J Mol Evol (1995) 40:181–189

JOURNAL OF MOLECULAR EVOLUTION



# A Maximum-Likelihood Approach to Analyzing Nonoverlapping and Overlapping Reading Frames

**Jotun Hein, Jens Støvlbæk**

Institute of Biological Sciences, Aarhus University, DK-8000 Aarhus, Denmark

Received: 20 July 1994 / Revised and Accepted: 1 September 1994

**Abstract.** A model is presented for sequence evolution on the basis of which one can analyze combinations of noncoding, singly coding, and multiply coding regions of aligned homologous DNA sequences. It is a generalization of Kimura's (*J. Mol. Evol.* 16:111–120, 1980) and Li et al.'s (*J. Mol. Evol.* 36:96–99, 1985) transition-transversion models with selection on replacement substitutions.

Based on a hierarchy of hypotheses, one will be able to estimate selection factors and transition and transversion distances for different combinations of regions ranging from many regions, each with their private set of parameters, to one set of parameters for all regions.

The method is demonstrated on two aligned HIV1 retroviruses.

**Key words:** Maximum likelihood — Reading frames — Sequence evolution

## Introduction

In the analysis of homologous sequences it is important to estimate how many events have occurred in their history, going back to their most recent common ancestor. The central problem is that several substitutions occurring at the same position in the molecule can never look like more than one substitution. Just counting the number of differences between two aligned molecules is bound to underestimate the real number of substitutions. As sequences analyzed get longer, they will most likely have a mosaic of noncoding, singly coding, and possibly multiply coding regions, as shown in Fig. 1 in the case of two aligned HIV1 viruses imposing a variety of restrictions on different regions. Several new complications occur relative to a region with only one reading frame. Are the substitution rates the same for different regions? Are the selection pressures on different regions the same? Can the selection effects in regions with overlapping reading frames be separated into the effects from each reading frame? Traditional models will be generalized to cover these situations.

A number of models (Fig. 2) have been devised to analyze sequence evolution. The first model was the Jukes-Cantor (1969) one-parameter model. It gives an estimate of how much evolution has actually occurred and the variance of this estimate. After the accumulation of sequence data it was realized that transitions (purine-to-purine or pyrimidine-to-pyrimidine substitutions) occur much more frequently than transversions. This led Kimura (1980) to introduce his two-parameter model, which was a significant step toward more realistic models.

The analysis of sequences coding for proteins is complicated by the selection against the substitutions that cause amino acid replacements. For single coding sequences this causes substitutions to be partitioned into synonymous (also called *silent*) and replacement (also called *nonsynonymous*) substitutions. To extend models from noncoding sequences to coding sequences is difficult because of the irregular restrictions introduced by

*Correspondence to:* J. Hein

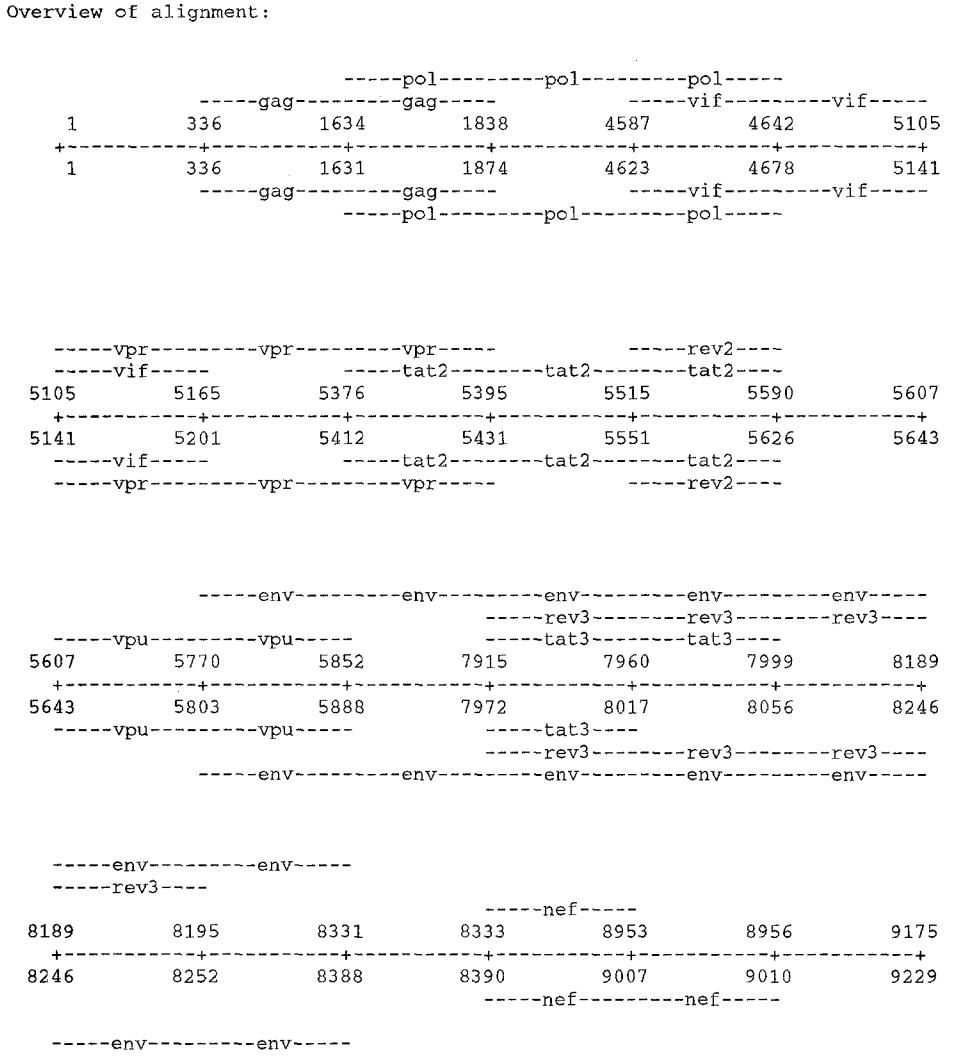



**Fig. 1.** Summary of alignment of homologous HIV1 genomes (Hein and Støvlbæk 1994). The first 336 base pairs of both HIVELI and HIVBRU are noncoding; then follow the single coding regions for *gag,* followed by a region that codes for both *gag* and *pol.* Each time the configuration of genes in one or both genes changes a new region starts. In this alignment there are 24 regions.

