A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames

Anne-Mette Krabbe Pedersen and Jens Ledet Jensen

Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Denmark

We present a model and methodology for the maximum-likelihood analysis of pairwise alignments of DNA sequences in which two genes are encoded in overlapping reading frames. In the model for the substitution process, the instantaneous rates of substitution are allowed to depend on the nucleotides occupying the sites in a neighborhood of the site subject to substitution at the instant of the substitution. By defining the neighborhood of a site to extend over all sites in the codons in both reading frames to which a site belongs, constraints imposed by the genetic code in both reading frames can be taken into account. Due to the dependency of the instantaneous rates of substitution on the states at neighboring sites, the transition probability between sequences does not factorize and therefore cannot be obtained directly. We present a Markov chain Monte Carlo procedure for obtaining the ratio of two transition probabilities between two sequences under the model considered, and we describe how maximum-likelihood parameter estimation and likelihood ratio tests can be performed using the procedure. We describe how the expected numbers of different types of substitutions in the shared history of two sequences can be calculated, and we use the described model and methodology in an analysis of a pairwise alignment of two hepatitis B sequences in which two genes are encoded in overlapping frames. Finally, we present an extended model, together with a simpler approximate estimation procedure, and use this to test the adequacy of the former model.

Introduction

In Felsenstein's maximum-likelihood framework for inferring evolutionary trees from DNA sequences, a fundamental assumption is that the substitution processes in the single-nucleotide sites are independent (Felsenstein 1981). When this is the case, transition probabilities between sequences can feasibly be obtained, because these probabilities become products of transition probabilities between nucleotides. If the independent substitution processes in the sites are assumed to be identical Markov processes described by a matrix of instantaneous rates Q^{nuc}, the matrix of transition probabilities between nucleotides separated by time t can be obtained as $P^{\text{nuc}}(t) = \exp(Q^{\text{nuc}}t)$.

For many sequences, the assumption of independent substitution processes in the nucleotide sites is in striking contradiction to biological reality. Protein-coding sequences present an obvious example: the rate of synonymous substitution is generally higher than that of nonsynonymous substitution (see Li, Wu, and Luo 1985 and references therein). Whether a substitution in a site is synonymous or nonsynonymous depends on what nucleotides occupy the other sites of the codon. Substitution processes in nucleotide sites belonging to the same codon are thus nonindependent.

Li, Wu, and Luo (1985) were among the first to describe a method for estimating evolutionary distances between coding sequences in which constraints imposed by the structure of the genetic code were taken into account. Their method relied on a partitioning of sites into degeneracy classes. A site was defined to be nondegenerate if all possible changes at the site were nonsynonymous, twofold-degenerate if one was synonymous and the other two were nonsynonymous, and fourfold-degenerate if all possible changes were synonymous. The classification of sites into degeneracy classes was based on one of the observed sequences, and the degeneracy class of a site was assumed to be constant over time. Having defined the degeneracy classes of the sites, the sequences were analyzed under the assumption that the substitution processes in the nucleotide sites were independent. This method is approximate: the classification of sites into degeneracy classes depends on which sequence one chooses to base the classification on, and the degeneracy class of a site is not constant over time, but changes as substitutions occur.

Goldman and Yang (1994) and Muse and Gaut (1994) described how the nonindependence introduced by the structure of the genetic code could be dealt with in an exact manner. They presented codon-based models in which the substitution processes in codons, rather than single-nucleotide sites, were assumed to be independent. The substitution processes in the codons were assumed to be identical, reversible Markov processes, described by a 61×61 matrix of instantaneous rates of codon substitution, Q^{codon}. In the matrix, entries corresponding to nonsynonymous substitutions could be modified relative to synonymous ones by multiplication of a factor representing the fractional reduction of amino-acid-altering relative to amino-acid-preserving rates. Due to the assumption of independent substitution processes in codons, transition probabilities between sequences in codon-based models factorize into a product of transition probabilities between codons. As in nucleotide-based models, these can be obtained by taking the exponential of the product of the rate matrix and the time, i.e., $P^{\text{codon}}(t) = \exp(Q^{\text{codon}}t)$. The basic idea behind codon-based models has since been utilized in the de-

Key words: overlapping reading frames, substitution process, dependent substitution rates, MCMC, hepatitis B.

Address for correspondence and reprints: Anne-Mette Krabbe Pedersen, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark. E-mail: annemet@imf.au.dk.

Mol. Biol. Evol. 18(5):763–776. 2001
 ISSN: 0737-4038

velopment of models for substitution processes in RNA sequences. Schöninger and von Haeseler (1994), Muse (1995), and Tillier and Collins (1995) have presented dinucleotide-based models for the analysis of RNA sequences that allow dependencies among the substitution processes in nucleotide sites that participate in base-pairings to be modeled. The substitution processes in dinucleotides are assumed to be independent, and the transition probability between sequences factorizes into a product of transition probabilities between dinucleotides.

The substitution processes in sequences in which more genes are encoded in overlapping reading frames are subject to constraints imposed by overlapping genetic codes. Due to the overlapping of the constraints, an assumption of independent substitution processes in small subsequences is inappropriate. An adequate description of the substitution process in these sequences can therefore not be obtained by exploiting the idea behind the codon-based models. A model for the substitution processes in sequences with multiple overlapping reading frames was suggested by Hein and Støvlbæk (1995), who extended the notion of the degeneracy class of a site to that of a combination of degeneracy classes (one for each reading frame to which a site belongs). They defined class-specific matrices of instantaneous rates of substitution, assumed independent Markov processes in the sites according to these matrices, and illustrated how a maximum-likelihood analysis of evolutionary trees under the model could be performed. As in the method of Li, Wu, and Luo (1985), the classification of sites was based on one of the observed sequences, and the class of a site was assumed to be fixed. The method thus inherits the shortcomings of Li, Wu, and Luo's (1985) method and deals with the constraints imposed by the overlapping genetic codes in an approximate manner.

In this study, we present a model for the substitution process in sequences in which two genes are encoded in overlapping frames that incorporates the constraints imposed by both of the overlapping genetic codes in an exact manner. This is achieved by allowing the instantaneous rates of substitution in a site to depend on what nucleotides occupy the sites in the neighborhood of the site at the instant of the substitution. Thus, the model does not rely on the degeneracy class notion and does not assume that the substitution processes in any subsequences are independent. Rather, the model contains parameters representing the degrees of selectional constraints operating in the different frames, and these parameters can be estimated. Due to the nonindependent instantaneous rates of substitution, the transition probability between two sequences does not factorize into products of transition probabilities between small subsequences. Rather, transition probabilities between full-length sequences must be considered. The model and methodology we describe here were obtained by generalizing a model and methodology we have previously presented (Jensen and Pedersen 2000).

In the *Methods* section, we describe the model for the substitution process in sequences with overlapping reading frames. We show that the substitution process is reversible and derive the stationary distribution of a sequence under the model. We further describe a Markov chain Monte Carlo (MCMC) procedure for estimating the ratio of two transition probabilities between two sequences under the described model. Together, these entities, the stationary distribution of a single sequence and the ratio of transition probabilities between two sequences, comprise the elements needed for a maximum-likelihood analysis to be performed. In the *Results* section, we analyze an alignment of two homologous hepatitis B subsequences in which the polymerase (P) and the envelope (S) genes are encoded in overlapping frames using the model and methodology presented in the preceding section. We obtain maximum-likelihood estimates of the parameters in the model and perform a likelihood ratio test of a hypothesis concerning the mode of substitution in the sequences. We calculate expected numbers of various types of substitutions. Finally, to check the adequacy of the model, we present a more general model, together with a simpler approximate estimation procedure.

Methods

The Model

In this section, we present a model for the substitution process in sequences with overlapping reading frames, in which constraints imposed by the two overlapping genetic codes are incorporated.

We consider an alignment of two homologous DNA sequences in which two genes are encoded in overlapping reading frames. We assume that the sequences have evolved from a common ancestral sequence through independent identical evolutionary processes that involve substitutions only. We further assume that the substitution process is a homogeneous Markov process and that substitutions happen sequentially, so that within an instant, the sequence may be changed in one nucleotide position only. We do not allow substitutions that generate stop codons in any of the reading frames considered, and we assume that no substitutions occur in the first and last codons of reading frame I in the alignment. With θ short for the set of parameters specifying the substitution process, the likelihood of observing the two sequences x and y at the tips of an evolutionary tree with branch lengths t_x and t_y is given by

$$L(\theta, t_x, t_y) = \sum_{s_0} p_{s_0}(\theta) P_{s_0 \to x}(\theta, t_x) P_{s_0 \to y}(\theta, t_y), \quad (1)$$

where the sum over s_0 is over all possible ancestral sequences and $p_{s_0}(\theta)$ is the probability under the model of the ancestral sequence being s_0. The parameters t_x and t_y are the time epochs separating sequences x and y respectively, from the ancestral sequence, and $P_{s_0 \to z}(\theta, t_z)$ is the transition probability between sequences s_0 and z, $z = x, y$.

