A Codon-based Model Designed to Describe Lentiviral Evolution

Anne-Mette K. Pedersen, Carsten Wiuf,†* and Freddy B. Christiansen**

*Institute of Biological Sciences, University of Aarhus, Denmark; and †Center for Biological Sequence Analysis, Department of Chemistry, Technical University of Denmark

A codon-based model designed to describe lentiviral evolution is developed. The model incorporates unequal base compositions in the three codon positions and selection against the CpG dinucleotide within codons to account for a deficit of this dinucleotide exhibited by lentiviral genes. The model is, to a large extent, able to account for the pattern of codon usage exhibited by the HIV1 genes *gag*, *pol*, and *env*, in spite of its parameter paucity. The model is extended to a similar model which operates on pentets (codons and their neighboring bases). The results obtained by the pentet model establish the importance of depression of CpGs across codon boundaries as well as within codons. The goodness of fit of the CpG depression model to the observed evolution in pairwise alignments of HIV1 sequences is assessed. The model provides a significantly better description of the observed evolution than the simpler models examined. The parameter estimates indicate that part of the unusually large biases in nucleotide frequencies observed in HIV1 genes is caused by selection against CpGs. We find that the estimates of expected numbers of substitutions, of transitions to transversions, and of synonymous to nonsynonymous substitution rates are robust to CpG depression, whereas the ratio of CpG-generating substitutions to other substitutions is strongly influenced by the choice of model.

Introduction

During the last decades, an increasing effort has been devoted to the development of statistical models for the evolutionary analysis of DNA sequences. A statistical approach clarifies the analysis because the assumptions of a method are explicitly stated and may be checked and because the reliability of the results obtained from an analysis may be evaluated.

The simplest models of the substitution processes in DNA sequences assume that the evolutionary processes in the different nucleotide sites are independent and identical and that the substitution process at each site is a Markov process defined by a rate matrix Q (see Zharkikh 1994 for a review). The simplest rate matrices, such as that suggested by Jukes and Cantor in 1969, assume that the different kinds of substitutions occur at the same rate. The Kimura two-parameter model (Kimura 1980) allows the rates of substitutions by transitions and transversions to differ—a widespread phenomenon (Kimura 1983, pp. 90–97). Both of these classical models made the implicit assumption of a uniform base composition. Felsenstein (1984) and Hasegawa, Kishino, and Yano (1985) suggested extensions of Kimura's model which relaxed this assumption. They suggested models where the substitution rates of a nucleotide x are assumed to be proportional to the equilibrium frequency of the nucleotide, π_x, where $x \in \{A, C, G, T\}$. Biased nucleotide compositions are found in many organisms (van Hemert and Berkhout 1995; Baumann 1996). More recently, the assumption of identical substitution processes in the different sites in an alignment has been relaxed to the extent that each site may be assigned a site-specific rate of substitution. Yang (1993) assumed that the rates of substitutions in the different sites were drawn independently from a Γ distribution, whereas Felsenstein and Churchill (1996) assigned the rates to the sites by a hidden Markov model. Extending a model to allow for rate variation does not alter the equilibrium frequencies assumed by the model.

More complicated models have been developed for the analysis of coding sequences. In general, synonymous substitutions accumulate at a much higher rate than nonsynonymous substitutions (Kimura 1983, pp. 90–97). Whether a nucleotide substitution is synonymous or nonsynonymous depends on the nucleotides that occupy the other positions of the codon at the instant of the substitution. Therefore, the substitution processes in the nucleotide sites of a codon cannot be independent. Rate matrices defined at the codon level, rather than at the nucleotide level, distinguish between synonymous and nonsynonymous substitutions and allow a description of the deviations from independence. The substitution process in codon-based models is thus described by a 64×64 (or 61×61, if stop codons are excluded) rate matrix in which the entry q_{ij} is defined as the rate at which codon j is substituted for codon i. The rates of substitution in the codon-based models given by Muse and Gaut (1994) are

$$q_{ij} = \begin{cases} \mu\pi_{(j)} & \text{if codons } i \text{ and } j \text{ encode the same amino acid} \\ \nu\pi_{(j)} & \text{if codons } i \text{ and } j \text{ encode different amino acids} \\ 0 & \text{if codons } i \text{ and } j \text{ differ by multiple substitutions,} \end{cases}$$

and those of Goldman and Yang (1994) are

$$q_{ij} = \begin{cases} \alpha f_j e^{-d_{aa_i aa_j}/V} & \text{if codons } i \text{ and } j \text{ differ by a transition} \\ \beta f_j e^{-d_{aa_i aa_j}/V} & \text{if codons } i \text{ and } j \text{ differ by a transversion} \\ 0 & \text{if codons } i \text{ and } j \text{ differ by multiple substitutions.} \end{cases}$$

Key words: codon-based model, lentivirus, CpG depression, codon usage, HIV1, goodness of fit.

Address for correspondence and reprints: Anne-Mette Krabbe Pedersen, Department of Ecology and Genetics, Institute of Biological Sciences, University of Aarhus, Ny Munkegade, Building 540, DK-8000 Aarhus C, Denmark. E-mail: annemet@pop.bio.aau.dk.

Mol. Biol. Evol. 15(8):1069–1081. 1998

Table 1
Nucleotide Frequencies in the Three Codon Positions of the Sequence HIV3202A12

Codon position	A	C	G	T
1	0.338	0.196	0.316	0.150
2	0.334	0.226	0.208	0.232
3	0.424	0.156	0.228	0.192

NOTE.—The frequencies given are for all genes pooled.

Here, (j) denotes the nucleotide in codon j where codon i and j differ and $\pi_{(j)}$ is the equilibrium frequency of this nucleotide. The Grantham distance between the amino acids aa_i and aa_j is $d_{aa_iaa_j}$, and V is a general variability parameter. The equilibrium frequency of the codon j is f_j. The models do not eliminate the possibility that a codon, through time, is substituted by a codon which differs at more than one site, but this is modeled as a series of single nucleotide substitutions. The models represent two extremes in terms of parametrization in that the Muse and Gaut model contains merely five parameters, whereas the Goldman and Yang model uses 63 parameters, of which the codon equilibrium frequencies comprise 60. An implicit assumption of the Muse and Gaut model is that the equilibrium frequency of a codon $i_1i_2i_3$ is $\pi_{i_1}\pi_{i_2}\pi_{i_3}$, $i_k \in \{A, C, G, T\}$, $k = 1, 2, 3$, whereas the Goldman and Yang model does not assume a special structure for the equilibrium frequencies.

Codon-based models are much slower computationally than nucleotide-based models. This is because the calculation of the transition probability matrix $P(t) = \exp(Qt)$, necessary for the calculation of the likelihood, is much more demanding when Q is a codon substitution rate matrix than when it is a nucleotide substitution matrix. Closed-form expressions for the transition probabilities have been found for the mentioned nucleotide-based models, but for the two codon-based models, no explicit solutions are available. Therefore, $P(t)$ for these models is approximated by using a Taylor expansion, i.e., $P(t) \approx \Sigma_{k=0}^{m}[(Qt)^k/k!]$, for a sufficiently large m. The computational burden associated with the maximization of the likelihood is considerable when $P(t)$ is calculated using numerical algorithms.

Computational complexity increases with the number of parameters, and so does the probable error in the estimates of the parameters (Huelsenbeck and Rannala 1997). The use of too simplified a model may lead to biased results (Zharkikh 1994). Modeling is therefore a balancing act: on one side weighs the demand for simplicity, while on the other side weighs the demand for complexity. Simplicity secures computational feasibility and minimal error estimates. Sufficient complexity is needed to assure reliable results.

We develop a codon-based model designed to describe lentiviral evolution. By incorporating features that are characteristic for lentiviral genes, we arrive at a relatively prudent model which, in spite of its simplicity, performs remarkably well. Rather than introducing an equilibrium frequency parameter for each codon (as in the Goldman and Yang model), we elaborate upon the frequency structure of the Muse and Gaut model. The ability of this new model to describe the codon frequencies observed in HIV1 genes is compared to that of simpler models, and we find that the extreme codon bias exhibited by lentiviruses is, to a large extent, accommodated in the model. We extend the model to a similar model that operates on pentets, i.e., the five nucleotides of a codon and its two neighboring sites, rather than three consecutive nucleotides, and we compare the performance of different versions of this model with respect to their abilities to account for the pentet frequencies observed in HIV1 genes. We analyze three coding regions in each of three pairwise alignments of HIV1 sequences using three versions of our codon-based model. Maximum-likelihood estimates obtained under the three models are compared, and we examine how well the models fit the data. Finally, the assessment of genetic distances by our model is compared to that of simpler models.