the genetic code (Fig. 3). A gordian knot solution to this was taken by Nei and Gojobori (1986). They partitioned both the positions and substitutions into synonymous and replacement components. Then the process for the whole sequences was partitioned into two processes—one synonymous and one replacement. Each process was described as a Jukes-Cantor process. This method was generalized to cover overlapping reading frames, but it fared badly as it ignored transition-transversion bias, which becomes especially acute in the overlapping case.

Li et al. (1985) extended Kimura's model to coding regions by assuming that the genetic code was more regular than it actually is. It was assumed that varying any nucleotide in a codon among the four possible nucleotides would give either the same amino acid (fourfold degenerate) (4); two amino acids, each pair differing by a transition (twofold degenerate) (2-2); and lastly, 4 different amino acids would be obtained (nondegenerate (1-1-1-1). This gives simplicity to the analysis of the problem (Fig. 3). It is problematic for following reasons:

First, the codons ATx code for three isoleucins and one methionine. They will be treated as 1-1-1-1 sites, implying that all substitutions within this group will be regarded as replacement.

Second, the stop codons cannot be expected to be selected against in the same manner as amino acid replacements, so they must either be ignored or given an unsatisfactory treatment. Here, they will be treated as ordinary replacement changes.

Third, there are several (2-1-1) cases (e.g., xTG). Fortunately, if one sequence is regarded as the ancestor, they can be reduced to the 2-2 case or the 1-1-1-1 case. If the starting nucleotide is one of the two synonymous codons, the probability distribution will be exactly as the 2-2 case and if the starting codon was one of the two singular codons, the probability distribution will be exactly as the

Rate-matrix:

| | | T | | O | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| F | A | $-2*\beta-\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| R | C | $\beta$ | $-2*\beta-\alpha$ | $\beta$ | $\alpha$ |
| O | G | $\alpha$ | $\beta$ | $-2*\beta-\alpha$ | $\beta$ |
| M | T | $\beta$ | $\alpha$ | $\beta$ | $-2*\beta-\alpha$ |

$a = \alpha*t$ $b = \beta*t$

X(a,b) = (1/4)(1 + exp(-4*b) + 2exp(-2*(a+b))) = Probability that two nucleotides are identical after time t.

Y(a,b) = (1/4)(1 + exp(-4*b) - 2exp(-2*(a+b))) = Probability that two nucleotides differ by a transition after time t.

Z(a,b) = (1/2)(1-2exp(-4*b)) = Probability that two nucleotides differ by a transversion after time t.

Expected number of substitutions occurred if the position averagely differs by P observable transitions and Q observable transversions.

Transitions: -0.5*ln(1-2P-Q) + 0.25*ln(1-2Q)

Transversions: - 0.25*ln(1-2Q)

**Fig. 2.** The Kimura two-parameter model. The Jukes-Cantor model is obtained by setting $a = b$ in the Kimura model. Empirically it is known that $a$ is higher than $b$. Each row sums to 0.0 as this is a rate matrix and not a transition probability matrix. The transition probability matrix, $X()$, $Y()$, and $Z()$, shows that after $t = 0.0$ no change will have occurred, but as $t$ increases the probability will start flowing into the nucleotide a transition away from the starting nucleotide. As $t$ becomes large $X()$, $Y()$, and $Z()$ will converge toward 0.25, 0.25, and 0.5, respectively. These quantities can be inverted to obtain estimates of transition and transversion coefficients from observed differences between the sequences. If two sequences averagely differ by $P$ transitions and $Q$ transversions, then the real—i.e., also parallel and superimposed substitutions, number of transitions, and number transversions—can be calculated by the formulae lowest in the figure.

1-1-1-1 case. Also, the codons CGG and GGG are synonymous, although they differ by a transversion in the first position. Following Li et al. (1985) this will be regarded as a transition.

Lastly, the model assumes that each matched nucleotide pair can be regarded independently of the other, which is not the case either: For instance, a mutation in the first nucleotide in a codon can convert the third nucleotide from a (2-2) type to a (4) type.

In the universal genetic code there are $4^2$ codons, when fixing one position. This position could experience three substitutions. This gives 48 possible codon pairs differing by one substitution. Of these two are 2-1-1 (xGA, xGG) with the transversion being silent and one is 3-1 (ATx). Further, seven codon pairs involve stop codons.