We still have to specify the precise form of the Markov process in the inner parts of the sequences, that is, in codons $2, \ldots, n - 1$, in a way that permits the

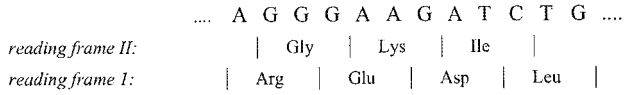

FIG. 1.—Example of a sequence in which two genes are encoded in overlapping reading frames.

constraints imposed by the two overlapping genetic codes to be incorporated. For this, consider a sequence in which two genes are encoded in overlapping reading frames. The substitution of a nucleotide in the sequence may lead to the alteration of the amino acid being encoded in both of the reading frames, in one of the reading frames only, or in none of the reading frames. Assume that the reading frames overlap as illustrated in figure 1. Whether a nucleotide substitution is synonymous or nonsynonymous with respect to one of the reading frames depends on what nucleotides occupy the other positions of the codon within that reading frame at the instant of the substitution. Whether a substitution in the first codon position in reading frame I is synonymous or nonsynonymous with respect to reading frame II depends on what nucleotides occupy the two immediately preceding nucleotide positions, that is, positions 2 and 3 of the preceding reading frame I codon. In order to determine whether a substitution in codon position 2 or 3 in reading frame I is synonymous or nonsynonymous with respect to reading frame II, the nucleotide immediately following the codon, that is, in position 1 of the next codon in reading frame I, must be known.

In order to incorporate constraints imposed by the operation of two overlapping reading frames in the model for the substitution process, we must allow the instantaneous rates of nucleotide substitution to be context-dependent. Numbering the positions by the codon number in reading frame I, the instantaneous rate of substitution of one of the three nucleotides in codon i should depend on the nucleotides occupying positions 2 and 3 of codon $i - 1$ and position 1 of codon $i + 1$. Let $z_i = (z_i^1, z_i^2, z_i^3)$ denote the ith codon in reading frame I of the inner part of a sequence ($i = 2, \ldots, n - 1$), where z_i^k is the nucleotide in codon position k, $k = 1, 2, 3$. Let $\tilde{z}_i$ be a codon that differs from $\tilde{z}_i$ in one nucleotide position only. In order to allow for unequal nucleotide frequencies, we assume that the instantaneous rate of substitution to codon $\tilde{z}_i$ is proportional to $\pi_{(\tilde{z}_i)}$, where $(\tilde{z}_i)$ is the target nucleotide, that is, the nucleotide occupying the position in $\tilde{z}_i$ at which it differs from z_i. We assume that the π_k's, $k \in \{A, C, G, T\}$, sum to 1. We further allow for transition/transversion bias by multiplying instantaneous rates of transitional substitutions by the factor K. With respect to these two features, our model is similar to that of Hasegawa, Kishino, and Yano (1985). The constraints imposed by the genetic codes are incorporated by multiplying all instantaneous rates that alter the amino acid in reading frame I only by f_{I}, those that alter the amino acid in reading frame II only by f_{II}, and those that alter the amino acids in both reading frames by $f_{\mathrm{I/II}}$, a procedure related to that used in the codon-based model for single coding sequences (Goldman and Yang 1994; Muse and Gaut 1994). We refer to the f parameters (f_{I}, f_{II}, and $f_{\mathrm{I/II}}$) as parameters for selective constraints. An f parameter larger than 1 indicates that amino-acid-altering substitutions in the associated reading frame are promoted, whereas if $f < 1$, synonymous substitutions are favored. When $f = 1$, there are no selective constraints in the corresponding reading frame. Note that using reading frame I for numbering the positions along the sequence has no influence on the instantaneous rates. Whether a substitution alters the amino acid in one of the reading frames is not related to the numbering used.

Let $q_{z_i,\tilde{z}_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1}$ denote the instantaneous rates of substitution from a sequence that has z_i as the ith codon in reading frame I to a sequence that is identical to the sequence except in codon i, in which it holds the codon $\tilde{z}_i$, at an instant when positions 2 and 3 of codon $i - 1$ and position 1 of codon $i + 1$ are z_{i-1}^2, z_{i-1}^3, and z_{i+1}^1, respectively. The model then states the following form for the instantaneous rates of substitution:

$$q_{z,\tilde{z}} = \begin{cases} 0 & \text{if } z \text{ and } \tilde{z} \text{ differ in more than one position,} \\ q_{z_i,\tilde{z}_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1} & \\ & \text{if } z \text{ and } \tilde{z} \text{ differ at one position in codon } i, \end{cases}$$

where

$$\begin{aligned} &q_{z_i,\tilde{z}_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1} \\ &\quad = \pi_{(\tilde{z}_i)} M((z_{i-1}^2, z_{i-1}^3, z_i, z_{i+1}^1), z_{i-1}^2, z_{i-1}^3, \tilde{z}_i, z_{i+1}^1)) 1_C(\tilde{z}_i) \\ &\qquad \times 1_C(z_{i-1}^2, z_{i-1}^3, \tilde{z}_i^1) 1_C(\tilde{z}_i^2, \tilde{z}_i^3, z_{i+1}^1), \end{aligned} \tag{2}$$

with

$$\begin{aligned} &M((z_{i-1}^2, z_{i-1}^3, z_i, z_{i+1}^1), (z_{i-1}^2, z_{i-1}^3, \tilde{z}_i, z_{i+1}^1)) \\ &\quad = K^{1_{\mathrm{ts}}} f_{\mathrm{I}}^{1_{\mathrm{non(I),syn(II)}}} f_{\mathrm{II}}^{1_{\mathrm{syn(I),non(II)}}} f_{\mathrm{I/II}}^{1_{\mathrm{non(I),non(II)}}}. \end{aligned}$$

Here, 1_{ts} is 1 for a transition (ts) and 0 for a transversion (tv), $1_{\mathrm{non(I),syn(II)}}$ is 1 for a substitution that is nonsynonymous in reading frame I and synonymous in reading frame II and 0 otherwise, and $1_{\mathrm{syn(I),non(II)}}$ and $1_{\mathrm{non(I),non(II)}}$ are defined similarly. The set C, on which the indicator functions are 1 in equation (2), consists of the 61 nonstop codons. Spelled out, we obtain the following representation of the instantaneous rates:

$$q_{z_i,\tilde{z}_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1} = \begin{cases} 0, & \text{STOP} \\ K\pi_{(\tilde{z}_i)}, & \text{no STOP, ts, syn(I), syn(II)} \\ \pi_{(\tilde{z}_i)}, & \text{no STOP, tv, syn(I), syn(II)} \\ f_{\mathrm{I}} K\pi_{(\tilde{z}_i)}, & \text{no STOP, ts, non(I), syn(II)} \\ f_{\mathrm{I}}\pi_{(\tilde{z}_i)}, & \text{no STOP, tv, non(I), syn(II)} \\ f_{\mathrm{II}} K\pi_{(\tilde{z}_i)}, & \text{no STOP, ts, syn(I), non(II)} \\ f_{\mathrm{II}}\pi_{(\tilde{z}_i)}, & \text{no STOP, tv, syn(I), non(II)} \\ f_{\mathrm{I/II}} K\pi_{(\tilde{z}_i)}, & \text{no STOP, ts, non(I), non(II)} \\ f_{\mathrm{I/II}}\pi_{(\tilde{z}_i)}, & \text{no STOP, tv, non(I), non(II).} \end{cases}$$

The instantaneous rate of a substitution which changes a codon ACA to a codon GCA in reading frame I at an instant when the codon considered is preceded by

the nucleotides TC and followed by the nucleotide A is thus $\pi_G K f_I$, since the substitution is a transition to a G that changes the amino acid coded for in reading frame I from a threonine to an alanine and does not change the amino acid encoded in reading frame II, since both of the codons TCA and TCG code for serines. The instantaneous rate of a substitution which changes a codon GAT in the context CC|...|A to GAA is $\pi_A f_{I/II}$, since the substitution is a transversion to an A and the codons ATA and AAA (in reading frame II) code for different amino acids, as do the codons GAT and GAA (in reading frame I). Note that the model can easily be modified to other kinds of overlapping genes, e.g., genes encoded in opposite directions. The only modification needed is in the translation via the genetic code.

The Stationary Distribution and Reversibility

We assume that the Markov process has reached equilibrium and let $\Pi(z)$ denote the equilibrium frequency of a sequence z. In this section, we show that the model presented above is reversible, identify the stationary distribution of a sequence under the model, and give a quick procedure for calculating the normalizing constant of the stationary distribution.

A Markov process with instantaneous rates $q_{z,\tilde{z}}$ is reversible and has Π as the stationary distribution if

$$\Pi(z)q_{z,\tilde{z}} = \Pi(\tilde{z})q_{\tilde{z},z}.$$

Under the model described above, the equilibrium frequency of a sequence with a stop codon in any of the two reading frames is 0, as are instantaneous rates of substitutions that generate stop codons. The above equality is thus satisfied for these cases. Moreover, the equality is trivially satisfied for sequences z and $\tilde{z}$ that differ in more than one nucleotide position, as in this case, the instantaneous rates are 0. Therefore, assume that the two sequences z and $\tilde{z}$ do not contain stop codons in either reading frame and differ at only one codon position in codon i, and consider

$$\Pi(z)q_{z_i,\tilde{z}_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1} = \Pi(\tilde{z})q_{\tilde{z}_i,z_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1}.$$

Since all factors in the instantaneous rates except $\pi_{(\tilde{z}_i)}$, which depends on the target nucleotide $(\tilde{z}_i)$, appear symmetrically, they cancel out, and we obtain

$$\Pi(z)\pi_{(\tilde{z}_i)} = \Pi(\tilde{z})\pi_{(z_i)}.$$

It is easily seen that this equality is satisfied if the equilibrium distribution of a sequence $\Pi(z)$ is a product over the π_k parameters corresponding to the nucleotide constituents of the sequence. Incorporating the exclusion of sequences with stop codons in the second reading frame, we obtain that under the model, the stationary distribution of a sequence $z = (z_2, \ldots, z_{n-1})$ with z_1 and z_n fixed is

$$\Pi(z) = \frac{1}{Z}\left(\prod_{i=2}^{n-1} \pi_{z_i^1}\pi_{z_i^2}\pi_{z_i^3} 1_C(z_{i-1}^2, z_{i-1}^3, z_i^1)\right) 1_C(z_{n-1}^2, z_{n-1}^3, z_n^1), \quad (3)$$

if $z_i \in \mathcal{C}$, $\forall i$. For all other sequences, the stationary distribution is 0. Here, Z is a normalizing constant different from 1, because sequences with stops in either reading frames are excluded. Without this exclusion, Z would indeed be 1, because the π_k's sum to 1. We have thus identified the equilibrium distribution and at the same time shown that the process is reversible. With reversibility, the likelihood value becomes independent of the placement of the root, and with this and the assumed equilibrium, the likelihood in equation (1) reduces to

$$L(\theta, t) = \Pi_x(\theta)P_{x\to y}(\theta, t), \quad (4)$$

where we write $\Pi_x(\theta)$ for $\Pi(x)$ to stress that in the likelihood $\Pi(x)$ is treated as a function of the parameters in the model, of which only θ, and not the branch length(s), is relevant for the stationary distribution, and t is the sum of the branch lengths t_x and t_y.