Methods

Lentiviral genes are characterized by two features: an extremely high frequency of the A nucleotide (up to 40% of the RNA genome [van Hemert and Berkhout 1995]) and an extremely low level of the CpG dinucleotide (Berkhout and van Hemert 1994). The

Table 2
Dinucleotide Frequencies in Codon Positions 1 and 2, 2 and 3, and 3 and 1, in Each of the Genes *gag*, *pol*, and *env*, in the Sequence HIV3202A12

		AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
gag	(1,2)	59	28	43	39	45	31	3	19	55	39	39	25	9	15	19	33
	(2,3)	64	18	48	38	59	24	5	25	41	21	30	12	49	15	31	21
	(3,1)	61	34	91	26	35	25	3	15	38	20	44	12	34	19	20	23
pol	(1,2)	117	53	51	86	72	49	1	33	106	54	65	64	29	11	36	57
	(2,3)	148	35	57	84	102	27	2	36	59	10	53	31	115	31	38	56
	(3,1)	132	62	173	57	58	18	4	23	48	31	50	21	68	44	62	32
env	(1,2)	100	51	49	64	46	25	0	30	59	42	46	48	16	21	41	52
	(2,3)	76	29	39	77	63	27	10	39	44	18	39	35	77	29	45	43
	(3,1)	85	46	85	44	58	15	4	26	42	27	41	22	79	13	64	38

bias toward the A nucleotide is more pronounced in third codon position than in first and second codon positions. Low frequencies of CpG are not restricted to neighboring positions within a codon but extend to pairs involving third and first positions of adjacent codons. Tables 1 and 2 show the extent of these biases for a typical sequence.

The molecular mechanism responsible for the generation of the A-rich genomes is unknown (Berkhout and van Hemert 1994). The depletion of CpG dinucleotides has been linked to the effect of methylation (Coulondre and Miller 1978) in that C nucleotides are more prone to methylation when placed in a CpG dinucleotide. The level of methylation of a gene has been found to be negatively correlated with the level of gene expression, and Shpaer and Mullins (1990) argued that the low levels of CpG in lentiviral genes are due to selection against CpGs rather than to a mutational bias.

The high A frequency and the CpG depression is mirrored in the pattern of codon usage in lentiviruses. Among synonymous codons, the A-rich codons are used more frequently than the A-poor codons. Codons that contain a CpG dinucleotide are virtually absent from lentiviral genes (van Hemert and Berkhout 1995). Codon counts for the three genes *gag*, *pol*, and *env* in the sequence HIV3202A12 are given in table 3. In regions of overlapping reading frames and in regions where genes overlap with long terminal repeats, the frequencies of A and CpG are less extreme. This is consistent with the hypothesis of selection against CpGs, because other functional constraints may limit the selection pressure in such regions (Shpaer and Mullins 1990).

The Model

We assume that the substitution processes in the codons are independent, identical Markov processes at equilibrium. The Markov process is defined by the rate matrix Q, with entries q_{ij} defined below. The nucleotide frequencies in the three codon positions are different, and we let the rate at which a nucleotide in codon position k is substituted by a nucleotide i_k be proportional to $\pi^k_{i_k}$, where $i_k \in \{A, C, G, T\}$, $\pi^k_A + \pi^k_C + \pi^k_G + \pi^k_T = 1$, and $k = 1, 2, 3$. We distinguish between a transition rate coefficient α and a transversion rate coefficient β. The rates of substitutions that lead to amino acid changes are modified relative to those that do not by multiplication of a selection pressure factor f. When $f > 1$, amino acid substitutions are promoted, and when $f < 1$, synonymous substitutions are favored. In a similar way, we allow for selection against the CpG dinucleotide by multiplying the rates that correspond to the generation of a CpG by the factor $1/\lambda$ and those that correspond to the loss of a CpG by λ. Entries that correspond to an unaltered CpG status are left untouched. If $\lambda > 1$, CpGs are selected against. If $\lambda < 1$, CpGs are selected for. If $\lambda = 1$, selection on CpGs vanishes.

Thus, the rate of substitution, q_{ij}, from a codon i to a different codon j, is defined as follows:

$$q_{ij} = \begin{cases} \alpha\pi^k_{(j)} & \text{transition, synonymous, CpG status unchanged} \\ \beta\pi^k_{(j)} & \text{transversion, synonymous, CpG status unchanged} \\ f\alpha\pi^k_{(j)} & \text{transition, nonsynonymous, CpG status unchanged} \\ f\beta\pi^k_{(j)} & \text{transversion, nonsynonymous, CpG status unchanged} \\ \lambda\alpha\pi^k_{(j)} & \text{transition, synonymous, CpG lost} \\ \lambda\beta\pi^k_{(j)} & \text{transversion, synonymous, CpG lost} \\ \lambda f\alpha\pi^k_{(j)} & \text{transition, nonsynonymous, CpG lost} \\ \lambda f\beta\pi^k_{(j)} & \text{transversion, nonsynonymous, CpG lost} \\ \frac{1}{\lambda}\alpha\pi^k_{(j)} & \text{transition, synonymous, CpG generated} \\ \frac{1}{\lambda}\beta\pi^k_{(j)} & \text{transversion, synonymous, CpG generated} \\ \frac{1}{\lambda}f\alpha\pi^k_{(j)} & \text{transition, nonsynonymous, CpG generated} \\ \frac{1}{\lambda}f\beta\pi^k_{(j)} & \text{transversion, nonsynonymous, CpG generated} \\ 0 & \text{the codons } i \text{ and } j \text{ differ at multiple sites} \end{cases}$$

where (j) is the nucleotide in codon j, which is different from that of codon i, and k is the codon position at which the codons differ. We omit the rows and columns that correspond to stop codons from the rate matrix. This is equivalent to using a rate matrix in which the entries that correspond to substitutions to and from stop codons are multiplied by factors of $1/\tau$ and τ, respectively, where τ is assigned the value ∞. The diagonal entries of the rate matrix $Q = \{q_{ij}\}$ are given by requiring that rows sum to zero. We refer to this model as the CpG depression codon-based model.

The rate of substitution from codon ACC to ACG is thus $(1/\lambda)\beta\pi^3_G$, since the two codons differ by a transversion (factor β), both codons encode threonine (no factor f), a CpG is generated (factor $1/\lambda$), and the substitution is to a G at codon position 3 (factor π^3_G). Note that in addition to substitutions between codons with no CpG, the substitution from the codon CCG to the codon CGG, or vice versa, also leaves the CpG status unchanged ($f\beta\pi^2_G$ and $f\beta\pi^2_C$, respectively).

Equilibrium Frequencies

The importance of the equilibrium frequencies postulated by a model can be illustrated by considering the following example. Assume that we have two almost identical sequences and that we want to test if a model is able to describe the evolutionary process that generated these sequences. Assume that the model postulates that the substitution processes in the individual codons are independent, identical Markov processes at equilibrium. Since the sequences are almost identical, the observation matrix for all possible pairs of codons in the alignment will have almost no nonzero entries, except along the diagonal. Consequently, the likelihood-maxi-

Table 3
Codon Counts for the Three Genes *gag, pol,* and *env* and for All Genes Pooled for the Sequence HIV3202A12

		gag					*pol*					*env*				
1st	2nd	A	C	G	T	2nd	A	C	G	T	2nd	A	C	G	T	3rd
A....		21	13	19	13		64	29	22	39		31	22	30	34	A
		8	8	11	5		5	8	2	12		18	16	11	10	C
		15	0	10	16		23	2	9	12		17	3	13	16	G
		15	7	3	5		26	14	16	24		44	16	17	17	T
C....		17	18	1	7		35	27	0	12		22	14	1	8	A
		2	4	0	3		6	9	0	4		4	9	3	13	C
		17	1	2	5		21	0	1	11		20	1	1	17	G
		9	8	0	4		12	13	0	6		8	6	0	10	T
G....		25	18	21	14		50	40	39	38		35	23	28	28	A
		7	9	8	3		16	9	4	7		11	8	5	8	C
		16	3	9	5		16	0	17	10		12	5	15	14	G
		7	9	1	3		29	13	7	10		20	16	10	10	T
T....		0	10	0	15		0	8	0	26		0	9	0	19	A
		1	3	2	4		8	1	0	8		6	2	6	11	C
		0	1	9	5		0	0	26	7		0	3	28	18	G
		7	1	8	9		22	2	11	16		14	8	17	15	T

mization problem under the model, that is, the maximization of the function

$$L = \prod_{(i,j)\in C} p(t)_{i,j}^{n_{i,j}},$$

where C is the set of possible codon patterns, $p(t)_{i,j}$ is the probability of observing the pattern (i,j) after time t, and $n_{i,j}$ is the number of times the pattern (i,j) is observed, will be similar to a multinomial problem in which the probability vector comprises the expressions for the equilibrium frequencies of the codons that are hypothesized by the model. This is because almost the only $p_{i,j}(t)$'s for which $n_{i,j}$ is not zero are the $p(t)_{i,i}$'s. If π_i is the equilibrium frequency of the codon i and $P(t)_{i,i}$ is the i,i'th entry in the transition probability matrix, then $p(t)_{i,i} = \pi_i P(t)_{i,i}$. When the sequences analyzed are almost identical, $P(t)_{i,i}$ is very close to 1, and therefore, $p(t)_{i,i}$ is approximately equal to the equilibrium frequency of codon i, π_i. A test of goodness of fit of the model, e.g., by the test suggested by Goldman (1993), will result in a clear-cut rejection if the equilibrium frequencies postulated by the model do not adequately describe the frequencies observed in the sequences. Although sequences analyzed differ at a larger proportion of the sites than indicated in the above example, a significant proportion of the sites in an alignment will typically be constant sites (otherwise one would be rather reluctant to accept the alignment!), and thus the performance of a model, as judged by its goodness of fit in general, will depend strongly on the assumptions of the model regarding equilibrium frequencies.