Given the general problems in making a model of a coding sequence, such as uneven codon usage and uneven selective constraint along the protein sequence, the above approximations are most likely minor. What will

First position:

| x = | A | C | G | T | |
|---|---|---|---|---|---|
| site xAA: | Lys | Gln | Glu | !!! | 1 : 1 : 1 : 1 |
| site xAC: | Asn | His | Asp | Tyr | 1 : 1 : 1 : 1 |
| site xAG: | Lys | Gln | Glu | !!! | 1 : 1 : 1 : 1 |
| site xAT: | Asn | His | Asp | Tyr | 1 : 1 : 1 : 1 |
| site xCA: | Thr | Pro | Ala | Ser | 1 : 1 : 1 : 1 |
| site xCC: | Thr | Pro | Ala | Ser | 1 : 1 : 1 : 1 |
| site xCG: | Thr | Pro | Ala | Ser | 1 : 1 : 1 : 1 |
| site xCT: | Thr | Pro | Ala | Ser | 1 : 1 : 1 : 1 |
| site xGA: | Arg | Arg | Gly | !!! | 2 : 1 : 1 |
| site xGC: | Ser | Arg | Gly | Cys | 1 : 1 : 1 : 1 |
| site xGG: | Arg | Arg | Gly | Trp | 2 : 1 : 1 |
| site xGT: | Ser | Arg | Gly | Cys | 1 : 1 : 1 : 1 |
| site xTA: | Ile | Leu | Val | Leu | 2 : 1 : 1 |
| site xTC: | Ile | Leu | Val | Phe | 1 : 1 : 1 : 1 |
| site xTG: | Met | Leu | Val | Leu | 2 : 1 : 1 |
| site xTT: | Ile | Leu | Val | Phe | 1 : 1 : 1 : 1 |

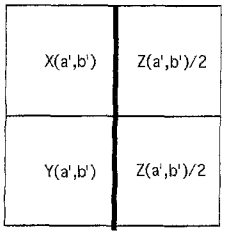
Second position:

| | | | | | |
|---|---|---|---|---|---|
| site AxA: | Lys | Thr | Arg | Ile | 1 : 1 : 1 : 1 |
| site AxC: | Asn | Thr | Ser | Ile | 1 : 1 : 1 : 1 |
| site AxG: | Lys | Thr | Arg | Met | 1 : 1 : 1 : 1 |
| site AxT: | Asn | Thr | Ser | Ile | 1 : 1 : 1 : 1 |
| site CxA: | Gln | Pro | Arg | Leu | 1 : 1 : 1 : 1 |
| site CxC: | His | Pro | Arg | Leu | 1 : 1 : 1 : 1 |
| site CxG: | Gln | Pro | Arg | Leu | 1 : 1 : 1 : 1 |
| site CxT: | His | Pro | Arg | Leu | 1 : 1 : 1 : 1 |
| site GxA: | Glu | Ala | Gly | Val | 1 : 1 : 1 : 1 |
| site GxC: | Asp | Ala | Gly | Val | 1 : 1 : 1 : 1 |
| site GxG: | Glu | Ala | Gly | Val | 1 : 1 : 1 : 1 |
| site GxT: | Asp | Ala | Gly | Val | 1 : 1 : 1 : 1 |
| site TxA: | !!! | Ser | !!! | Leu | 2 : 1 : 1 |
| site TxC: | Tyr | Ser | Cys | Phe | 1 : 1 : 1 : 1 |
| site TxG: | !!! | Ser | Trp | Leu | 1 : 1 : 1 : 1 |
| site TxT: | Tyr | Ser | Cys | Phe | 1 : 1 : 1 : 1 |

Third position:

| | | | | | |
|---|---|---|---|---|---|
| site AAx: | Lys | Asn | Lys | Asn | 2 : 2 |
| site ACx: | Thr | Thr | Thr | Thr | 4 |
| site AGx: | Arg | Ser | Arg | Ser | 2 : 2 |
| site ATx: | Ile | Ile | Met | Ile | 3 : 1 |
| site CAx: | Gln | His | Gln | His | 2 : 2 |
| site CCx: | Pro | Pro | Pro | Pro | 4 |
| site CGx: | Arg | Arg | Arg | Arg | 4 |
| site CTx: | Leu | Leu | Leu | Leu | 4 |
| site GAx: | Glu | Asp | Glu | Asp | 2 : 2 |
| site GCx: | Ala | Ala | Ala | Ala | 4 |
| site GGx: | Gly | Gly | Gly | Gly | 4 |
| site GTx: | Val | Val | Val | Val | 4 |
| site TAx: | !!! | Tyr | !!! | Tyr | 2 : 2 |
| site TCx: | Ser | Ser | Ser | Ser | 4 |
| site TGx: | !!! | Cys | Trp | Cys | 2 : 1 : 1 |
| site TTx: | Leu | Phe | Leu | Phe | 2 : 2 |

**Fig. 3.** Classification of sites in single coding regions. There are 48 quadruplets of codons differing by one substitution. Three of these (xGA, xGG, and ATx) fall outside the scheme that can be easily analyzed and will be treated as if they are more regular than they are. seven (xAA, xAG, xGA, TxA, TxG, TAx, and TGx) involved stop codons, that will be treated as if it was a 21st amino acid. Five (xGG, xTC, xTG, TxA and TGx) are 2-1-1 cases but can be regarded as 2-2 or 1-1-1-1 cases depending on the initial nucleotide. Two (xGA and xGG) are 2-1-1 cases, where the two synonymous codons differ by a transversion, that will be treated as a transition. As there is some overlap between these cases they only total to 12 cases.