In order to calculate the likelihood value eq. (4), we must be able to calculate the value of the normalizing constant Z of the stationary distribution eq. (3). This normalizing constant can be found by summing up the equilibrium frequencies of all possible sequences. By first summing over (z_i^2, z_i^3), $i = 2, \ldots, n-1$, we can derive an explicit form for Z. The details are in appendix A, where we end up with the formula (A.9)

$$Z = (c_1\lambda_1^{n-2}\nu_1^{z_1^1} + c_2\lambda_2^{n-2}\nu_2^{z_1^1})\frac{1}{\pi_{z_n^1}},$$

where all the terms are defined in appendix A.

Calculation of the Transition Probability Between Two Sequences

We now specify how the transition probability from a sequence x to a sequence y under the model described above can be calculated. Since the instantaneous rates of substitution under the model depend on the states at neighboring sites at the instant of the substitution, the probability of transition between the full sequences does not reduce to a product of ''marginal'' transition probabilities, such as a product of transition probabilities between nucleotides or codons. The substitution processes in all the sites must be considered simultaneously. As argued in Jensen and Pedersen (2000), we will have to resort to simulations in order to calculate the transition probability. Furthermore, in order to reduce the variance of the simulated values, we simulate the ratio of two probabilities $P_{x\to y}(\theta_1, t_1)/P_{x\to y}(\theta_2, t_2)$ instead of simulating a transition probability directly. If the ratio can be evaluated for two sets of parameter values, likelihood ratio tests can be obtained, and maximum-likelihood estimates of the parameters in the model can be found by maximizing the ratio

$$\frac{\Pi_{\theta_1}(x)P_{x\to y}(\theta_1, t_1)}{\Pi_{\theta_2}(x)P_{x\to y}(\theta_2, t_2)}$$

as a function of (θ_1, t_1) for fixed (θ_2, t_2). In the following, we describe a procedure for obtaining the ratio of two

transition probabilities using an MCMC simulation technique.

Let χ_t be the space of paths between sequences x and y separated by time t, and let L denote a particular path in χ_t. A path is a specification of the number of substitutions, the positions in which the substitutions occur, what nucleotides replace existing nucleotides in these substitutions, and the times ($\in(0, t)$) at which the substitutions occur. Let μ_t be the measure on χ_t which, for a fixed number of substitutions r and fixed positions and fixed nucleotides of these substitutions, corresponds to ordinary integration on the space $(0, t)^r$ for the substitution times. For a positive number s, we denote by $(s/t)L$ the path in χ_s obtained by scaling all the substitution times in L by s/t. Let $q_\theta(t; L)$ be the contribution from the path L to the transition probability $P_{x\to y}(\theta, t)$. A detailed description of q is given in appendix B.

In appendix B, we derive the representation

$$\frac{P_{x\to y}(\theta_1, t_1)}{P_{x\to y}(\theta_2, t_2)} = \tilde{E}\left(\frac{t_1^r q_{\theta_1}\left(t_1; \frac{t_1}{t_2}L\right)}{t_2^r q_{\theta_2}(t_2; L)}\right), \tag{5}$$

where r is the number of substitutions in the path L, and $\tilde{E}$ denotes the mean value under the distribution $\tilde{P}$ on the space χ_{t_2} having density

$$\frac{q_{\theta_2}(t_2; L)}{\int_{\chi_{t_2}} q_{\theta_2}(t_2; L)\, d\mu_{t_2}} \tag{6}$$

with respect to μ_{t_2}. We can thus obtain an approximation of the ratio of the two transition probabilities by calculating $[t_1^r q_{\theta_1}(t_1; (t_1/t_2)L)]/[t_2^r q_{\theta_2}(t_2; L)]$ for a large number of paths L drawn from $\tilde{P}$. The farther apart (θ_1, t_1) and (θ_2, t_2) are, the larger the variance and the more paths are needed for the ratio to be obtained with reasonable precision. It is thus necessary when maximizing as a function of (θ_1, t_1) to alter the parameters in the simulation measure (θ_2, t_2) as (θ_1, t_1) moves away from (θ_2, t_2).

We now specify how to simulate from equation (6). We use an MCMC method (Gilks, Richardson, and Spiegelhalter 1996), that is, we construct a Markov chain on the path space χ_{t_2} that has $\tilde{P}$ as its stationary distribution. A path L is the collection of paths L_i^j of the nucleotides in the jth codon position of codon number i in reading frame I. We construct the Markov chain by running through the codons from number 2 to number $n-1$ while we update the path L_i for the ith codon. The updating of L_i is done by proposing a new path L_i' from a distribution P_i with density q_i. The new path L_i' is accepted with probability

$$\alpha = \min\left(1, \frac{\tilde{q}_{\theta_2}(L_i' \,|\, L_{i-1}, L_{i+1})/q_i(L_i')}{\tilde{q}_{\theta_2}(L_i \,|\, L_{i-1}, L_{i+1})/q_i(L_i)}\right), \tag{7}$$

where $\tilde{q}_{\theta_2}$ is given in appendix B. From a computational-cost point of view, the important thing here is that $\tilde{q}_{\theta_2}$ depends on L_i (or L_i') and the two neighboring paths L_{i-1} and L_{i+1} only. Having completed a run through the alignment, we have performed a transition in our Markov chain on sequence paths. By continuing the procedure many times, we obtain a sample of sequence paths which contains paths throughout the support of equation (6) in the correct proportions. In particular, if we propose a path L_i' that gives rise to a stop codon in reading frame II, then $\tilde{q}_{\theta_2}(L_i' \,|\, L_{i-1}, L_{i+1})$ will be zero, and therefore the path is not accepted.

The choice of an initial path L to start the Markov chain is not important. We have obtained a start path by simulating paths L_i from P_i, $i = 2, \ldots, n-1$, and continuing until a sequence path without stop codons in reading frame II has been obtained. The exact form of the path proposal distribution for codon i, q_i, is given by the following three steps:

1. The number of substitutions k_i is taken from a modified Poisson distribution with intensity γ_i.
2. The substitution times $t_i(r)$, $r = 1, \ldots, k_i$, are taken from a uniform distribution on the interval from 0 to t_2.
3. The nucleotide position and the new nucleotide for each substitution is chosen from a set of allowed substitutions $\mathcal{A}_r$.

As for the modification of the Poisson distribution in step 1, note that if the ith codons in the two sequences are identical, paths with one substitution are impossible. If the codons differ at d_0 positions, there must be at least d_0 substitutions in the path leading from one codon to the other. We thus modify the Poisson distributions and propose a number of substitutions for the path between the ith codons from

$$p_i(k) = P(N_i = k) = \begin{cases} \dfrac{\gamma_i^k}{k!}e^{-\gamma_i}/(1-e^{-\gamma_i}) & \\ \qquad x_i = y_i, \qquad k = 0, 2, 3, \ldots, & \\ \dfrac{\gamma_i^k}{k!}e^{-\gamma_i}\Big/\left(1-\displaystyle\sum_{l=0}^{d_0^i-1}\frac{\gamma_i^l}{l!}e^{-\gamma_i}\right) & \\ \qquad x_i \neq y_i, \qquad k \geq d_0^i, & \end{cases}$$

where d_0^i is the number of codon positions at which the ith codons in the two sequences differ. The intensity γ_i is described below.