All of the models mentioned, including the CpG depression codon-based model, are reversible models. That is, they satisfy that $\pi_i q_{ij} = \pi_j q_{ji}$, where π_i is the equilibrium frequency of the nucleotide or codon i, and q_{ij} is the i, j'th entry in the rate matrix. Only parameters that appear asymmetrically about the diagonal of a rate matrix in a reversible model affect the equilibrium frequencies. Parameters that are symmetrically distributed are merely local linear transformations of the time scale on which substitution rates are measured. For instance, the Jukes and Cantor and the Kimura two-parameter models for nucleotide substitutions have no asymmetrically distributed parameters, and the equilibrium frequencies of each of the nucleotides are therefore ¼. In the more complex nucleotide substitution matrices, such as that suggested by Felsenstein (1984) and Hasegawa, Kishino, and Yano (1985), the π_k parameters are asymmetrically distributed and are the nucleotide equilibrium frequencies. In Muse and Gaut's codon-based model, only the π_k parameters are asymmetrically distributed, and the equilibrium frequency of the codon ijk is $\pi_i\pi_j\pi_k$, for $i, j, k \in \{A, C, G, T\}$. The equilibrium frequency of the codon m in Goldman and Yang's model is π_m—again, these are the only nonsymmetrically distributed parameters.

In the CpG depression codon-based model, the λ and the $\pi_{i_k}^k$ parameters are asymmetrically distributed. By solving for the codon equilibrium frequencies, in this model we get the equilibrium frequency of codon $i_1i_2i_3$ as $\kappa\pi_{i_1}^1\pi_{i_2}^2\pi_{i_3}^3$ if the codon contains a CpG dinucleotide and $\lambda^2\kappa\pi_{i_1}^1\pi_{i_2}^2\pi_{i_3}^3$ if the codon contains no CpG dinucleotide, where $i_k \in \{A, C, G, T\}$, $k = 1, 2, 3$, and κ is the normalizing constant that makes the equilibrium codon frequencies add to 1:

$$\kappa = \frac{1}{\lambda^2(1 - (\pi_C^1\pi_G^2 + \pi_C^2\pi_G^3) - \Pi_{\text{stop}}) + (\pi_C^1\pi_G^2 + \pi_C^2\pi_G^3)}.$$

When $\lambda = 1$, the equilibrium frequency of any codon $i_1i_2i_3$ is $\pi_{i_1}^1\pi_{i_2}^2\pi_{i_3}^3/(1 - \Pi_{\text{stop}})$, as would be the case for an extended version of Muse and Gaut's model, in which the three codon positions were allowed to have different sets of nucleotide equilibrium frequencies. The larger λ is, the smaller are the frequencies of CpG-containing codons.

In the CpG depression codon-based model the parameter $\pi_{i_k}^k$, $i_k \in \{A, C, G, T\}$, $k = 1, 2, 3$ is not the equilibrium frequency of the nucleotide i_k in codon position k. By summing the equilibrium frequencies of all

Table 4
χ^2 Test Statistics and ML Estimates of λ^2 for the Observed Versus Expected Codon Frequencies Under the Triplet Models H_{-CpG} and H_{+CpG} Performed on Each of the Genes *gag, pol,* and *env* in the Sequence HIV3202A12 (seq1)

		χ^2			$\hat{\lambda}^2$		
GENE	MODEL	Seq1	Average	Min–max	Seq1	Average	Min–max
gag	H_{-CpG}	135.68	127.37	105.08–140.85	—	—	—
	H_{+CpG}	68.04	73.18	63.65–82.82	9.39	8.58	4.42–12.21
pol	H_{-CpG}	285.52	291.24	224.98–321.24	—	—	—
	H_{+CpG}	129.53	132.50	108.29–149.26	28.50	24.16	10.04–45.50
env	H_{-CpG}	196.41	187.04	164.54–221.76	—	—	—
	H_{+CpG}	104.61	101.33	82.36–118.47	7.61	7.24	5.28–11.23

NOTE.—Averages and minimal and maximal values are given for the tests performed on 28 sequences.

codons that have a given nucleotide at a given position, we find the following equilibrium nucleotide frequencies:

	A	*C*	*G*	*T*
position 1	$\kappa\lambda^2\pi_A^1\gamma_1$	$\kappa\lambda^2\pi_C^1(\gamma_1 - \delta_1)$	$\kappa\lambda^2\pi_G^1\gamma_1$	$\kappa\lambda^2\pi_T^1\gamma_1$
position 2	$\kappa\lambda^2\pi_A^2$	$\kappa\lambda^2\pi_C^2(1 - \delta_{2'})$	$\kappa\lambda^2\pi_G^2(1 - \delta_{2''})$	$\kappa\lambda^2\pi_T^2$
position 3	$\kappa\lambda^2\pi_A^3\gamma_3$	$\kappa\lambda^2\pi_C^3\gamma_3$	$\kappa\lambda^2\pi_G^3(\gamma_3 - \delta_3)$	$\kappa\lambda^2\pi_T^3\gamma_3$

where

$$\gamma_1 = (1 - \pi_C^2\pi_G^3) + \frac{1}{\lambda^2}\pi_C^2\pi_G^3$$

$$\gamma_3 = (1 - \pi_C^1\pi_G^2) + \frac{1}{\lambda^2}\pi_C^1\pi_G^2$$

$$\delta_1 = \left(1 - \frac{1}{\lambda^2}\right)\pi_G^2$$

$$\delta_{2'} = \left(1 - \frac{1}{\lambda^2}\right)\pi_G^3$$

$$\delta_{2''} = \left(1 - \frac{1}{\lambda^2}\right)\pi_C^1$$

$$\delta_3 = \left(1 - \frac{1}{\lambda^2}\right)\pi_C^2$$

The equilibrium frequency of A in codon position 1 is a weighed average of the equilibrium frequencies of the two general types of codons: ACG and A(nonCG), where the weights are determined by the codon equilibrium frequencies of these types of codons. When $\lambda = 1$, the equilibrium frequency of the nucleotide A in the first position is π_A^1 (except for the usual effect of the stop codon frequencies in the normalizing constant). If λ is large, the codons without CpGs will weigh more heavily than those with CpGs in the determination of the overall equilibrium frequency of a nucleotide.

Results

Expected Versus Observed Codon Frequencies in HIV1 Genes

Below, we compare χ^2 tests of goodness of fit of the observed to expected codon counts for the CpG depression codon-based model, H_{+CpG}, and the simpler model obtained when λ is restricted to be equal to 1, H_{-CpG}. The expected codon counts under a model are found by inserting the maximum-likelihood estimates obtained under a model that hypothesizes that codons are multinomially distributed, $m(n,\pi)$, where n is the number of codons in the gene considered and π is the vector of equilibrium frequencies under the model, in the expression for the codon equilibrium frequencies and multiplying by n.

Tests were performed on each of the *gag, pol,* and *env* genes for each of the complete HIV1 genome sequences that were available from the HIV sequence database in Los Alamos, N. M., at the time. A total of 26 sequences, with genes annotated, was obtained. Stop codons and regions of the genes in which there was overlap with other genes were excluded from the analysis. The number of codons in the single-coding region(s) of the *gag*, *pol*, and *env* genes are approximately 500, 900, and 700, respectively.

The sequences are highly correlated due to common ancestry, and therefore the results for the different sequences are quite similar. In table 4, χ^2 test statistics obtained for the genes in the sequence HIV3202A12 are given, along with the maximum-likelihood estimate obtained for the parameter λ^2 (λ only appears in the codon equilibrium frequencies as λ^2). Means, maximum and minimum values of χ^2 test statistics, and estimates of λ^2 for the analyses performed on all the sequences are presented here as well.