One reading frame analysis
(1-1-1-1) sites: a' = f*a b' = f*b
(2-2) sites: a' = a b' = f*b
(4) sites: a' = a b' = b

Two reading frame analysis: Selection factors to be multiplied on the basic transition/transversion (a,b) coefficients if selection factors operate independently. Otherwise substitute fA*fB with fAB

First reading frame

| S e c o n d | (1-1-1-1) | (2-2) | (4) |
|---|---|---|---|
| (1-1-1-1) | fA*fB, fA*fB | fB, fA*fB | fB, fB |
| (2-2) | fA, fA*fB | 1, fA*fB | 1, fB |
| (4) | fA, fA | 1, fA | 1, 1 |

ex.: first reading frames (2-2), second reading frame (1-1-1-1)

a' = fB*a
b' = fA*fB*b

**Fig. 4.** Scheme for selection parallel with transition/transversion boundaries. $X(a',b')$, $Y(a',b')$, and $Z(a',b')$ are the probability for no change, a transition change, and a transversion change after time $t$, respectively. Change here means overall change; the nucleotide could have mutated many times in the time interval $[0\text{-}t]$, but this only registers the relationship between which nucleotide occupies the position at time 0 and time $t$. Thus, at $t = 0$ all probability mass will be in the upper left cell. Transition will cause a vertical movement, while transversions will move probability mass horizontally. A full transition probability matrix would need $4 \cdot 4$ entries, but in the simple Kimura model all rows are permutations of each other and each can be summarized in this $2 \cdot 2$ matrix. In the situation with one reading frame all amino-acid-changing substitutions will be assumed to be reduced with a factor, $f$, relative to the situation had they not changed the amino acid. Ignoring a few irregularities of the genetic code allows this $f$ to be incorporated directly into Kimura's model. If two reading frames, $A$ and $B$, are present, there will be two distinct selection factors, $f_A$ and $f_B$, that can be incorporated as well. If there were interaction between the joint effects of replacements in both reading frames, then $f_A \cdot f_B$ should be replaced with a $f_{AB}$. This is readily generalized to more reading frames.

be done in this article is to generalize Li et al.'s (1985) model to any combination of noncoding and overlapping reading frames occurring in genomic DNA, and maximum likelihood will be used to test hypotheses and to estimate the parameters.

## Basic Model

The assumptions are that all sites belong to type (4), (2-2), or (1-1-1-1), that transitions and transversions occur biochemically with rates $\alpha$ and $\beta$, and that there is time $t/2$ back to the most recent common ancestor of the two sequences. Thus the expected numbers of transitions and transversions per unselected site are $a = t \cdot \alpha$ and $b = t \cdot \beta$, respectively.

alpha-globin from rabbit and mouse.

| Sites | Total | Conserved | Transitions | Transversions |
|---|---|---|---|---|
| 1-1-1-1 | 274 | 246(.8978) | 12(.0438) | 16(.0584) |
| 2-2 | 77 | 51 (.6623) | 21(.2727) | 5(.0649) |
| 4 | 78 | 47 (.6026) | 16(.2051) | 15(.1923) |

Li (1993) Analysis:

| | Transitions | Transversions |
|---|---|---|
| 1-1-1-1 sites: | 0.0479 | 0.0621 |
| 2-2 sites: | 0.4365 | 0.0691 |
| 4 sites: | 0.3400 | 0.2428 |

Ks = 0.6307 Ka = 0.1116 Ka/Ks = 0.177

Maximum Likelihood Analysis:

a = 0.3003 b = 0.1871 2*b = 0.3742
(a + 2*b) = 0.6745 f = 0.1663

| | Transitions | Transversions |
|---|---|---|
| 1-1-1-1 | a*f = 0.0500 | 2*b*f = 0.0622 |
| 2-2 | a = 0.3004 | 2*b*f = 0.0622 |
| 4 | a = 0.3004 | 2*b = 0.3741 |

Expected number of replacement substitutions 35.49
synonymous 75.93

Replacement sites : 246 + (0.3742/0.6744)*77 = 314.72
Silent sites : 429 - 314.72 = 114.28

Ks = .6644 Ka = .1127

**Fig. 5.** The observed differences in the three types in the aligned rabbit and mouse alpha-globin are analyzed first by Li et al.'s (1985) method and then by maximum-likelihood analysis using the same underlying model.

It is not possible to separate the product of rate and time in these calculations, so only $a$ and $b$ are of interest in the sequel and there will be no reference to absolute time or rates. It is also assumed that any replacement substitution is accepted with a probability $f$ (selection factor) expected for purifying selection to lie between 0.0 and 1.0. This can readily be incorporated into Kimura's scheme by defining $a'$ and $b'$ for the three site types. In a 1-1-1-1 site, both transitions and transversions will change the amino acid and therefore we will have $a' = f \cdot a$ and $b' = f \cdot b$; in a 2-2 site only transversions will be amino acid changing, yielding $a' = a$ and $b' = f \cdot b$; and lastly, a 4 site will have no selection factor incorporated: $a' = a$ and $b' = b$ (Fig. 4).

The underlying model in this paper is identical to Li et al.s (1985), but is here generalized to cover overlapping reading frames. A second difference is in parameter estimation: We use maximum likelihood, while Li et al. used formulas derived by Kimura (1980). We opted for maximum likelihood, as it has good statistical properties (Edwards 1972), allows for the calculation and estimation of all interesting quantities, and was easy to generalize to more complex situations.

The parameters of the model are transition ($a$) and transversion ($b$) coefficients and the strength of selection ($f$). Let $X_t(a, b, f)$ be the probability that a position of type t is identical in the two sequences. Let $x_t$ be the observed number of identities in sites of type $t$. $Y_t$ and $y_t$ signify transition differences and $Z_t$ and $z_t$ transversion differences. It must be pointed out that this accounting is done using the first sequence to define the site types. The exact numbers could change slightly if the second sequence is used instead. The change is typically insignificant. We then have

$$X_{1-1-1-1}(a, b, f) = X(a \cdot f, b \cdot f)$$

$$X_{2-2}(a, b, f) = X(a \cdot f, b)$$

$$X_4(a, b, f) = X(a, b)$$

$Y_t(a, b, f)$ and $Z_t(a, b, f)$ are defined analogously.