In both steps 1 and 3 we will use intensities $\tilde{q}_{y,w}^i$ of a change from a codon y to a codon w given by

$$\tilde{q}_{y,w}^i = \tilde{q}_{y,w|x_{i-1}^2, x_{i-1}^3, x_{i+1}^1},$$

where $\tilde{q}_{y,w|x_{i-1}^2,x_{i-1}^3,x_{i+1}^1}$, is defined as in equation (2), except that we treat substitutions to stop codons in the second reading frame $((x_{i-1}^2, x_{i-1}^3, w^1)$ or $(w^2, w^3, x_{i+1}^1))$ as substitutions to a 21st amino acid, rather than giving them intensity 0. We have adopted this approach in order to keep the proposal distribution simple; that is, we are not using the paths L_{i-1} and L_{i+1} during the proposal step. For the Poisson intensity γ_i in step 1, we take

$$\gamma_i = \left(\sum_{w\in C, w\neq x_i} \tilde{q}_{x_i,w}^i\right)t_2.$$

In step 3, let k_i be the number of substitutions from step

1, and let $z_i(r)$, $r = 0, \ldots, k_i$, be the codon after the rth substitution with $z_i(0) = x_i$ and $z_i(k_i) = y_i$. When $z_i(1), \ldots, z_i(r-1)$ have been chosen, we choose $z_i(r)$ according to the probabilities

$$P(z_i(r) = z) = \frac{\tilde{q}^i_{z_i(r-1),z}}{\sum_{w \in \mathcal{A}_r} \tilde{q}^i_{z_i(r-1),w}}.$$

The set of allowed substitutions for substitution number r, $\mathcal{A}_r$, can be described as follows: let $d_i(z)$ be the number of nucleotide positions at which codon z differs from codon y_i. Then,

$$\mathcal{A}_r = \{w \in \mathcal{C} \mid w \neq z_i(r-1), d_i(w) \leq k_i - r, d_i(w) = 1 \text{ if } r = k_i - 1\}.$$

The following examples illustrate the role of $\mathcal{A}_r$. Assume that for a potential path between identical codons ($x_i = y_i$), the number of substitutions chosen is two. For the first substitution, we may choose any combination of position and nucleotide, as long as the nucleotide chosen is different from the one that occupies that position at the moment, and as long as we do not create a stop codon. Irrespective of the choices for the first substitution, the choices for the second (and last) substitution are completely fixed, since we must get to the target codon via this substitution. Similarly, in a path containing one substitution between codons that differ in one position, the choices of position and nucleotide for the substitution are both completely fixed. For paths between codons that differ at one position for which we have chosen a number of two substitutions, the first must occur in the position at which the two codons differ (since otherwise it would generate an additional difference, making it impossible to get to the target codon with the remaining one substitution). Furthermore, the first substitution must not generate the target codon, since if it does, we have no way of assigning the last substitution. Note that we may generate a path for which the set of allowed substitutions for a certain substitution is empty. This is the case for a path with two substitutions between the codons TCA and TTA: the first substitution must be in position 2 (as otherwise we would generate an additional difference), it cannot be to a T (because then the target codon is reached prematurely), and it cannot be to an A or a G (as rates to the codons TAA and TGA are 0, since the codons are stop codons). A path for which an empty set of allowed substitutions is created will be discarded; that is, it will never be accepted.

With this procedure for generating paths, the density of proposing a path L_i for codon i is

$$q_i(L_i) = \left(p_i(k_i) \prod_{r=1}^{k_i-1} \frac{\tilde{q}_{z_i(r-1),z_i(r)}}{\sum_{w \in \mathcal{A}_r} \tilde{q}_{z_i(r-1),w}} \right) k_i! \left(\frac{1}{t_2} \right)^{k_i}.$$

Having obtained expressions for the equilibrium frequency of a sequence under the model, and here an approximation of the likelihood ratio of transition probabilities, we have the means for performing maximum-likelihood estimation and likelihood ratio tests.

Results

Below, we present results from a maximum-likelihood analysis of a pairwise alignment of two hepatitis B sequences, in which the model and methodology presented in the *Methods* section were used. We refer to the model presented above as ''the full model''. We give maximum-likelihood estimates of the parameters in the full model and perform a likelihood ratio test of a model of multiplicatively operating selective constraints (referred to as ''the multiplicative model'') under the full model. The multiplicative model is accepted. In the subsequent subsection, we show how the expected numbers of various types of substitutions may be calculated and give the values obtained under the multiplicative model. In the last subsection, we present an extension of the full model, referred to as the ''extended model,'' that we use to check the adequacy of the full model. For the extended model, we describe and use a simpler, but approximate, estimation method than that used for the full and multiplicative models.

Maximum-Likelihood Analysis

The hepatitis B viral genome is circular and partially double-stranded, with the longer strand consisting of approximately 3,200 nt (Ganem 1996).

Every nucleotide in the genome is within a coding region, and more than half of the sequence is translated in more than one reading frame. The genome has four open reading frames (ORFs): P, C, S, and X. The P ORF encodes the viral polymerase, the C ORF encodes the structural protein of the nucleocapsid, the X ORF encodes a putative regulatory protein, and the S ORF encodes the viral surface glycoproteins. The S ORF is completely embedded in the P ORF, and the C and X genes partially overlap with the P ORF and also themselves partially overlap (see fig. 2).

Two full genome sequences were obtained from the GenBank database (http://ncbi.nlm.nih.gov/genbank) (accession numbers AF151735 [type ayw2] and X75663 [type adw4q]). An alignment of the parts of the sequences in which the S (surface) and P (polymerase) genes are encoded in overlapping frames was obtained automatically using GENAL (Hein and Støvlbæk 1994). The alignment exhibited an insertion of 11 consecutive codons after the first 2 codons in the reading frame of the S gene. Analysis was restricted to the region following the insertion, a region spanning 1,152 nt, or 384 codons, in the P gene reading frame. The paths of the first and last codons in this region were assumed to be fixed with no substitutions. In the region analyzed, the sequences differed at 13% of the nucleotide positions. The 150 differing nucleotide positions were distributed among 119 reading frame I codons, and 78 transitional differences were exhibited, with the remaining 72 being transversional differences. Among the differing nucleotide positions, 70 fell in first codon positions in reading frame I (third codon positions in reading frame II), 32

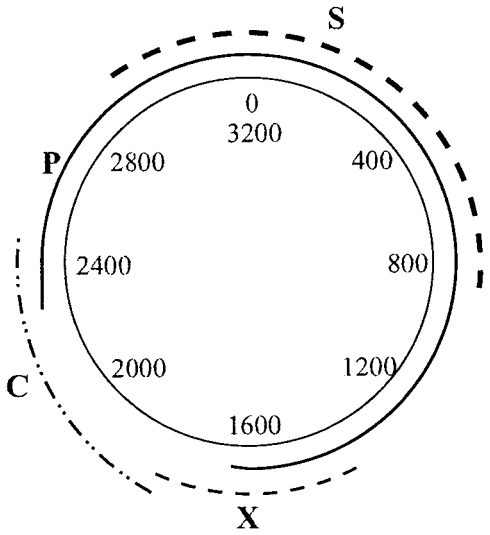

FIG. 2.—Schematic representation of the circular genome of hepatitis B. Approximate locations of the four open reading frames C, P, X, and S are shown.

in second codon positions in reading frame I (first codon positions in reading frame II), and 48 in third codon positions of reading frame I (second codon positions of reading frame II).

The MCMC algorithm was implemented in C. The Markov chain appeared to converge quickly toward its stationary distribution (results not shown). In the maximization of the likelihood ratio as a function of (θ_1, t_1), we used a stepwise procedure: we started with a certain set of initial values for the parameters for the simulation measure (θ_2, t_2). We performed a first rough maximization (round I) with high threshold (10^{-4}) in which 10,000 samples were used in each calculation of the ratio of the two transition probabilities. We used the resulting ''rough'' maximum-likelihood estimates as new values of (θ_2, t_2) in a new round of maximization (round II) with a lower threshold (10^{-5}). In this round, each evaluation of the likelihood value was based on 100,000 samples. We then proceeded to round III, in which we used the obtained parameter estimates from round II as new values of (θ_2, t_2), while the number of samples used for evaluation of the ratio of transition probabilities was kept at 100,000, and the threshold was again lowered (10^{-6}). In each round, the starting values of (θ_1, t_1) were set equal to those of the simulation measure (θ_2, t_2).

In order to examine the efficiency and dependency of the MCMC sampler and maximization scheme on the starting values of the parameters, we compared the outcomes of four different runs. Runs A, B, and C had initial values of $\pi_k = 0.25$, $k \in \{A, C, G, T\}$, but the initial values for the remaining parameters varied. Initial parameter values of run C* were identical to those of run C:II to the first four decimals, but the runs were started with different random seeds. Initial parameter values and maximum-likelihood estimates obtained after each round in each of the four runs are given in table 1. Also given are the numbers of iterations (full parameter vector updatings) used in the maximization procedure (Powell's method; Press et al. 1992) to reach the given maximum-likelihood estimates. The intensities are scaled so that, at equilibrium, the expected number of substitutions out of a codon is 1, and thus the parameter t gives the expected number of substitutions in total per codon between the two sequences. How this scaling is obtained is described in the next section.

The maximum-likelihood estimates obtained in the four different runs were similar, and the results did not indicate any dependency of the Gibbs sampler on the starting values for the procedure (rows A:III, B:III, and C:III), nor was the result sensitive to the

Table 1
Maximum-Likelihood Estimates of the Parameters in the Full Model After each of Three Rounds in Four Different Runs

Run	Round[a]	$\hat{\pi}_A$	$\hat{\pi}_C$	$\hat{\pi}_G$	$\hat{\pi}_T$	$\hat{t}$	$\hat{K}$	$\hat{f}_P$	$\hat{f}_S$	$\hat{f}_{P/S}$	No. Iter[b]
A.....	Start	0.250	0.250	0.250	0.250	0.527	1.200	0.200	0.200	0.050	
	I	0.244	0.267	0.218	0.271	0.478	1.714	0.212	0.151	0.062	2
	II	0.237	0.270	0.221	0.272	0.460	1.660	0.214	0.151	0.062	2
	III	0.238	0.269	0.221	0.272	0.455	1.620	0.213	0.152	0.061	2
B.....	Start	0.250	0.250	0.250	0.250	0.341	2.000	0.500	0.500	0.500	
	I	0.239	0.274	0.207	0.280	0.442	1.449	0.560	0.406	0.178	7
	II	0.242	0.270	0.218	0.270	0.468	1.685	0.244	0.164	0.066	17
	III	0.238	0.269	0.221	0.272	0.454	1.618	0.230	0.163	0.066	9
C.....	Start	0.250	0.250	0.250	0.250	0.396	3.000	0.100	0.100	0.100	
	I	0.249	0.267	0.217	0.267	0.471	1.662	0.217	0.150	0.064	10
	II	0.238	0.270	0.221	0.271	0.455	1.645	0.215	0.153	0.063	2
	III	0.238	0.269	0.221	0.272	0.454	1.632	0.217	0.156	0.063	3
C*....	Start	0.238	0.270	0.221	0.271	0.455	1.645	0.215	0.153	0.063	
	III	0.238	0.269	0.221	0.272	0.456	1.638	0.216	0.154	0.062	2

[a] Threshold in the maximization procedure and numbers of samples per transition probability calculation in rounds I, II, and III were 10^{-4} and 10,000, 10^{-5} and 100,000, and 10^{-6} and 100,000, respectively.