The increase in fit obtained by allowing CpGs to be depressed is quite impressive: χ^2 values are reduced by one- to two-thirds as a consequence of the extension of the model by a single parameter, λ. The χ^2 statistics are still quite large, however, in that they should be compared with the $\chi^2(51)$ and $\chi^2(50)$ distributions for the fixed and the free λ models, respectively. The 95th percentiles in these models are 68.7 and 67.5. Even with CpG depression, the codon-based model is still rejected for the three genes, although the short gene *gag* is just at the border of rejection. Selection against CpG-containing codons is common to the three genes analyzed, although the force of the selection differs. Selection against CpGs is strongest in the *pol* gene. The estimates obtained for λ^2 in the different genes vary considerably between sequences.

In order to better understand the role of the λ parameter in the determination of the expected codon equilibrium frequencies under the CpG depression codon-based model, it is instructive to take a look at the maximum-likelihood estimate of λ, or rather, λ^2. Under the model of multinomially distributed codon counts, $m(n, \pi)$, where π is the vector of codon equilibrium frequencies expected under the model and n is the total number of codons in a studied sequence, an expression for the maximum-likelihood estimate of λ^2 may be found explicitly in terms of the π_i^k parameters and the observed numbers of codons as

$$\hat{\lambda}^2(\pi) = \frac{K_2}{K_1},$$

where

$$K_1 = \frac{\Sigma_{C_pG}\, x_{i_1 i_2 i_3}/x\ldots}{(\pi_C^1\pi_G^2 + \pi_C^2\pi_G^3)}$$

$$K_2 = \frac{\Sigma_{\text{not}C_pG}\, x_{i_1 i_2 i_3}/x\ldots}{1 - (\pi_C^1\pi_G^2 + \pi_C^2\pi_G^3) - \Pi_{\text{stop}}}$$

K_1 is the ratio of the observed frequency to the expected frequency (when independence of neighboring positions are assumed) of codons that contain a CpG dinucleotide, and K_2 is the similar ratio for codons that do not contain a CpG. When this expression for λ^2 is inserted in the expected codon equilibrium frequencies, we get $K_1\pi_{i_1}^1\ \pi_{i_2}^2\pi_{i_3}^3$ if the codon contains a CpG dinucleotide and $K_2\pi_{i_1}^1\pi_{i_2}^2\pi_{i_3}^3$ if the codon contains no CpG dinucleotide. The increase in fit is thus obtained because the CpG depression model allows expected frequencies of CpG-containing codons to be decreased relative to the rest of the codons and simultaneously the estimates of the $\pi_{i_k}^k$ to be determined primarily by the more frequent codons. Figure 1 shows the observed CpG-containing codon counts and the counts expected under the models H_{+CpG} and H_{-CpG}.

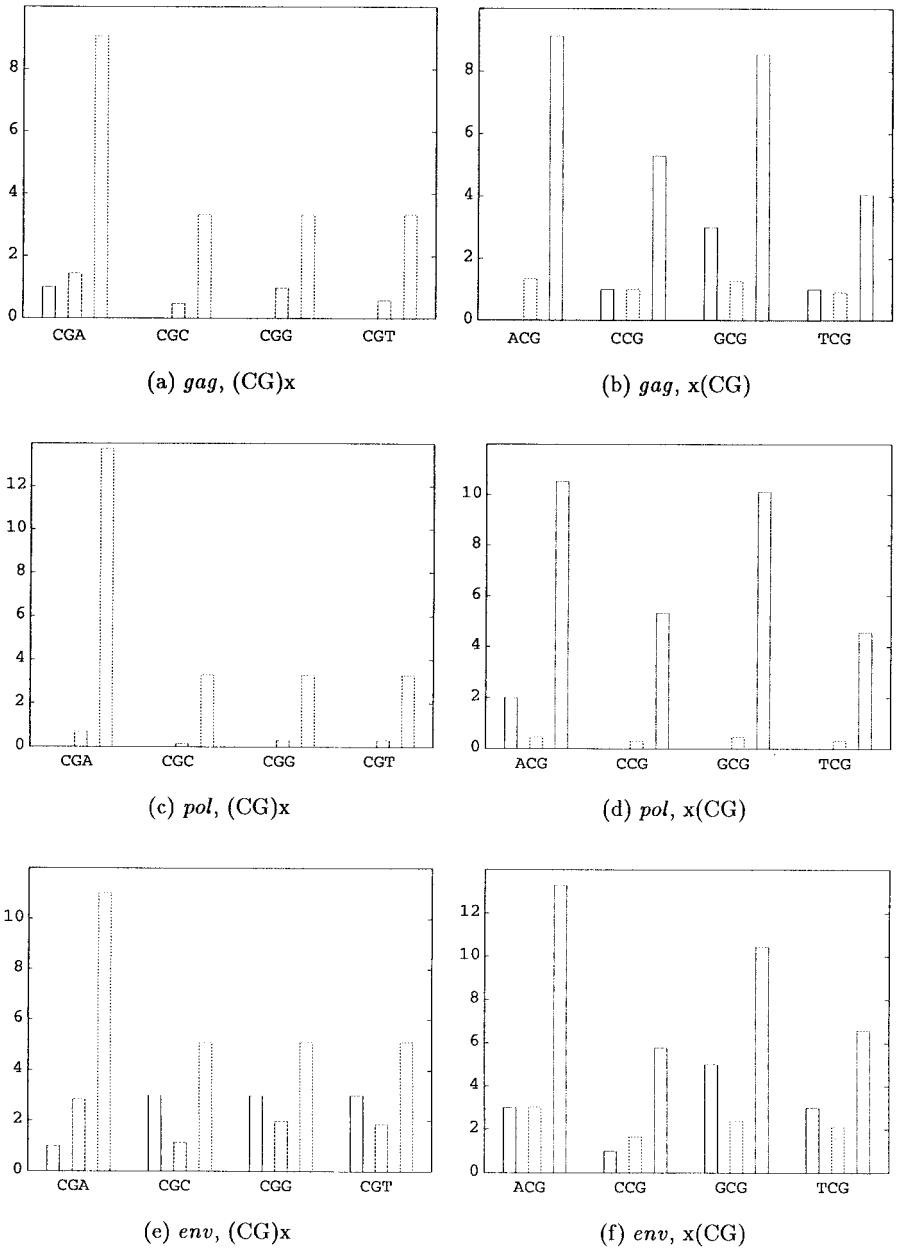
A CpG Depression Pentet-based Model

The dinucleotide counts shown in table 2 suggest that CpG depression is a feature that is equally important between and within codons. Further improvement in the description of codon equilibrium frequencies is likely to be obtained with a model that allows the equilibrium frequency of a codon to depend on the neighbor codons. In particular, equilibrium frequencies of codons that are preceded by C-ending codons are expected to be different from those that are preceded by non–C ending codons, since the frequency of the G nucleotide in the first codon position in these codons is likely to differ. A similar pattern is expected for the C nucleotide in the third position in that its frequency depends on whether it is followed by G- or a non–G starting codon. We measure the effect of the CpG dinucleotide depression at codon boundaries by adopting a pentet approach. In this approach, a pentet $i_1i_2i_3i_4i_5$ is thought of as a codon $i_2i_3i_4$ and its immediate nucleotide neighbors, i_1 and i_5, on each side. We consider three models, expressed as hypotheses H_1, H_2, and H_3, of the rate matrices operating on pentets. These models extend the idea behind the matrix that defines the CpG depression codon-based model. The first matrix incorporates no CpG depression, corresponding to the codon model with $\lambda = 1$. In the second matrix, only entries in the matrix corresponding to the generation or loss of a CpG within the three codon sites are modified, and this corresponds to the codon model with a freely varying λ. The third matrix is the full pentet model, and all entries corresponding to the generation or loss of a CpG within a pentet are modified—those where a CpG is generated or lost within the codon positions, as well as those in which a CpG is generated or lost at codon boundaries, that is, in dinucleotide sites i_1i_2 and i_4i_5. Since the first position in a pentet is also the last position in the preceding codon and the last position in a pentet is also the first position of the following codon, the frequency parameters were forced to be equal in the first and fourth pentet positions and in the second and fifth pentet positions, i.e., $\pi_i^1 = \pi_i^4$ and $\pi_i^2 = \pi_i^5$, $i \in \{A, C, G, T\}$. As in the CpG depression codon-based model, the pentet models thus contain three sets of nucleotide frequencies.