The likelihood function will look as follows:

$$L(\text{observations}, a, b, f) =$$

$$C(t) \cdot \prod_t X_t(a, b, f)^{x_t} \cdot Y_t(a, b, f)^{y_t} \cdot Z_t(a, b, f)^{z_t}$$

Purifying selection will slow down all events in 1-1-1-1 sites, only transversions in 2-2 sites, and should have no effect in 4 sites. Properly quantified, this slowdown can be used to estimate the strength of selection and then to recover the underlying biochemical substitution rate.

The reasoning behind Li et al.'s (1985) and our method is illustrated in the simple case of rabbit and mouse alpha-globin (Fig. 5). There are 429 positions in the alignment of the two alpha-globin exons and they are classified into the three types. Within each type the aligned nucleotides can either be identical or differ by a transition or a transversion. Li et al. (1985) applies Kimura's (1980) correction formulae to estimate how many transitions and transversions have occurred in each position. It is readily seen that more events occur in 4 sites than in 1-1-1-1 sites. It is also seen that there is a much stronger transition/transversion bias in 2-2 sites than in other sites, which is to be expected.

To obtain rates of replacement per site, it is necessary to have a concept of replacement sites (analogously for synonymous events). In Li et al. all (1985) 1-1-1-1 sites and two-thirds of the 2-2 sites were replacement sites. The reason for the two-thirds is that of the three possible changes, two (the transversions) in 2-2 sites cause a replacement. This causes an overcounting of replacement because transversions occur more rarely than two-thirds of the time because of the transversion/transition bias. This led Li (1993) to partition 2-2 sites in replacement and synonymous sites so they reflected this bias. The Li (1993) method is here used to calculate replacement changes per replacement site, *Ka* and synonymous changes per synonymous site, *Ks*. It is seen that replacement changes occur 0.177 times less frequently than synonymous changes.

The likelihood for this dataset will be the product of four factors:

$$\begin{aligned} L(\text{observations}, a, b, f) = \\ C(429, 274, 77, 78) \cdot \{X(a \cdot f, b \cdot f)^{246} \\ \cdot Y(a \cdot f, b \cdot f)^{12} \cdot Z(a \cdot f, b \cdot f)^{16}\} \\ \cdot \{X(a, b \cdot f)^{51} \cdot Y(a, b \cdot f)^{21} \cdot Z(a, b \cdot f)^{5}\} \\ \cdot \{X(a, b)^{47} \cdot Y(a, b)^{16} \cdot Z(a, b)^{15}\} \end{aligned}$$

The first combinatorial factor, which is noninformative for the parameters, is the probability of the distribution of the three different kinds of site types along the 429-base-pair-long molecule.

The second factor is the probability of what is observed in the 1-1-1-1 sites. There are 246 identities, 12 transition differences, and 16 transversion differences. The $X(\ ,\ )$, $Y(\ ,\ )$, and $Z(\ ,\ )$ series has parameters $a \cdot f$ and $b \cdot f$ as both transitions and transversions causes amino acid changes and both these changes are dampened by factor $f$.

The third factor is analogous to the second factor, except now only transversions are dampened by factor $f$.

The fourth factor is analogous to the second and third factors, except now neither transitions nor transversion causes amino acid changes and there is no selection against these events and thus no $f$ in these terms. In the likelihood approach the three estimates are the parameter values that maximize the likelihood function. The kinds of changes in each type of site can be easily calculated. Quantities like *Ks* and *Ka* can also be calculated when sites have been partitioned into replacement and synonymous sites. Again 2-2 sites are partitioned to reflect the transition/transversion bias.

Applied to the alpha globins, the two methods give very similar results, which is reassuring, when the method is generalized to apply to overlapping reading frames.

This scheme is easily carried on to overlapping reading frames (Fig. 4) and will be illustrated in the case of two overlapping reading frames.

Sites:

| | | gag | | | |
|---|---|---|---|---|---|
| | | 1-1-1-1 | 2-2 | 4 | Total |
| pol | 1-1-1-1 | 64 | 31 | 34 | 129 |
| | 2-2 | 40 | 7 | 0 | 47 |
| | 4 | 27 | 2 | 0 | 29 |
| | Total | 131 | 40 | 34 | 205 |

Estimated parameters under assumption of common a and b for all regions and one selection factor per gene:

a = 0.084764 b = 0.024590 2*b = 0.04918 (a + 2*b) = 0.13394

$f_{gag}$ = 0.403 $f_{pol}$ = 0.229

Estimated events in presence of both genes:

| | | gag | | |
|---|---|---|---|---|
| | | synonymous | replacement | Total |
| pol | synonymous | 0.762 | 3.22 | 3.982 |
| | replacement | 1.644 | 1.08 | 2.652 |
| | Total | 2.408 | 4.30 | 6.708 |

Estimated events if the genes were nonoverlapping:

| | synonymous | replacement | Total |
|---|---|---|---|
| gag | 8.31 | 7.85 | 16.16 |
| pol | 7.79 | 4.48 | 12.27 |

**Fig. 6.** Analysis of the *gag-pol* region under the assumption that the two overlap (which they do), and under the assumption that they are single.

Now, there will be a selection factor for each protein, $f_A$ and $f_B$, and one for both $f_{A,B}$, and each site will have to be classified according to its kind in each reading frame. If selection works independently on the two genes $f_{A,B} = f_A \cdot f_B$.