[b] Number of iterations (full parameter vector updatings) in the maximization procedure.

random seed (rows C:III and C*:III). After round III, values obtained for the π_k, $k \in \{A, C, G, T\}$, parameters varied by <1%, those of t and K by <5%, and those of the remaining parameters (f_P, f_S, and $f_{P/S}$) by <8%. In contrast to the similarity of the final parameter values, the computational time requirements of the runs differed markedly. As can be seen from the number of iterations used in the maximization procedures of the runs (table 1, last column), the computational effort involved was positively correlated with the distance between the parameter values in which the run was started and the ''true'' values. This demonstrated the importance of having ''good'' starting values.

The degrees of variation in the final parameter values from the four different runs reflect the amount of information contained in the data concerning the different parameters. The variance is expected to be larger on the purely evolutionary parameters (t, K, f_P, f_S, and $f_{P/S}$) than on the parameters π_k, $k \in \{A, C, G, T\}$, which are determined mainly by sequence composition (equilibrium distribution). Among the purely evolutionary parameters, the variance is most likely larger on the f_P, f_S, and $f_{P/S}$ than on the t and K parameters, since the former are related to a finer partitioning of evolutionary events. It is likely that some of the variation obtained among the final values of the f_P, f_S, and $f_{P/S}$ parameters is due to too few samples in the calculation of the transition probability. As the likelihood surface is flat in the directions corresponding to the purely evolutionary parameters, a relatively high precision in the calculation of the transition probability is necessary for reliable estimates of these parameters to be obtained. The precision may be increased by augmenting the number of samples used. The pattern of variation among round II and III parameter values illustrates this (table 1): reasonable values for the π_k, $k \in \{A, C, G, T\}$, parameters could be obtained after round I, as these differed only slightly (<5%) from the final values obtained after round III. As for the t and K parameters, their values were ill determined after round I (see run B); however, after round II, they differed by no more than a few percent. The values for the remaining parameters (f_P, f_S, and $f_{P/S}$) obtained after round II differed by up to 14% from the final values.

The maximum-likelihood estimates of the parameters showed that in the double coding region analyzed, selection against amino-acid-altering substitutions was stronger in the S than in the P reading frame ($f_S \approx 0.15$ vs. $f_P \approx 0.22$). A similar finding has been reported by Yang, Lauder, and Lin (1995). Parameter estimates further indicated that selection against substitutions that altered the amino acid encoded in both reading frames was particularly strong ($f_{P/S} \approx 0.06$). Under the model above, with f_P and f_S varying freely, we performed a log likelihood ratio test for the null hypothesis that selection against amino acid substitution in the double coding region acts multiplicatively; that is, $f_{P/S} = f_P \cdot f_S$. Parameter estimates under the null hypothesis were $\pi_A = 0.238$, $\pi_C = 0.270$, $\pi_G = 0.219$, $\pi_T = 0.273$, $t = 0.450$, $K = 1.590$, $f_P = 0.346$, and $f_S = 0.250$. Use of the values from run A:III for the parameters under the model with f_P and f_S varying freely led to a $-2 \log Q$ test statistic of 1.305. This gave a P value of approximately 0.25, and the null hypothesis of multiplicatively operating selective constraints was thus accepted.

Expected Numbers of Various Types of Substitutions

If we scale the intensities so that the average rate of substitution per codon at equilibrium equals 1, the time t between sequences will effectively be measured as expected numbers of substitutions per codon. Let (s^1, s^2, s^3), (s^4, s^5, s^6), and (s^7, s^8, s^9) be three consecutive codons in reading frame I, and consider the septet of nucleotides $s = (s^1, s^2, s^3, s^4, s^5, s^6, s^7)$. The scaling is then obtained by requiring that

$$\sum_{s \in \mathcal{S}} \text{prob}_s \sum_{w \in C} q_{(s^4,s^5,s^6),w|s^1,s^2,s^6} = 1,$$

where prob_s denotes the stationary probability of septet s, and $\mathcal{S}$ is the set of all septets. A procedure for calculating prob_s is given in appendix A.

We can now write the instantaneous rate of synonymous substitutions (that is, substitutions that are synonymous in both reading frames) per codon as

$$\rho_{\text{syn(I)},\text{syn(II)}} = \sum_{s \in \mathcal{S}} \text{prob}_s \left\{ \sum_{w \in M(s)} q_{(s^4,s^5,s^6),w|s^2,s^3,s^7} \right\}, \quad (8)$$

where the set $M(s)$ consists of those $w = (w^1, w^2, w^3)$ for which the change $(s^4, s^5, s^6) \rightarrow (w^1, w^2, w^3)$ is synonymous, the change $(s^2, s^3, s^4) \rightarrow (s^2, s^3, w^1)$ is synonymous, and the change $(s^5, s^6, s^7) \rightarrow (w^2, w^3, s^7)$ is synonymous. The expected number of synonymous substitutions per codon is obtained as $\rho_{\text{syn(I), syn(II)}}t$, with the maximum-likelihood estimates as parameter values. Expected numbers of other types of substitutions are obtained by restricting the summation in equation (8) appropriately.

Expected proportions of different kinds of substitutions per codon under the model with multiplicatively acting selection factors were calculated by inserting the maximum-likelihood estimates in the formulas above. The obtained values are given in table 2. Also given are values for a situation with the same parameter values of π_k, $k \in \{A, C, G, T\}$, and K, but with no selective constraints, that is, with $f_P = f_S = f_{P/S} = 1.0$. Expected numbers of the various types of substitutions per codon can be obtained by multiplication with $\hat{t}$. For the maximum-likelihood values under the model with multiplicative selective constraints, the ratio of transitional to transversional substitutions was 0.534/0.466 = 1.146, and that of synonymous to any type of nonsynonymous substitutions was 0.058/(1 − 0.058) = 0.062. With respect to reading frame I, the ratio of synonymous to nonsynonymous rates was (0.058 + 0.314)/(1 − 0.058 − 0.314) = 0.592, whereas with respect to reading frame II, it was (0.058 + 0.439)/(1 − 0.058 − 0.439) = 0.988. The corresponding values for the case that was similar but had no selective constraints (second row of table 2) were 0.848, 0.012, 0.379, and 0.385, respectively.

Table 2
Expected Proportions of Various Types of Substitutions per Codon Under the Multiplicative Selectional Constraints Model for Two Sets of Parameter Values: Maximum-Likelihood Estimators (MLEs) and No Selective Constraints (NSCs)

	Transitions	Transversions	Syn[a]	Non(P)[b]	Non(S)[c]	Non(PS)[d]
MLEs[e]	0.534	0.466	0.058	0.439	0.314	0.189
NSCs[f]	0.459	0.541	0.012	0.266	0.263	0.458

[a] Substitutions that are synonymous in both reading frames.
[b] Substitutions that are nonsynonymous in the P reading frame only.
[c] Substitutions that are nonsynonymous in the S reading frame only.
[d] Substitutions that are nonsynonymous in both reading frames.
[e] Parameter values equal to MLEs under the multiplicative-selectional-constraints hypothesis.
[f] $f_P = f_S = f_{P/S} = 1.0$; remaining parameter values MLEs under the multiplicative-selectional-constraints hypothesis.

As compared with the hypothetical situation with similar parameter values but no selective constraints, the transition/transversion rate ratio was thus raised by a factor of 1.146/0.848 = 1.351, and the ratio of overall synonymous to nonsynonymous rates was raised by a factor of 0.062/0.012 = 5.167. The ratio of synonymous to nonsynonymous rates with respect to reading frames I and II, respectively, were raised by factors of 0.592/0.379 = 1.562 and 0.988/0.385 = 2.566. These ratios further establish the stronger degree of selective constraints for the evolution of the S gene than the part of the P gene in which the S gene overlaps.

Model Check

In this section, we present an extension of the model described in the *Methods* section and describe a simpler, but approximate, estimation method. We use the extended model to check the adequacy of the full model assumed above. In the extended model, all dinucleotide interactions and position specific nucleotide intensity parameters are allowed for.