In addition to the models H_1, H_2, and H_3, we consider two models $\tilde{H}_1$ and $\tilde{H}_3$ in which we disregard the effect of codon position. The rate matrices in these models are similar to that of model H_3, with the restriction that $\pi_i^1 = \pi_i^2 = \pi_i^3$, $i \in \{A, C, G, T\}$. In model $\tilde{H}_1$, we assume no selection of CpGs and set $\lambda = 1$.

The equilibrium frequencies of the pentets in the models are given in table 5, where κ_m ($\tilde{\kappa}_m$) is the normalizing constant in model H_m, $m = 1, 2, 3$ ($\tilde{H}_k$, $k = 1, 3$). Note that in spite of the fact that model H_2 and H_3 have the same number of parameters, 10, model H_3 has three expressions for pentet equilibrium frequencies, whereas model H_2 only distinguishes between two types of pentets. We used the χ^2 test value based on the observed and expected codon counts under the models as a measure of discrepancy from the models. Expected pentet counts were found by the maximum-likelihood estimates, assuming a model that hypothesizes that pentets are drawn independently from a multinomial distribution, $m(n, \Pi^k)$ or $m(n, \tilde{\Pi}^k)$, where Π^k ($\tilde{\Pi}^k$) is the vector of pentet equilibrium frequencies in model k, $k = 1, 2, 3$ ($k = 1, 3$). Since the pentets overlap, they are not independent. However, this assumption is not essential to the analysis we make—the assumption is merely used to find suitable parameter estimates, and the χ^2 is not used as a test statistic but merely as an intuitive measure of discrepancy. Alternatively, estimates could have been made by minimizing the sum of quadratic deviances.

Due to the huge number of pentets ($4 \times 64 \times 4 = 1024$), we pooled the pentet counts for all genes and measured the discrepancy between these combined counts and those expected under each of the three pentet models. Results for the sequence HIV3202A12 are given in table 6, along with the means, minimum and maximum values of the χ^2's, and estimates of the λ^2 parameters obtained from analyses of all 28 sequences. The results clearly show that CpG depression is not a phenomenon that is restricted to within codons—it is equally pronounced at codon boundaries. The difference in



(a) *gag*, (CG)x (b) *gag*, x(CG)

(c) *pol*, (CG)x (d) *pol*, x(CG)

(e) *env*, (CG)x (f) *env*, x(CG)

FIG. 1.—Histograms of observed and expected codon counts for the codons containing CpG in the genes *gag*, *pol*, and *env* of the sequence HIV3202A12. Three boxes are plotted at each codon: the first is the observed codon count, the second is the count expected under the model H_{+CpG}, and the third is the count expected under model H_{-CpG}. Some CpG-containing codons are absent in the genes presented, resulting in an invisible first box.

Table 5
Expressions for the Equilibrium Frequencies of Pentets in the Different Pentet Models

	TYPE OF PENTET		
MODEL	Contains No CpG	Contains One CpG	Contain Two CpGs
H_1	$\kappa_1\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$	$\kappa_1\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$	$\kappa_1\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$
H_2	$\lambda^2\kappa_2\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$	$\kappa_2\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$	$\kappa_2\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$
H_3	$\lambda^4\kappa_3\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$	$\lambda^2\kappa_3\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$	$\kappa_3\pi_{i_1}^3\pi_{i_2}^1\pi_{i_3}^2\pi_{i_4}^3\pi_{i_5}^1$

NOTE.—Model H_1 allows no CpG depression, H_2 allows for depression of CpGs within codons, and H_3 allows for depression of CpGs within codons as well as across codon boundaries. Expressions for the equilibrium frequencies of the pentets in the models $\tilde{H}_1$ and $\tilde{H}_3$, in which effects of codon positions are disregarded, are obtained from those of H_1 and H_3 by setting $\pi_i^1 = \pi_i^2 = \pi_i^3$, $i \in \{A, C, G, T\}$.

Table 6
χ^2 Test Statistics for the Observed Versus Expected Pentet Frequencies and Estimates of λ^2 Under the Models $\tilde{H}_1$, $\tilde{H}_3$, H_1, H_2, and H_3 for All Genes Pooled in the Sequence HIV3202A12 (seq1)

	χ^2			$\hat{\lambda}^2$		
MODEL	Seq1	Average	Min–max	Seq1	Average	Min–max
$\tilde{H}_1$	2391.1	2355.25	1919.1–2526.0	—	—	—
$\tilde{H}_3$	2100.2	2066.7	1673.3–2215.5	6.0	6.0	4.9–8.0
H_1	2080.0	1977.0	1646.4–2080.0	—	—	—
H_2	1798.9	1705.5	1461.7–1798.9	9.2	9.0	6.7–13.2
H_3	1455.9	1388.2	1249.6–1479.7	12.2	11.6	9.4–14.0

NOTE.—Averages and minimal and maximal values are given for the tests performed on 26 sequences.

discrepancies between observed and expected numbers of pentets between the models H_2 and H_3 is of the same order of magnitude as that obtained for the models H_1 and H_2. Moreover, of the models H_2 and H_3, that have the same number of parameters, H_3 provides a much better explanation of the observed pentet counts. The χ^2 values obtained for the models $\tilde{H}_1$ and $\tilde{H}_3$ are considerably higher than those for H_1 and H_3, and this indicates that even though the CpG depression is strong, the effect of codon positions is more pronounced.

We conclude that an evolutionary model that is able to take into account both the between-codon and the within-codon CpG depressions is a promising candidate in the search for a model which better fits the codon frequencies observed in the sequences and thereby also better describes the evolutionary process that has generated the sequences. Unfortunately, the pentet model presented here does not serve as such a model—an evolutionary analysis using this model is not mathmatically tractable. The problem arises because the pentets overlap. In calculating the likelihood of observing two (or more) sequences in the codon-based models, we need to make the assumption that codons are independent. With this assumption, the probability of evolving one codon from another is all we need to derive the probability of evolving any full sequence from another sequence, given the model. The likelihood of two observed sequences is obtained by multiplying these transition probabilities and an appropriate set of codon equilibrium frequencies. The probabilities of evolving one pentet into another cannot easily produce the probability of evolving one sequence into another because overlapping pentets are not independent. The lack of independence cannot be overcome by conditioning, that is, by calculating the probabilities of evolving a given codon from another, conditioned on the type of pentet in which the codon is placed, because the pentet of a codon may change over time as substitutions occur.

Evolutionary Analyses of HIV1 Genes

Three pairwise alignments of a subtype-B sequence, HIVSF2, were analyzed. The sequence HIVSF2 was aligned with another subtype-B sequence, HIVLAI; with a subtype-D sequence, HIVELI; and with a subtype-A sequence HIVMAL, using the alignment program GENAL (Hein and Støvlbæk 1994). Three versions of the CpG depression codon-based model, H_A, H_B, and H_C, were used, and the analysis focused on the single-coding regions of each of the genes *gag*, *pol*, and *env*. The *gag* region contained 431 codons, *pol* contained 915 codons in all three pairwise alignments, and the *env* region contained 751, 753, and 748 codons in the alignments of HIVLAI and HIVSF2, HIVELI and HIVSF2, and HIVMAL and HIVSF2, respectively. Model H_A assumes that the nucleotide frequencies in the three codon positions are identical (i.e., $\pi_i^1 = \pi_i^2 = \pi_i^3$, $i \in \{A, C, G, T\}$) and that there is no selection against CpGs (i.e., $\lambda = 1$). In model H_B, the frequencies of the nucleotides are allowed to differ in the three codon positions and still, $\lambda = 1$. Model H_C is the unrestricted CpG depression codon-based model. Maximum-likelihood estimates were found and goodness of fit tests were performed for each of the models on each of the genes in each pairwise alignment by the method developed by Goldman (1993). In the goodness of fit tests performed, the unrestricted model, H_0, that is used as a basis for the tests was, in all cases, a model in which each possible pattern of codons in a column in the alignment is given a probability, p_k, on which there are no restrictions except that the p_k's add to 1. Two hundred simulations were made for each goodness of fit test.

The maximum-likelihood estimates of the $\pi_{i_k}^k$ parameters from the alignment of the sequences HIVELI and HIVSF2 are given in table 7. The estimates of these parameters from the remaining two alignments were very similar to those in table 7. In table 8, maximum-likelihood estimates of the additional parameters in the three models are given. The last three columns of table 8 provide the maximum-log likelihood values of each of the three models, and the likelihood ratio test statistics, $-2 \log Q$, for model H_B under H_A, model H_C under H_B, and for model H_i under the unrestricted model H_0, $i = A, B, C$. Again, only results obtained from the HIVELI–HIVSF2 alignment are shown, since similar patterns were exhibited by the results for the remaining alignments.