To illustrate the effect of overlapping genes relative to not overlapping genes, how many synonymous and nonsynonymous substitutions are expected to have occurred in the overlap between *gag* and *pol* are calculated. The parameters are estimated under the assumptions that transition and transversion coefficients were the same for all regions. If there were no overlap, the expected number of synonymous substitutions in the *gag* region would be calculated as follows: In the 4 sites (34) it would be the expected number of (transitions + transversion) per site multiplied by the number of sites, i.e., $34 \cdot 0.1339 = 4.554$. In the 2-2 sites (40) it would be the expected number of transitions (only) per site multiplied by the number of sites, i.e., $40 \cdot 0.0848 = 3.391$. A similar calculation can be made for replacement substitutions, except in this case selection only allows a fraction (0.403) of the substitutions to occur.

When overlap is assumed, calculations can be illustrated as follows: The expected number of substitutions that are synonymous in the *gag* gene and nonsynonymous in the *pol* gene would be 1.644. They could only occur in the 2-2,1-1-1-1 (31), 2-2,2-2 (7), 4,1-1-1-1 (34), or the 2-2,4 (0) sites. The expected number occurring in the 2-2,1-1-1-1 sites would be: (the number of sites) times (the number of transitions per site) times (the fraction of replacement substitutions expected to survive purifying selection against the *gag* gene) = $31 \cdot 0.085 \cdot 0.229 = 0.602$. The calculations can be carried out for all classes of genes and the results are shown at the bottom in Fig. 6.

The addition of *gag* reading frame within the *pol* reading frame diminishes the expected number of both replacement and synonymous substitutions. They are lowered from 7.79 and 4.48 to 3.98 and 2.65, respectively. This is not surprising, as both synonymous and nonsyn-

onymous substitutions in the *pol* gene will be selected against, as they can be nonsynonymous in the *pol* gene. This should be taken into account when calculating rates and selection factors. The reverse the situation: To add the *pol* gene to a preexisting *gag* gene will have an even larger effect, as the selection against *pol* is stronger than against *gag*.

## A Hierarchy of Models

An interval of aligned nucleotide pairs with the same configuration of reading frames along the aligned DNA is called a *region.* In Fig. 1, there are 24 regions. The first goes from 1 to 336 and doesn't code for anything in either sequence. The second region goes from 336 to 1,633 and codes for *gag* in both sequences; then follows a double coding region coding for both *gag* and *pol.* A region with $k$ nonhomologous coding regions could have $3^k$ different types of sites if all combinations of types were possible from the single coding case. Some types cannot be realized with the genetic code, e.g., a site cannot be fourfold degenerate in two reading frames in the same direction.

Each site will be occupied by a nucleotide pair, where the first will be interpreted as the initial state of the Markov chain and the second nucleotide as the resultant state.

In the two aligned sequences there are 4 noncoding regions, 10 singly coding regions, 8 doubly coding, and 2 triply coding regions.

The model lumps the two transversions together, so all that is observed at each position is whether the nucleotides are identical, differ by a transition, or differ by a transversion. The likelihood function for these observations is

$$L(\text{observations}, p) =$$

$$C(r,t) \cdot \prod_r \prod_t X_{r,t}(p)^{x_{r,t}} \cdot Y_{r,t}(p)^{y_{r,t}} \cdot Z_{r,t}(p)^{z_{r,t}}$$

where $X_{r,t}(p)$, $Y_{r,t}(p)$, and $Z_{r,t}(p)$ (Fig. 2) are the probabilities of identity, of differing by a transition, or of differing by a transversion, respectively, for a position of type t in region r. The factors are multiplied over all $r$ (regions) and for each region all $t$ (types of sites) and $p$ includes all parameters. The parameter $p$ includes $a$, $b$, and the possible $f$'s for that site. $x_{r,t}$ is the number of matched nucleotides with identical nucleotides in region $r$ and in a site of type $t$. Analogously, $y_{r,t}$ is the number of transitions and $z_{r,t}$ is the number of transversion differences. $C(r,t)$ is the multinomial coefficient, which is uninformative about the parameters. The parameters can now be estimated by maximum likelihood, giving estimates together with their standard deviation obtained from the curvature of the likelihood function around the maximum (Edwards 1972). As hypotheses using less and less parameters are accepted as descriptions of the data, the standard deviations will fall. As a measure of distance the total amount of substitution, $a + 2b$, is given. The maximization procedure used was Powell's, as described in Press et al. (1992).

### *1. The Full Model*

Each region has its own $a$ and $b$, and a substitution causing replacement in genes $i$, $j$, and $k$ will be selected on with a factor $f_{i,j,k}$. If $n$ nonhomologous genes were present in the region, it would be $2^n$-1 selection intensities. The observations for the first three regions are shown in Fig. 7. The first region is noncoding and there is only one type of site of which there are 335. Among these there are 324 aligned nucleotide pairs that are identical, 7 that differ by a transition, and 4 that differ by a transversion. There are two parameters, $a$ and $b$, to explain three observations that must sum to 335 and this allows a perfect fit. The Markov process used cannot give probabilities of identity less than 0.25 and in such cases the data cannot be fitted perfectly, but in less-diverged cases this will not occur.

In the second region there are three types of sites: All substitutions change the amino acid, i.e., nondegenerate (1-1-1-1), and there are 832 of these. There are 288 aligned nucleotide pairs (sites) where a transition will not change the amino acid but a transversion will (twofold degenerate [2-2]). And lastly there are 169 positions where a substitution will not change the amino acid (fourfold degenerate [4]).