In the extended model, we use the intensities $q_{z_i,\tilde{z}_i|z_{i-1}^2,z_{i-1}^3,z_{i+1}^1}$ from equation (2), with the nucleotide intensity $\pi_{(\tilde{z}_i)}$ replaced by a more general term which includes position specific nucleotide intensities π_k^j, $k \in \{A, G, C, T\}$, $\pi_A^j + \pi_G^j + \pi_C^j + \pi_T^j = 1$, $j = 1, 2, 3$, and dinucleotide interactions. By a dinucleotide interaction we mean a function of two neighboring nucleotides. We denote these functions by

$$\gamma_1(z_{i-1}^3, z_i^1), \ \gamma_2(z_i^1, z_i^2), \quad \text{and} \quad \gamma_3(z_i^2, z_i^3),$$

respectively. Precisely, the extended model is now defined by using equation (2) with the term $\pi_{(\tilde{z}_i)}$ replaced by

$$H(z_i, \tilde{z}_i \,|\, z_{i-1}^3, z_{i+1}^1) = \left\{ \frac{\gamma_1(z_{i-1}^3, z_i^1)\gamma_2(z_i^1, z_i^2)\gamma_3(z_i^2, z_i^3)\gamma_1(z_i^3, z_{i+1}^1)}{\gamma_1(z_{i-1}^3, \tilde{z}_i^1)\gamma_2(\tilde{z}_i^1, \tilde{z}_i^2)\gamma_3(\tilde{z}_i^2, \tilde{z}_i^3)\gamma_1(\tilde{z}_i^3, z_{i+1}^1)} \right\}^{1/2} \pi_{(\tilde{z}_i)}^{\tilde{j}},$$

where $\tilde{j}$ is the position at which $\tilde{z}_i$ differs from z_i (note that of the eight terms inside the brackets, four terms cancel because $\tilde{z}_i$ and z_i differ at one position only).

The model considered in the *Methods* section corresponds to the model here with no dinucleotide interactions, that is, with $\gamma_j(a, b) \equiv 1$, for all $(a, b) \in \{A, G, C, T\}^2$, and identical position-specific nucleotide intensities, that is, $\pi_a^j = \pi_a$, $a \in \{A, G, C, T\}$, $j = 1, 2, 3$. A model where the only dinucleotide interactions are selection against CpG dinucleotides is obtained when $\gamma_j = 1$ except for the values $\gamma_j(C, G)$, $j = 1, 2, 3$. In this model, instantaneous rates of substitution that generate (respectively, eliminate) a CpG in frame j, $j = 1, 2, 3$, are multiplied by $\{1/\gamma_j(C, G)\}^{1/2}$ (respectively, $\gamma_j(C, G)^{1/2}$) relative to instantaneous rates of substitution that leave the CpG count unaltered.

In the extended model, the stationary distribution of a sequence $z = (z_2, \ldots, z_{n-1})$ is

$$\Pi(z) = \frac{1}{Z}\gamma_1(z_1^3, z_2^1)1_{\mathcal{C}}(z_1^2, z_1^3, z_2^1) \times \prod_{i=2}^{n-1} \pi_{z_i^1}^1 \pi_{z_i^2}^2 \pi_{z_i^3}^3 \gamma_2(z_i^1, z_i^2)\gamma_3(z_i^2, z_i^3)\gamma_1(z_i^3, z_{i+1}^1) \times 1_{\mathcal{C}}(z_i^2, z_i^3, z_{i+1}^1) \tag{9}$$

for $z_i \in \mathcal{C}$ for all i. This can be verified by inspection in a manner similar to that used for the model in the *Methods* section. Like the model in the *Methods* section, the extended model allows an explicit formula for the normalizing constant to be derived (see eq. A.4 in appendix A), and that allows expected numbers of various types of substitutions to be calculated (e.g., eq. 8). For the latter, stationary probabilities of subsequences under the extended model are needed—these are derived in appendix A.

To check the adequacy of the model assumed in the *Methods* section, we compare its performance with that of the extended model. For parameter estimation in the extended model we use the following simplified procedure: we first estimate the dinucleotide interactions γ_j and the position-specific nucleotide intensities π_a^j using the stationary distribution under the extended model. We base the estimation on one of the two sequences (we use ayw2). The full extended model, however, has too many parameters to be useful. When fitting the extended model, we will make the dinucleotide interactions as simple as possible; that is, we will only include those in the model that increase the fit of the data to the model significantly. For this, we use the following stepwise selection procedure: We start by analyzing the conditional distribution of z_{i+1}^1 given (z_i^1, z_i^2, z_i^3), given in equation

Table 3
Entries Selected in the Stepwise Procedure for Estimating the Interaction Parameters in the Interaction Functions $\gamma_1(\cdot, \cdot)$, $\gamma_2(\cdot, \cdot)$ and $\gamma_3(\cdot, \cdot)$ Under the Extended Model, Along with the Corresponding Parameter Estimates and Twice the Increase in Log Likelihood Obtained by Including the Interaction in the Model

γ_1			γ_2			γ_3		
Entry	$\gamma_1(\cdot, \cdot)$	$2(l_2 - l_1)$	Entry	$\gamma_2(\cdot, \cdot)$	$2(l_2 - l_1)$	Entry	$\gamma_3(\cdot, \cdot)$	$2(l_2 - l_1)$
CG	0.31	11.0	CG	0.086	36.8	CG	0.33	1.11
GA	3.04	11.4	AA, AG	2.1, 0.57	11.8	TC	3.4	13.0
Full[a]		4.3			8.1			12.8

[a] Twice the increase in log likelihood obtained by including all remaining entries of interactions in the model.

(A.14) in appendix A, in order to estimate the interaction γ_1. We start with the model where $\gamma_1 \equiv 1$ and thereby obtain an estimate of $\pi_a^1 r^a$, $a \in \{A, C, G, T\}$ (see appendix A). We next calculate the score function (numerically) for the different entries of γ_1 and choose the entry with the largest absolute value. We include this entry as a parameter in the conditional distribution and see if this provides a significantly better description of the distribution. This procedure is continued until a reasonable fit has been obtained (see below). Next, the conditional distributions (A.15) and (A.16) from appendix A are treated in the same way in order to estimate the interactions γ_2 and γ_3.

The result of the stepwise procedure for estimating the interaction parameters and the position-specific nucleotide intensities are given in tables 3 and 4. It is clear from table 3 that there was a significant CG depression in the data—the sequence considered the first entry to be included in the interaction γ_j was the CG entry in all three positions. The stepwise procedure additionally includes four types of dinucleotide interactions. Furthermore, the estimates of the position-specific nucleotide intensities π^j varied among the three codon positions (table 4). The results from this analysis show that the simpler model does not do well with respect to describing a single sequence, meaning that the simple model is only a rough approximation of the true model.

Having estimated the interactions γ_j and the position-specific nucleotide intensities π^j in the extended model using the stationary distribution, we could return to the MCMC method of the *Methods* section to estimate the remaining purely evolutionary parameters t, K, f_P, f_S, and $f_{P/S}$. However, we mention here another approximate estimation method that will allow us to consider the fit of the model as well. We split the observable changes in the aligned codons in the two sequences into a number of disjoint groups. In particular, we used the nine groups obtained by dividing codons into groups with single and multiple changes and further dividing the group exhibiting single changes into transition and transversion groups, and dividing each of these two groups into groups based on the four combinations of synonymous and nonsynonymous changes in the two reading frames. The observed numbers N_g, $g = 1, \ldots, 9$, in the nine groups are approximately independent and Poisson distributed with, say, mean μ_g. Thus, we form an approximate likelihood based on the Poisson approximation and use this to estimate the remaining parameters (t, K, f_P, f_S, $f_{P/S}$). The means μ_g can be approximated by a simple forward simulation of the Markov process describing the evolution; that is, we must simulate exponential waiting times and simulate the jump type.

Table 4
Parameter Estimates of the Position-Specific Nucleotide Intensity Parameters Obtained Under the Extended Model in which all Selected Interactions Have Been Included

Pos[a]	π^j_A	π^j_G	π^j_C	π^j_T
$j = 1$.....	0.19	0.31	0.34	0.16
$j = 2$.....	0.23	0.22	0.26	0.29
$j = 3$.....	0.15	0.27	0.33	0.25

[a] Codon position in reading frame I.

The estimated evolutionary parameters for the extended model, are $t = 0.476$, $K = 1.946$, $f_P = 0.175$, $f_S = 0.161$, and $f_{P/S} = 0.033$. Observed numbers of codons falling into the nine groups are given in column 3 of table 5, and expected numbers under the extended model with corresponding parameter estimates are given in the fourth column. The expectations under the extended model fit well with the observed numbers in the nine categories ($-2 \log Q = 3.8 \approx \chi(4)$). In the fifth column, results are given for a model that is similar to the extended model except that the selectional constraints are assumed to act multiplicatively ($f_{P/S} = f_P f_S$). Here, parameter values identical to those in the column before are used, except that $f_{P/S} = 0.175 \cdot 0.161 = 0.028$. Even though no fitting of the parameters to the extended model with multiplicatively operating selection parameters has been done, the $-2 \log Q$ test statistic is only marginally augmented (to 4.1), and the result of multiplicatively operating selection factors found in the analyses based on the simple model of the *Methods* section are confirmed. The last column gives the expected number of codons in the nine categories under the model of the *Methods* section using the maximum-likelihood estimates from the subsection above. With a $-2 \log Q$ value of 9.3, the simple model is a good approximation of the extended model with respect to the evolutionary part, as measured by the expectations of the various types of changes.