The likelihood ratio test statistics for model H_A under model H_B and for model H_C under H_B are approximately $\chi^2(6)$ and $\chi^2(1)$ distributed, respectively. The 95th percentiles of these distributions are 12.6 and 3.84, and neither the hypothesis of equal base frequencies in the three codon positions nor the hypothesis of $\lambda = 1$ is accepted for any of the genes in any of the alignments analyzed (table 8). The hypothesis $\lambda = 1$ is rejected by a larger margin than the hypothesis $\pi_i^1 = \pi_i^2 = \pi_i^3$, $i \in$

Table 7
ML Estimates of the $\pi_{i_k}^k$ Parameters, $k = 1, 2, 3$, $i_k \in \{A, C, G, T\}$, Obtained Under the Models H_A, H_B, and H_C in the Analyses of the Three Single-coding Regions of the *gag*, *pol*, and *env* Genes, in the HIVELI–HIVSF2 Alignment

Gene	Model	$\hat{\pi}_A^1$	$\hat{\pi}_C^1$	$\hat{\pi}_G^1$	$\hat{\pi}_T^1$	$\hat{\pi}_A^2$	$\hat{\pi}_C^2$	$\hat{\pi}_G^2$	$\hat{\pi}_T^2$	$\hat{\pi}_A^3$	$\hat{\pi}_C^3$	$\hat{\pi}_G^3$	$\hat{\pi}_T^3$
gag	H_A	.390	.173	.236	.200	.390	.173	.236	.200	.390	.173	.236	.200
	H_B	.331	.175	.297	.197	.368	.206	.210	.216	.474	.140	.198	.187
	H_C	.315	.213	.283	.189	.331	.237	.237	.195	.446	.133	.245	.177
pol	H_A	.409	.145	.215	.232	.409	.145	.215	.232	.409	.145	.215	.232
	H_B	.317	.158	.301	.224	.401	.169	.178	.251	.506	.110	.165	.219
	H_C	.305	.189	.289	.217	.371	.192	.204	.232	.485	.106	.200	.210
env	H_A	.373	.158	.219	.250	.373	.158	.219	.250	.373	.158	.219	.250
	H_B	.373	.143	.247	.237	.333	.190	.216	.261	.409	.145	.195	.251
	H_C	.359	.175	.237	.230	.306	.219	.236	.239	.387	.139	.235	.239

{*A, C, G, T*}. The goodness of fit tests restate these findings. The results for the tests of each of the models on each of the genes in the alignment of the sequences HIVELI and HIVSF2 are given in figure 2. Similar results were obtained for the remaining alignments (results not shown). The margin by which model H_A is rejected is reduced by approximately ⅓ when the model is extended to H_B, that is, to allow for different base frequencies in the three codon positions. A similar reduction in the margin of rejection is obtained when model H_B is extended to H_C, that is, to allow for selection against CpGs. This goes for all the single-coding regions of the genes *gag*, *pol*, and *env* in all three alignments analyzed. Moreover, the CpG depression codon-based model is almost able to account for the evolution observed in the relatively short gene, *gag*—the model is accepted by tests at the 99% confidence levels for the HIVELI–HIVSF2 (fig. 2) and HIVMAL–HIVSF2 (not shown) alignments, and at the 95% level for the HIVLAI–HIVSF2 alignments (not shown). The performance of this model is worse on the longer *pol* and *env* genes.

The estimates of the parameters α and β and of the selection factor *f* are relatively unaffected by the choice of model (table 8). The estimates $\hat{\alpha}$ and $\hat{\beta}$ rise with the number of parameters in the model, whereas the estimate of *f* is largest for model H_A and smallest for model H_B. The estimates of these parameters differ by at most 10% in the different models. The ratio $\hat{\alpha}/\hat{\beta}$ of the estimates of the transition to the transversion rate coefficient obtained are different in the three genes but similar in the analyses of the three alignments of the same gene. This transition/transversion ratio lies within the intervals 3.66–4.58, 8.23–9.10, and 2.64–3.18 for the *gag*, *pol*, and *env* genes in the three alignments. An exception is the *pol* gene in the alignment of the subtype A sequence HIVMAL to HIVSF2, where the ratios are within the interval 4.69–5.08. Thus, a more pronounced skewness toward transitions is seen in the *pol* gene. The estimates of λ in model H_C show the same pattern of differences between genes and similarity among the same genes in different alignments. The estimates of *f* vary somewhat more. The estimate $\hat{\lambda}$ is larger for *pol* than for *gag* and *env*. The selection against substitutions that change the amino acid is considerably stronger in *gag* and *pol* than it is in *env*.

The choice of model has a strong effect on the estimates of the $\pi_{i_k}^k$ parameters. In model H_B, compared to H_A, the estimate of the frequency of base A in the third position of the codon and of the frequency of base G in the first position are considerably higher, and the estimate of the frequency of base C in the second position is somewhat higher. The frequency of A in the first position, G in the second, and C and G in the third position are lowered under H_B, relative to H_A. The effect of extending the codon-based model with unequal nucleotide frequencies in the three codon positions (H_B) to allow for selection against CpGs (H_C) is that $\hat{\pi}_C^1$, $\hat{\pi}_C^2$, $\hat{\pi}_G^2$ and $\hat{\pi}_G^3$ are considerably increased. This effect is balanced by a general reduction in the estimates of the remaining $\pi_{i_k}^k$ parameters.

Expected number of substitutions per codon, ρ, the ratios of the rates of synonymous to nonsynonymous

Table 8
ML Estimates of α*t*, β*t*, λ, *f*, and Log $\hat{L}$ Values Obtained Under the Models H_A, H_B, and H_C in the Analyses of the Three Single-coding Regions of the *gag*, *pol*, and *env* Genes, in the Alignment HIVELI–HIVSF2

Gene	Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\lambda}$	$\hat{f}$	$\log\hat{L}$	$-2\log Q_{(i+1)i}$	$-2\log Q_{i0}$
gag.	H_A	.394	.087	—	.307	−2043.5	—	480.9
	H_B	.417	.091	—	.290	−2027.6	31.8	449.1
	H_C	.427	.095	3.052	.293	−1997.9	59.4	389.7
pol	H_A	.634	.073	—	.177	−4243.4	—	746.2
	H_B	.704	.077	—	.163	−4185.3	116.6	630.0
	H_C	.718	.081	4.046	.164	−4126.4	117.8	512.1
env.	H_A	.572	.214	—	.624	−4206.4	—	1079.1
	H_B	.585	.222	—	.600	−4196.2	20.4	1058.6
	H_C	.599	.226	2.806	.606	−4148.1	96.2	962.5

NOTE.—The last two columns contain the $-2\log Q$ test statistics for model H_A under H_B, for model H_B under H_C, and for model H_x under H_0, $x = A, B, C$.