With the first type of sites there are 802 identical positions, 22 that differ by a transition, and 14 that differ by a transversion. The numbers in parenthesis are the expected number in the stochastic model if the maximum likelihood were used. It is seen that the largest contribution to chi-square comes from too-few conserved sites in the fourfold degenerate sites.

There are two independent observations for each type of site, six in total, and there are three parameters in the Markov process used to describe the evolution of the sequence. This gives 3 degrees of freedom and the observed −2logQ is 6.32, corresponding to a test probability of 0.09.

The maximum likelihood estimates of the parameters: The number of transitions per site, $a$, is 0.088 with a standard deviation of 0.012; the number of a specific transversion per site, $b$, is 0.021 with a standard deviation of 0.004. So transitions occur approximately four times as often as transversions. The fraction of replacement substitutions accepted by selection is estimated to be 0.367 with a standard deviation of 0.057. The total number of substitutions per site, $(a + 2b)$, is also shown and is about 0.13.

In the region where *gag* and *pol* overlap there are only seven types observed. The double completely degenerate is impossible and the combination of a (2-2) site in the

```
Region 1: Noncoding

Sites          Conserved     Transitions          Transversions
335              324           7                     4

Estimates:
a=    0.0216  st.dv.=0.0083
b=    0.0060  st.dv.=0.0030
a+2b=0.0337
-log likelihood=55.6063


Region 2: gag

Site type      total(1389)        conserved      transitions      transversions

1-1-1-1        832              802 (799.35)     22 (25.88)       14 (12.77)
2-2            288              253 (260.71)     28 (22.90)        7 (4.39)
4              169              154 (149.03)     12 (13.15)        3 (6.83)

Estimates:
f=    0.3674  st.dv.=0.0567
a=    0.0880  st.dv.=0.0116
b=    0.0211  st.dv.=0.0043
a+2b=0.1337
-log likelihood=357.9564
-2lnQ=6.3294
3 degrees of freedom
test probability=0.0966


Region 3: gag (1) and pol (2)

Site type           total      conserved     transitions    transversions

1-1-1-1,1-1-1-1      84        82 (61.62)      2 (1.88)       0 (0.50)
1-1-1-1,2-2          31        27 (27.74)      3 (3.02)       1 (0.24)
1-1-1-1,4            34        30 (29.83)      2 (3.25)       2 (0.92)
2-2,1-1-1-1          40        35 (34.95)      5 (4.74)       0 (0.31)
2-2,2-2               7         6 (6.17)       1 (0.77)       0 (0.05)
4,1-1-1-1            27        23 (22.97)      4 (3.12)       0 (0.91)
4,2-2                 2         2 (1.72)       0 (0.22)       0 (0.07)

Estimates:
f1=   0.8685  st.dv.=0.3462
f2=   1.0867  st.dv.=0.3877
f12=  0.2434  st.dv.=0.1421
a=    0.1256  st.dv.=0.0325
b=    0.0164  st.dv.=0.0094
a+2b=0.1577
-log likelihood=71.0838
-2lnQ=7.3166
9 degrees of freedom
test probability=0.6042

Summed loglikelihood for all 24 regions = -3057.4976
Total -2lnQ = 130.2885
108 degrees of freedom
Test probability: 0.1271
```

**Fig. 7.** Calculations for the first three regions, when each region has a transition and transversion coefficient and without the assumption that selection factors work independently. (*Hypothesis: a* and *b* specific for each region. One specific selection factor, *f*, for each combination of genes in each region.)

first gene with a completely degenerate site is not observed. This gives a total of five parameters: three selection parameters, $f_{gag}$, $f_{pol}$, and $f_{gag,pol}$, and $a$ and $b$. There are $7 \cdot 2 = 14$ independent cells of observations giving a test with 9 degrees of freedom. The $f_{gag}$ is estimated to 0.869, $f_{pol}$ to 1.087, and $f_{gag,pol}$ to 0.24. Their standard deviations are quite large.

The situation for region 16 with three overlapping regions is analogous. Of the $3^3 = 27$ combinations of sites for three genes, only ten were observed. This gives 20 independent cells and takes nine parameters (seven $f$'s and $a$ and $b$) and a goodness of fit with 11 degrees of freedom.

For all regions there are 96 parameters, as there 24 regions, each with a transition and a transversion rate (i.e., 48), 10 singly coding regions, each with one selection parameter (i.e., 10), 8 doubly coding regions, each with 3 selection factors (i.e., 24), and, lastly, 2 triply coding regions, each with 7 selection parameters (i.e., 14). There are 204 cells that must be explained with these parameters, giving 108 degrees of freedom. For each noncoding region there are three observations: identities, transition differences, and transversion differences. For single coding regions there are three types, each having three observations. For doubly coding there are less than nine types depending on the region. In the *gag-pol* region, there are seven observed types of sites, each with three observations. In triply coding regions, there are less

```
Region 3: gag (1) and pol (2)

Site type      total(225)   conserved      transitions     transversions

1-1-1-1,1-1-1-1  84   82(61.62)(60.89)  2(1.88)(2.48)    0(0.50)(0.63)

1-1-1-1,2-2      31   27(27.74)(28.05)  3(3.02)(2.64)    1(0.24)(0.30)

1-1-1-1,4        34   30(29.83)(30.37)  2(3.25)(2.86)    2(0.92)(0.77)

2-2,1-1-1-1      40   35(34.95)(35.32)  5(4.74)(4.28)    0(0.31)(0.39)

2-2,2-2           7    6( 6.17) (5.43)  1(0.77)(1.50)    0(0.05)(0.07)

4,1-1-1-1        27   23(22.97)(23.38)  4(3.12)(2.84)    0(0.91)(0.78)

4,2-2             2    2( 1.72)( 1.52)  0(0.22)(0.42)    0(0.07)(0.06)

Estimates:
Without independence                With  independence
f1=  0.8685  st.dv.=0.3462        f1=  0.3336  st.dv.=0.0735
f2=  1.0867  st.dv.=0.3877        f2=  0.4306  st.dv.=0.1300
f12= 0.2434  st.dv.=0.1421        f12= f1*f2 = 0.1436
a=   0.1256  st.dv.=0.0325        a=   0.2829  st.dv.=0.0735
b=   0.0164  st.dv.=0.0094        b=   0.0346  st.dv.=0.0202
a+2b=0.1577                       a+2b=0.3175
-log likelihood=71.0838            -log likelihood=71.6646
                                  -2lnQ=0.5808

                                        1 degrees of freedom
                                       test probability=0.72

                Summed log likelihood: -135.9318
                          -2lnQ = 5.65
                      16 degrees of freedom
                    test probability: 0.9915
```

