O. Barndorff-Nielsen, W.S Kendall and MNM. van Lieshout (Eds), Chapman and Hall.

GOLDMAN, N. (1993) Statistical tests of models of DNA substitutions. *J. Mol. Evol.* **36**, 182–198.

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

HAESELER, A. von, and SCHÖNIGER, M. (1998) Evolution of DNA or amino acid sequences with dependent sites. *J. Comp. Biol.* **5**, 149–163.

KYPR, J., MRAZEK, J., and REICH, J. (1989) Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. *Biochim. Biophys. Acta.* **1009**, 280–282.

MUSE, S. V., and GAUT, B. S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 715–724.

PEDERSEN, A. K., WIUF, C., and CHRISTIANSEN, F.B. (1998) A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15**, 1069–1081.

THORNE, J., KISHINO, H., and FELSENSTEIN, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124.

YANG, Z. (1993) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites. *Mol. Biol. Evol.* **10**, 1396–1401.