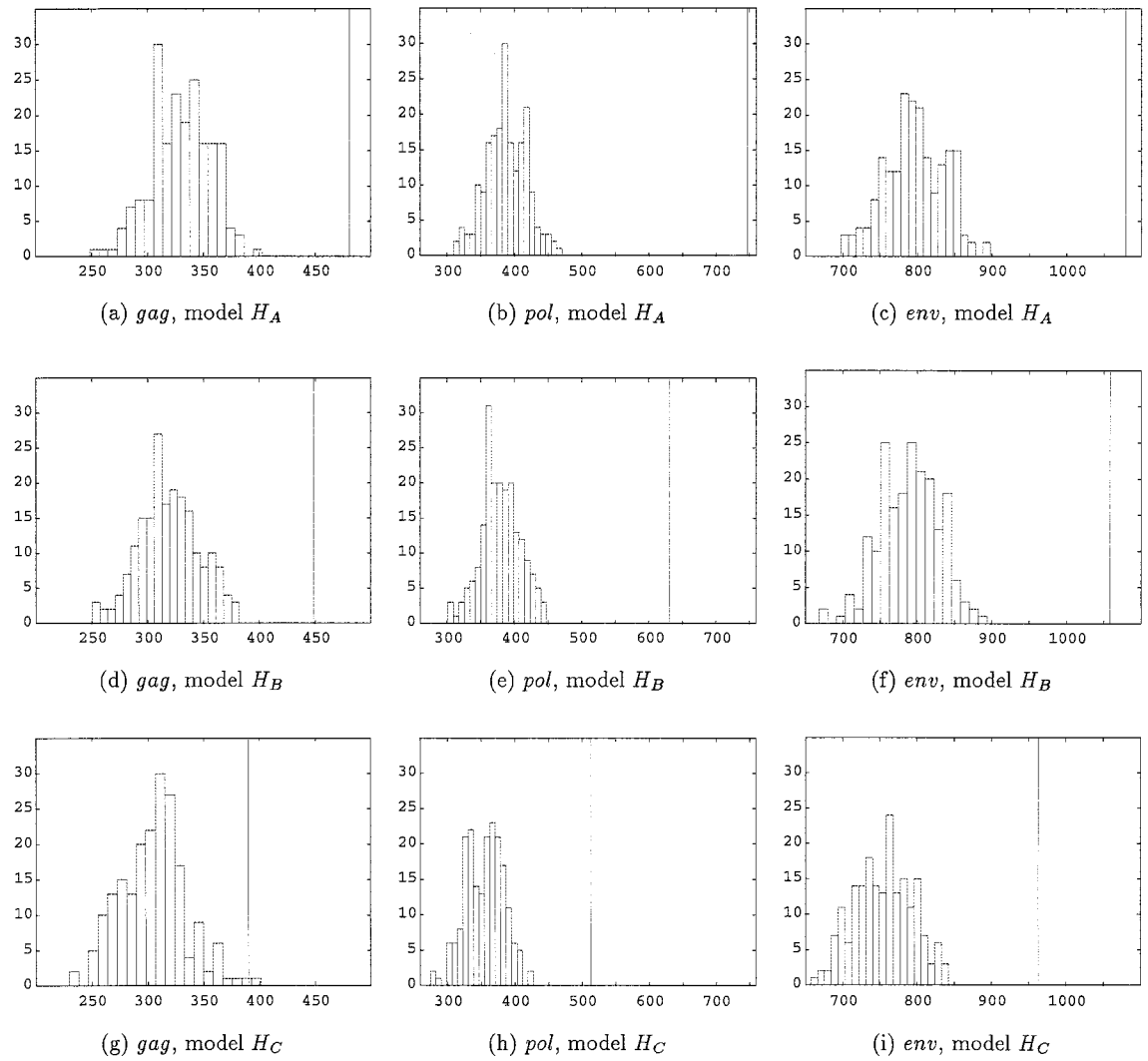


FIG. 2.—Goodness of fit tests for the models H_A, H_B, and H_C on each of the regions in which the genes *gag*, *pol*, and *env* are encoded in the alignment of the sequences HIVELI and HIVSF2. Vertical lines indicate observed $-2\log Q$'s (table 8). The histograms are the distributions of the $-2\log Q$'s obtained from simulations.

substitutions, ρ_s/ρ_n, and the ratios of transitions to transversions, ρ_{ts}/ρ_{tv}, were calculated for the three models. The estimates of these quantities were obtained by inserting the maximum-likelihood estimates of the parameters in the expressions

$$\rho = \sum_{\substack{j \\ j \neq i}} \pi_i q_{ij}$$

$$\rho_s/\rho_n = \left(\sum_{\substack{j,j \neq i \\ aa_i = aa_j}} \pi_i q_{ij} \right) \left(\sum_{\substack{j,j \neq i \\ aa_i \neq aa_j}} \pi_i q_{ij} \right)$$

$$\rho_{ts}/\rho_{tv} = \left(\sum_{\substack{j,j \neq i \\ ts}} \pi_i q_{ij} \right) \left(\sum_{\substack{j,j \neq i \\ tv}} \pi_i q_{ij} \right),$$

where π_i is the equilibrium frequency of codon i and q_{ij} is the ij'th entry in the rate matrix Q. The conditions under the summation signs in the ratios indicate that only entries corresponding to synonymous, nonsynonymous, transition, and transversion substitutions, respectively, are included in the sum.

The estimates of ρ, ρ_s/ρ_n, and ρ_{ts}/ρ_{tv} were largely unaffected by the choice of model (results not shown). The rate ρ was slightly increased in model H_B relative to H_A (up to 2%), but no systematic change appeared between model H_B and H_C. In H_B, the ratio of synonymous to nonsynonymous substitutions was a few percent higher, and the ratio of transitions to transversions was a few percent lower compared to the values under H_A and H_C.

We also calculated estimates of the ratios of CpG-generating substitutions to all other substitutions, ρ_{+CpG}/ρ_{-CpG} (table 9). These are obtained by inserting the maximum-likelihood estimates of the parameters in the expression

$$\rho_{+C_pG}/\rho_{-C_pG} = \left(\sum_{\substack{j,j \neq i \\ +C_pG}} \pi_i q_{ij} \right) \left(\sum_{\substack{j,j \neq i \\ -C_pG}} \pi_i q_{ij} \right),$$

where the subscripts $+CpG$ and $-CpG$ indicate that only entries corresponding to the generation of a CpG and no generation of a CpG, respectively, are included

Table 9
Estimates of ρ_{+CpG}/ρ_{-CpG} Calculated Using the ML Estimates Obtained Under the Models H_A, H_B, and H_C, in the Analysis of the Alignment HIVELI–HIVSF2

Model	*gag*	*pol*	*env*
H_A	0.076	0.064	0.063
H_B	0.081	0.070	0.063
H_C	0.039	0.024	0.032

in the sums. These estimates are highly affected by the choice of model. These ratios are in general increased by 5–9% when moving from model H_A to H_B and decreased by 50–70% in model H_C relative to H_B.

Plots of ρ, ρ_s/ρ_n, ρ_{ts}/ρ_{tv}, and ρ_{+CpG}/ρ_{-CpG} as functions of λ under the CpG depression codon-based model are given in figure 3. The remaining parameters were fixed at the maximum-likelihood values obtained in the analysis of the alignment HIVELI–HIVSF2. As functions of λ, the expected number of substitutions per codon, the ratio of the rates of synonymous and nonsynonymous substitutions, and the ratio of transitions to transversions per codon are almost constant for $\lambda > 1$. The rate ρ is reduced by 7.6%, 8.6%, and 4.8% in *gag*, *pol*, and *env*, respectively, when the value of λ is raised from 1 to the maximum-likelihood values, 3.052, 4.046, and 2.806 (table 8). The ρ_s/ρ_n ratios are decreased by 9.4%, 7.6%, and 8.2% and the ρ_{ts}/ρ_{tv} ratios are increased by less than 1% when the quantities are compared for the same λ values. In contrast, the ρ_{+CpG}/ρ_{-CpG} ratios are highly affected by the value of λ. This ratio is decreased by 62.8% when the maximum-likelihood value of λ in the *gag* single-coding region is used relative to the value for $\lambda = 1$, and by 72.4% and 60.9% for the similar comparisons for the *pol* and *env* genes.

The values of ρ, ρ_s/ρ_n, and ρ_{ts}/ρ_{tv} obtained under model H_B and H_C did not exhibit the differences expected from figure 3. These values appear to be robust to deviations from the model's assumption regarding selection against CpGs. This robustness was confirmed by