Discussion

The model presented for the substitution process in DNA sequences in which two genes are encoded in

Table 5
Observed and Expected Numbers of Nine Groups of Codons Under the Three Models "Ext," "Ext-Mult," and "Full," along with Twice the Decrease in Log Likelihood When Going from the "Means-Free" Model to Each of These Three Models

No. of Changes[a]	Group[b]	Observed	Expected		
			Ext[c]	Ext-Mult	Full[e]
1					
	ts, syn(I), syn(II)	6	5.6	5.7	5.0
	ts, syn(I), non(II)	19	22.2	22.4	16.0
	ts, non(I), syn(II)	27	28.8	29.0	27.4
	ts, non(I), non(II)	7	9.0	8.2	9.8
	tv, syn(I), syn(II)	3	2.9	2.9	2.2
	tv, syn(I), non(II)	11	12.1	12.2	10.6
	tv, non(I), syn(II)	14	18.1	18.2	20.0
	tv, non(I), non(II)	12	13.2	11.7	19.9
Multiple	Any	22	16.8	16.0	16.4
$-2 \log Q$ ...			3.8	4.1	9.3

[a] Number of nucleotides differing in the codons in sequences 1 and 2.

[b] Type of (single-nucleotide) difference (transition [ts] or transversion [tv], synonymous [syn], or nonsynonymous [non] in either of the two reading frames).

[c] Extended model, f_P and f_S free. Parameter estimates obtained from the simple approximate estimation procedure (tables 3 and 4) are used.

[d] Extended model, $f_{P/S} = f_P f_S$. Parameter estimates obtained form the simple approximate estimation procedure (tables 3 and 4) are used, except for the value of $f_{P/S}$, which here is $f_{P/S} = f_P f_S = 0.175 \cdot 0.161 = 0.028$.

[e] The "full" model (see *Methods*). MLEs are used (table 1).

overlapping reading frames takes into account constraints imposed by the genetic code in both of the reading frames. A model for the analyses of sequences with three genes encoded in overlapping frames, in which constraints imposed by three overlapping codes are incorporated, can be obtained simply by extending the neighborhood of dependency one nucleotide to the right. A procedure for calculating the transition probability between two sequences under such a model can be obtained by minor modifications of the procedure presented here. By combining models of the above types, one can achieve a model for sequences with combinations of non-, single-, double-, and triple-coding regions in which constraints imposed by the various combinations of overlapping reading frames are allowed for. In the case of hepatitis B, analyses of the substitution process in the full genome under such a model should be feasible, given the limited size of the genome (3.2 kb) and the small number of genes.

The methodology described for calculating the transition probability between two sequences has two drawbacks. First of all, it allows for the analysis of pairs of sequences only. The development of a procedure for obtaining the likelihood of observing a set of (more than two) sequences at the tips of a given binary tree under a model with dependent substitution rates has yet to be developed. The second drawback is that the methodology is computationally very demanding. It should be possible to reduce the computational time requirements considerably by parallelizing computations. Computational requirements of a similar procedure for more than two sequences related by a tree will be increased due to the larger number of branch length parameters. It is, however, possible that the increase due to more parameters will be counterbalanced by an increase of information in the data regarding the more poorly determined purely evolutionary parameters, which would allow transition probabilities to be approximated by smaller samples of paths.

Given the computational requirements for inference under the presented model, it would be of considerable interest to compare results obtained with this model and methodology to those obtained using the more heuristic and much quicker models and procedures. Computational demands, however, seriously limit the possibility of performing simulation-based bias studies.

Acknowledgments

A.K.P. was supported by grants from the Danish Natural Science Research Foundation and the Carlsberg Foundation. Two anonymous reviewers and the associate editor are thanked for many helpful comments.

APPENDIX A

In this appendix, we first find the normalizing constants Z in the stationary distributions (3) and (9) of a sequence. We next turn to a rewriting of the stationary distribution that allows us to calculate the stationary probability of a subsequence.

Since the model in equation (3) is a special case of the model in equation (9), we first state the formulas for the earlier model. The normalizing constant Z is found by summing up the equilibrium frequencies over all $z_2, \ldots, z_{n-1}$. If we sum first over (z_i^2, z_i^3), $i = 2, \ldots, n-1$, we obtain

$$Z = \sum_{z_2^1, z_3^1, \ldots, z_{n-1}^1} \gamma_1(z_1^3, z_2^1) 1_C(z_1^2, z_1^3, z_2^1) V(z_2^1, z_3^1) \times V(z_3^1, z_4^1) \cdots V(z_{n-1}^1, z_n^1), \quad \text{(A.1)}$$

where, for $(a, b) \in \{A, G, C, T\}^2$,

$$V(a, b) = \sum_{s^2, s^3 \in \{A,G,C,T\}^2} \pi_{s^2}^2 \pi_{s^3}^3 \pi_b^1 \gamma_2(a, s^2) \gamma_3(s^2, s^3) \times \gamma_1(s^3, b) 1_C(a, s^2, s^3) 1_C(s^2, s^3, b). \quad \text{(A.2)}$$

To evaluate equation (A.1), we use the eigenvalues and eigenvectors of the 4×4 matrix V. Let $\lambda_1, \ldots, \lambda_4$ and $v_1, \ldots, v_4$ be the eigenvalues and left eigenvectors, respectively, with λ_1 being the largest eigenvalue. Writing $v_i = (v_i^A, v_i^G, v_i^C, v_i^T)$, we thus have

$$\sum_{a \in \{A,G,C,T\}} v_i^a V(a, b) = \lambda_i v_i^b.$$

Let $w = (w^A, w^G, w^C, w^T)$ be the vector with

$$w^a = \gamma_1(z_1^3, a) \pi_a^1 1_C(z_1^2, z_1^3, a), \quad \text{(A.3)}$$

where $a \in \{A, G, C, T\}$, and define coefficients $c_1, \ldots, c_4$ by

$$w = c_1 v_1 + c_2 v_2 + c_3 v_3 + c_4 v_4.$$

Then we get, from equation (A.1),

$$Z = \sum_{i=1}^{4} c_i \lambda_i^{n-2} v_i^{z_n^1} \frac{1}{\pi_{z_n^1}^1}. \quad \text{(A.4)}$$

Let us now specialize equation (A.2) to the simple model in equation (A.1), corresponding to $\gamma_j \equiv 1$ and $\pi_a^j = \pi_a$. The matrix V from equation (A.2) becomes

$$V = \begin{pmatrix} \alpha \\ \alpha \\ \alpha \\ \alpha - \beta \end{pmatrix}, \quad \text{(A.5)}$$

where the vectors $\alpha = (\alpha_A, \alpha_G, \alpha_C, \alpha_T)$ and $\beta = (\beta_A, \beta_G, \beta_C, \beta_T)$ are given by

$$\alpha = (\pi_A - \pi_T\pi_A\pi_A - \pi_T\pi_G\pi_A, \pi_G - \pi_T\pi_A\pi_G, \pi_C, \pi_T) \quad \text{and}$$

$$\beta = (\pi_A, \pi_G, \pi_C, \pi_T)\frac{\pi_{stop}}{\pi_T},$$

with $\pi_{stop} = \pi_T\pi_A\pi_A + \pi_T\pi_G\pi_A + \pi_T\pi_A\pi_G$. A simple calculation shows that the eigenvalues are

$$\lambda_1 = \frac{1}{2}(1 - 2\pi_{stop} + \sqrt{1 - 4\pi_{stop}}),$$

$$\lambda_2 = \frac{1}{2}(1 - 2\pi_{stop} - \sqrt{1 - 4\pi_{stop}}),$$

$$\lambda_3 = 0, \qquad \lambda_4 = 0, \quad \text{(A.6)}$$

with corresponding left eigenvectors

$$v_1 = \alpha + \gamma_1\beta, \qquad v_2 = \alpha + \gamma_2\beta,$$

$$v_3 = (1, -1, 0, 0), \qquad v_4 = (1, 0, -1, 0), \quad \text{(A.7)}$$

where

$$\gamma_1 = \frac{\pi_T}{2\pi_{stop}}(-1 + \sqrt{1 - 4\pi_{stop}}),$$

$$\gamma_2 = \frac{\pi_T}{2\pi_{stop}}(-1 - \sqrt{1 - 4\pi_{stop}}). \quad \text{(A.8)}$$

The vector w from equation (A.3) becomes

$$w = (\pi_A 1_C(z_1^2, z_1^3, A), \pi_G 1_C(z_1^2, z_1^3, G), \pi_C, \pi_T) = c_1v_1 + c_2v_2 + c_3v_3 + c_4v_4,$$

and the normalizing constant in equation (A.4) is

$$Z = (c_1\lambda_1^{n-2}v_1^{z_n^1} + c_2\lambda_2^{n-2}v_2^{z_n^1})\frac{1}{\pi_{z_n^1}}. \quad \text{(A.9)}$$

In particular, if (z_1^2, z_1^3) is different from (T, A) and (T, G) we find that $c_1 = c_2 = 1/\sqrt{1 - 4\pi_{stop}}$.