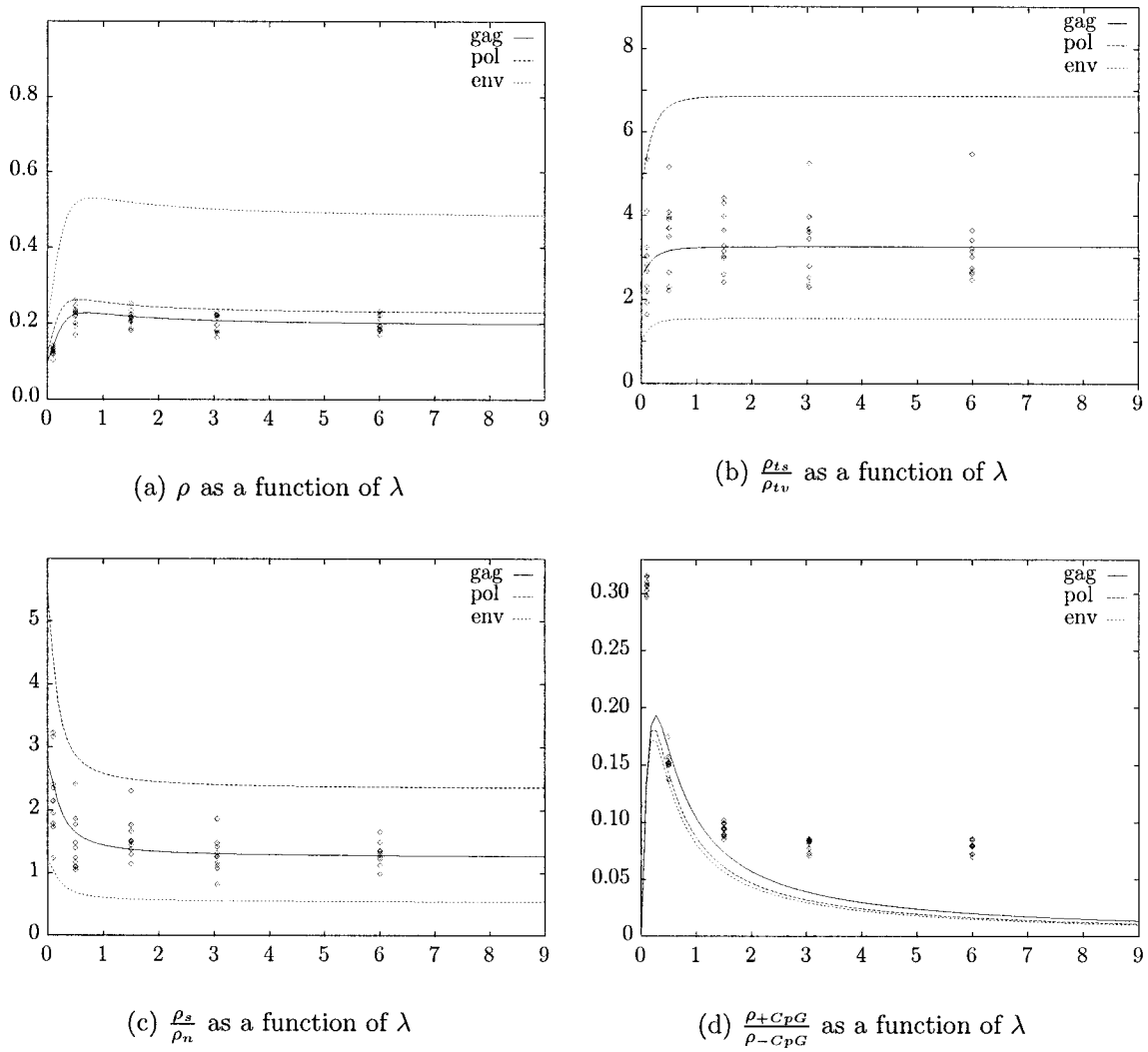


FIG. 3.—Plots of different quantities of interest as functions of λ under the CpG depression codon-based model. Maximum-likelihood estimates obtained in the analysis of the *gag* gene in the HIVELI–HIVSF2 alignment were used for the remaining parameters. The results of the analysis of robustness are shown by diamonds; data simulated under model H_C for various values of λ were analyzed using model H_B ($\lambda = 1$).

simulations. For each of five values of λ, 0.1, 0.5, 1.5, 3.05, and 6.0, we simulated 10 alignments of 421 codons each according to the CpG depression model, using the maximum-likelihood estimates of the remaining parameters obtained in the analysis of the *gag* region in the HIVELI–HIVSF2 alignment. Each simulated data set was analyzed using model H_B ($\lambda = 1$). For each analysis, ρ, ρ_s/ρ_n, ρ_{ts}/ρ_{tv}, and ρ_{+CpG}/ρ_{-CpG} were calculated; this data is summarized in figure 3. The values of the quantities ρ, ρ_s/ρ_n, and ρ_{ts}/ρ_{tv} obtained using the model H_B are well in accordance with the values under H_C as given by the graphs (fig. 3). The simulations confirm the sensitivity of the estimate of the quantity ρ_{+CpG}/ρ_{-CpG} to the choice of model—the estimate obtained when the model H_B is used to analyze data generated under H_C is seriously biased.

In conclusion, H_C provides a significantly better description of the evolutionary processes in the HIV1 genes analyzed than model H_B, which in turn performs significantly better than H_A. H_C is almost able to describe the evolution exhibited by the *gag* gene. There are major differences in the estimates of the parameters $\pi^k_{i_k}$ obtained under the three models and some differences in the remaining parameter estimates. The estimates of genetic distance, ρ, and of the synonymous to nonsynonymous and the transition to transversion rates, ρ_s/ρ_n and ρ_{ts}/ρ_{tv}, are little affected by the model used, whereas that of ρ_{+CpG}/ρ_{-CpG} is highly affected.

Discussion

By incorporating two features characteristic of lentiviruses—unequal nucleotide compositions in the three codon positions and selection against CpGs—in a codon-based model, we arrived at a model under which the expected codon frequencies are considerably closer to those observed in the genes *gag*, *pol*, and *env* in the HIV1 genome than are those of simpler models. The frequencies expected almost fit those observed in the *gag* genes. By extending the model to operate on pentets, we demonstrated that a further improvement in the description is obtained when the selection against CpGs is extended to operate across codon boundaries. Thus, part of the codon usage observed in the examined lentiviral genes can be explained by a dinucleotide bias—a bias that is not restricted to within-codon positions but is present across codon boundaries as well. The general idea in the modeling is simple, and the model can easily be extended to incorporate other selective pressures similar to that against CpGs. Exact expressions for the equilibrium frequencies of the codons under models of this type are readily obtainable—they are of the same form as those obtained in the CpG depression codon-based model.

The CpG depression codon-based model provides a significantly better description of the evolutionary process relating pairs of the genes in the HIV1 sequences examined than simpler models do. Indeed, the model was just at the border of being accepted by a goodness of fit test as an adequate description of the evolution exhibited by the *gag* genes in the pairwise alignments analyzed. The performance of the model on the longer *pol* and *env* genes was less impressive. We believe that the increase in fit obtained when the full CpG depression codon-based model is used, rather than the model with $\lambda = 1$, is mainly due to the better description of the equilibrium frequencies of the codons and much less due to a better explanation of the substitutions seen in the alignments. The worse performance of the model on the longer genes indicates remaining features that need to be properly modeled in order to obtain an adequate model for the evolution of the HIV1 genes. The results of the pentet approach argue for the development of a model that allows an evolutionary analysis without the assumption of independence of the substitution processes in individual codons and thus allows the incorporation of selection against CpGs at codon boundaries.

Substantial evidence exists for rate variation across sites in HIV1 genes (Starcich et al. 1986), and extension of the model to allow sites to have different rates of substitution is likely to result in improved performance. However, when only pairs of sequences are considered, we suspect that the improvement will be moderate. The effect of rate variation is most obvious in alignments of many sequences, where some columns may be highly variable and others rather conserved—for pairwise alignments, columns exhibit either a constant pattern or two different nucleotides. Thus, a pairwise alignment provides limited information about the rate of substitution at individual sites. Pairwise alignments, however, may pinpoint highly variable or almost invariable regions in the alignment. The proper way to model rate variation in pairwise alignments may therefore be the hidden Markov model suggested by Felsenstein and Churchill (1996), in which the actual order of the sites in the sequence are of importance, rather than the method developed by Yang (1993), where rates of substitutions in different sites are drawn independently from a Γ-distribution.

Selection against CpGs provides more homogeneous estimates of $\pi^k_{i_k}$, $i_k \in \{A, C, G, T\}$, $k = 1, 2, 3$ (H_B to H_C in table 7). Thus, part of the bias in nucleotide frequencies observed in the genes is caused by the selection against CpGs. The $\pi^k_{i_k}$ parameters in the model H_B ($\lambda = 1$) may be viewed as nucleotide frequencies obtained after selection against CpGs has occured. In the full CpG depression codon-based model, these parameters reflect the frequencies before this selection occurs.

For the data sets analyzed and the sets of parameters used in the simulations, the estimate of the genetic distance between two sequences is not affected by assuming $\lambda = 1$ in the full CpG depression codon-based model. Neither are the estimates of the ratios of the rates of synonymous to nonsynonymous substitutions or of transitions to transversions affected. As expected, the estimate of the ratio of the CpG-generating to non–CpG generating substitution rates is seriously affected by the choice of model; this ratio was overestimated by a factor of two to three when the $\lambda = 1$ model was used. These studies on bias are rather preliminary; the effects need to be assessed for a wider range of parameter values.

Acknowledgments

We thank Jotun Hein for many useful discussions. Morten Lauritsen and Ellen Agerbo Jensen are thanked for computing assistance. C. W. was supported by the Danish National Research Foundation.

LITERATURE CITED

BAUMANN, U. 1996. Biases and balances within the genetic code. Biochem. Biophys. Res. Comm. **219**:543–547.

BERKHOUT, B., and F. J. VAN HEMERT. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased composition of HIV proteins. Nucleic Acids Res. **22**:1705–1711.

COULONDRE, C., and J. H. MILLER. 1978. Molecular basis of base substitution hotspots in *Escherichia coli.* Nature **274**: 775–780.

FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: a justification. Evolution **46**:16–24.

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

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

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. 1994. Genomic alignment. J. Mol. Evol. **38**:310–316.

HEMERT, F. J. VAN, and B. BERKHOUT. 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J. Mol. Evol. **42**:132–142.

HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–231.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. M. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

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

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

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

SHPAER, E. G., and J. I. MULLINS. 1990. Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression. Nucleic Acids Res. **18**:5793–5797.

STARCICH, B. R., B. H. HAHN, G. M. SHAW et al. (11 coauthors). 1986. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell **45**:637–648.

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

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. Mol. Biol. Evol. **39**:315–329.

STANLEY A. SAWYER, reviewing editor

Accepted May 11, 1998