We next turn to a closer study of the stationary measure in equations (3) and (9). The stationary measure can be written as a product of conditional densities. To this end, we consider the matrix V in equation (A.2) again and let $r = (r^A, r^G, r^C, r^T)$ be the positive right eigenvector corresponding to the largest eigenvalue λ_1. For the simple model with V given in equation (A.5), we find

$$r = \left(1, 1, 1, \frac{1}{2}(2\pi_T - 1 + \sqrt{1 - 4\pi_{stop}})/\pi_T\right). \quad \text{(A.10)}$$

Considering the chain $\{z_i^1\}$, $i = 1, 2, \ldots$, one finds that this is a homogeneous Markov chain with transition matrix

$$T(a, b) = P(z_{i+1}^1 = b \,|\, z_i^1 = a) = (V(a, b)r^b)/(\lambda_1 r^a), \quad a, b \in \{A, G, C, T\}.$$

Since

$$\sum_a v_1^a r^a T(a, b) = \frac{r^b}{\lambda_1}\sum_a v_1^a V(a, b) = v_1^b r^b,$$

the stationary density p^0 for this Markov chain is

$$p^0(a) = P(z_i^1 = a) = \frac{v_1^a r^a}{\sum_b v_1^b r^b}, \quad a \in \{A, G, C, T\}. \quad \text{(A.11)}$$

Furthermore, the conditional density $p_{231|1}$ of $(z_i^2, z_i^3, z_{i+1}^1)$ given z_i^1 is

$$\frac{r^{z_{i+1}^1}}{\lambda_1 r^{z_i^1}}\pi_{z_i^2}^2\pi_{z_i^3}^3\pi_{z_{i+1}^1}^1\gamma_2(z_i^1, z_i^2)\gamma_3(z_i^2, z_i^3)\gamma_1(z_i^3, z_{i+1}^1) \times 1_C(z_i^1, z_i^2, z_i^3)1_C(z_i^2, z_i^3, z_{i+1}^1). \quad \text{(A.12)}$$

Thus, the stationary frequency of the septet $(z_i, z_{i+1}, z_{i+2}^1)$ is

$$p^0(z_i^1)p_{231|1}(z_i^2, z_i^3, z_{i+1}^1 \,|\, z_i^1)p_{231|1}(z_{i+1}^2, z_{i+1}^3, z_{i+2}^1 \,|\, z_{i+1}^1), \quad \text{(A.13)}$$

which can be used for evaluating the expected numbers of various types of substitutions. For the simple model, we use equation (A.13) with $\gamma_j \equiv 1$, $\pi_a^j = \pi_a$ in equation (A.12), with r given in equation (A.10), and with v_1 given in equation (A.7).

Finally, in connection with finding a suitable extended model, we use the following conditional distributions:

$$P(z_{i+1}^1 = b \,|\, (z_i^1, z_i^2, z_i^3) = (a, s^2, s^3)) = \frac{r^b\pi_b^1\gamma_1(s^3, b)}{g(s^2, s^3)}1_C(s^2, s^3, b), \quad \text{(A.14)}$$

$$P(z_i^3 = s^3 \,|\, (z_i^1, z_i^2) = (a, s^2)) = \frac{g(s^2, s^3)\pi_{s^3}^3\gamma_3(s^2, s^3)}{h(a, s^2)}1_C(a, s^2, s^3), \quad \text{(A.15)}$$

$$P(z_i^2 = s^2 \,|\, z_i^1 = a) = \frac{h(a, s^2)\pi_{s^2}^2\gamma_2(a, s^2)}{\lambda r^a}, \quad \text{(A.16)}$$

where the functions g and h are normalizing functions, defined so that equations (A.14) and (A.15) are densities.

APPENDIX B

We first derive formula (5). From the definition of $q_{\theta_1}(t_1; L)$, we have

$$P_{x\to y}(\theta_1, t_1) = \int_{\chi_{t_1}} q_{\theta_1}(t_1; L)\mu_{t_1}(dL)$$

$$= \int_{\chi_{t_2}} \left(\frac{t_1}{t_2}\right)^r q_{\theta_1}\left(t_1; \frac{t_1}{t_2}L\right)\mu_{t_2}(dL),$$

where r is the number of substitutions in the path L. We then find

$$\frac{P_{x\to y}(\theta_1, t_1)}{P_{x\to y}(\theta_2, t_2)}$$

$$= \frac{\displaystyle\int_{\chi_{t_2}} \left(\frac{t_1}{t_2}\right)^r q_{\theta_1}\left(t_1; \frac{t_1}{t_2}L\right)\mu_{t_2}(dL)}{\displaystyle\int_{\chi_{t_2}} q_{\theta_2}(t_2; L)\mu_{t_2}(dL)}$$

$$= \int_{\chi_{t_2}} \frac{t_1^r q_{\theta_1}\left(t_1; \frac{t_1}{t_2}L\right)}{t_2^r q_{\theta_2}(t_2; L)} \frac{q_{\theta_2}(t_2; L)}{\displaystyle\int_{\chi_{t_2}} q_{\theta_2}(t_2; \tilde{L})\mu_{t_2}(d\tilde{L})}\mu_{t_2}(dL)$$

$$= \tilde{E}\left(\frac{t_1^r q_{\theta_1}\left(t_1; \frac{t_1}{t_2}L\right)}{t_2^r q_{\theta_2}(t_2; L)}\right),$$

where $\tilde{E}$ is the mean under the measure $\tilde{P}$ defined in equation (6).

The weight $q_\theta(t; L)$ of a path L with r substitutions is the product of the densities of r waiting times, times the product of r jump probabilities, times the probability that the last waiting time exceeds t. Since the intensities in these waiting times are the sum over all the positions of the intensity for an event at this position, one sees that $q_\theta(t; L)$ becomes a product

$$q_\theta(t; L) = \prod_{i=2}^{n-1} q_\theta(t, i, 1)q_\theta(t, i, 2)q_\theta(t, i, 3), \quad \text{(B.1)}$$

where $q_\theta(t, i, 1)$ depends on $(L_{i-1}^2, L_{i-1}^3, L_i)$, and $q_\theta(t, i, 2)$ and $q_\theta(t, i, 3)$ depend on (L_i, L_{i+1}^1). To give the exact form of these terms, define s_i to be the total number of substitutions in the paths L_{i-1}^2, L_{i-1}^3, L_i^1, L_i^2, L_i^3, and L_{i+1}^1. Let $u_i(r)$ be the time the rth among these substitutions occurs, and let $z_{i-1}^2(r)$, $z_{i-1}^3(r)$, $z_i(r)$, and $z_{i+1}^1(r)$, respectively, be the nucleotide contents of the second and third positions of codon $i - 1$, the three positions in codon i, and the first position in codon $i + 1$ after the rth substitution. Set

$$(z_{i-1}^2(0), z_{i-1}^3(0), z_i(0), z_{i+1}^1(0)) = (x_{i-1}^2, x_{i-1}^3, x_i, x_{i+1}^1)$$

and

$$(z_{i-1}^2(s_i), z_{i-1}^3(s_i), z_i(s_i), z_{i+1}^1(s_i)) = (y_{i-1}^2, y_{i-1}^3, y_i, y_{i+1}^1).$$

With these definitions, we can write $q_\theta(t, i, j)$ as

$$q_\theta(t, i, j)$$

$$= \left\{\prod_{r=1}^{s_i} (q_{z_i(r-1),z_i(r)|z_{i-1}^2(r-1),z_{i-1}^3(r-1),z_{i+1}^1(r-1)})^{1(z_i^j(r)\neq z_i^j(r-1))}\right.$$

$$\times \exp\{-q^j_{z_i(r-1)|z_{i-1}^2(r-1),z_{i-1}^3(r-1),z_{i+1}^1(r-1)}$$

$$\left.\times (u_i(r) - u_i(r-1))\}\right\}$$

$$\times \exp\{-q^j_{z_i(s_i)|z_{i-1}^2(s_i),z_{i-1}^3(s_i),z_{i+1}^1(s_i)}(t - s_i)\}$$

where

$$q^j_{z|a^2,a^3,b^1} = \sum_{\{w\in C|w^k=z^k,k\neq j\}} q_{z,w|a^2,a^3,b^1}.$$

The $\tilde{q}_{\theta_2}(L_i \,|\, L_{i-1}, L_{i+1})$ term from equation (7) comes from the conditional distribution of L_i given (L_{i-1}, L_{i+1}) and is found from equation (B.1) to be

$$\tilde{q}_{\theta_2}(L_i \,|\, L_{i-1}, L_{i+1})$$

$$= q_{\theta_2}(t_2, i-1, 2)q_{\theta_2}(t_2, i-1, 3)q_{\theta_2}(t_2, i, 1)$$

$$\times q_{\theta_2}(t_2, i, 2)q_{\theta_2}(t_2, i, 3)q_{\theta_2}(t_2, i+1, 1).$$

The term $\tilde{q}_{\theta_2}(L_i' \,|\, L_{i-1}, L_{i+1})$ is defined as above, with L_i replaced by L_i'.

LITERATURE CITED

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

GANEM, D. 1996. *Hepadnaviridae* and their replication. Pp. 2703–2737 *in* B. N. FIELDS, D. M. KNIPE, and P. M. HOWLEY, eds. *Fields Virology*. Vol. **2**, 3rd edition. Lippincott-Raven, Philadelphia.

GILKS, W. R., S. RICHARDSON, and D. J. SPIEGELHALTER. 1996. Introducing Markov chain Monte Carlo. Pp. 1–9 *in* W. R. GILKS, S. RICHARDSON, and D. J. SPIEGELHALTER, eds. Markov chain Monte Carlo in practice. Chapman and Hall, London.

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

HASEGAWA, M., M. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HEIN, J., and J. STØVLBÆK. 1995. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. J. Mol. Evol. **40**:181–189.

JENSEN, J. L., and A. K. PEDERSEN. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. **32**:499–517.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MUSE, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139**: 1429–1439.

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

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, and B. P. FLANNERY. 1992. Numerical recipes in Fortran. Cambridge University Press, Cambridge, England.

SCHÖNINGER, M., and A. VON HAESELER. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3**:240–247.

TILLIER, E. R. M., and R. A. COLLINS. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12**:7–15.

YANG, Z., I. J. LAUDER, and H. J. LIN. 1995. Molecular evolution of the hepatitis B virus genome. J. Mol. Evol. **41**: 587–596.

KEITH CRANDALL, reviewing editor

Accepted December 11, 2000