**Fig. 8.** In the region where *gag* overlaps *pol*, it is now tested whether $f_{gag} \cdot f_{pol} = f_{gag,pol}$ which gives a reduction of one parameter, a −2logQ of 1.162, is found and independence is accepted with a test probability of 0.2811. Both $f_{gag}$ and $f_{pol}$ are then estimated considerably lower, to 0.333 and 0.431, respectively. The additional parenthesis relative to Fig. 7 denotes how many was expected if the hypothesis of independence were true. So there are 84 double nondegenerate sites. Of these 82 were conserved. Under the model without independence there would be expected to be 61.62 and if independence were assumed 60.89 would be expected.

than 27 types. In the *env-rev3-tat3* region ten types were observed, each with three observations. The total −2logQ is 130.28, corresponding to a test probability of 0.12 and an overall fit for the model.

### 2. *Independent Selection Factors*

The selection factors are independent for each gene so $f_{ijk}$ can be written as $f_i \cdot f_j \cdot f_k$ and a region with $n$ genes would need $n$ selection intensities. The test for the region coding for *gag* and *pol* is illustrated in Fig. 8. For all regions this would sum to 80 parameters, i.e., a reduction of 16 parameters relative to the model without independence of selection factors. The total −2logQ is 5.65, corresponding to a test probability of 0.9915 and an overall accept of independence.

### 3. *One Gene, One Selection Factor*

Each gene now has its own selection intensity, which must be the same for different regions. Since there are nine genes there will be nine selection intensities. For all regions this would sum to 57 parameters, i.e., a reduction of 22 parameters relative to the model without independence of selection factors. The total −2logQ is 29.71, corresponding to a test probability of 0.1257 and an overall accept of describing each gene with one selection factor. The selection factors for the nine genes are shown in Fig. 9.

### 4. *Constant* a/b *Ratio Along the Sequence*

The parameters $a$ and $b$ are usually interpreted as the basic biochemical substitution rates and can reasonably be expected to be the same for different regions. However, often the sequences have been subject to recombination and gene conversion; the most recent common ancestor might be found at different times for different regions. This could lead to different sizes of $a$ and $b$ for different regions, but it should not influence their ratio, so the hypothesis that these ratios were constant was also tested. This corresponds to a reduction of 23 parameters relative to the hypothesis above, giving 34 independent parameters. The total −2logQ is 47.14, leading to a rejection of this hypothesis with test probability 0.0021. The *a/b* ratio showed no easily interpretable deviations but was highly variable from region to region.

### 5. *Constant* a *and* b *Along the Sequence*

Assuming that $a$ and $b$ are constant for all regions, would have 11 free parameters. The total −2logQ is 92.11 with 23 degrees of freedom, giving a strong rejection. This hypothesis was tested, although a weaker hypothesis had already been rejected, because an overall measure of dis-

```
GAG:   0.339 (st.dv. 0.048)
POL:   0.226 (st.dv. 0.028)
VIF:   0.349 (st.dv. 0.077)
VPR:   0.431 (st.dv. 0.122)
TAT:   1.261 (st.dv. 0.249)
REV:   0.394 (st.dv. 0.093)
VPU:   0.576 (st.dv. 0.117)
ENV:   0.604 (st.dv. 0.041)
NEF:   0.723 (st.dv. 0.101)
```

**Fig. 9.** The selection factors for different genes. It is seen that longer genes (*gag, pol, env*) have their selection factors better determined than shorter genes (that is, they have smaller standard deviations). It is also seen that the *pol* gene is the slowest evolving. *Tat* has an $f$ higher than 1.0, but it is not significantly higher than 1.0. If it was, it would have implied positive selection and that replacement substitutions were favored over synonymous substitutions. The large envelope gene (*env*) evolves considerably faster than the other two large genes (*pol, gag*).

tance ($a + 2b$) would be very convenient. For these two retroviruses this says that the divergence is slightly above 0.1339 with a standard deviation of 0.0066.

## Additional Tests

If a region matches nonhomologous reading frames (Fig. 1), two parameters will influence the pattern of evolution: how long a time the reading frame has been present in the region and how strong the selection has been against replacement substitutions. The simplest possible situation is when there is one reading frame in the first sequence and none in the second. This situation illustrates the central problem: If the selection against replacements isn't known, then it is difficult to distinguish between a gene that has been there for a long time, but experienced weak selection, from a gene that has been there for a short time, but experienced strong selection. If the strength of selection is known, it is possible to estimate under which fraction of the evolution experienced, the reading frame has been present, and furthermore, if a molecular clock can be assumed, it can then be determined if the reading frame has been inserted or deleted.

Another question that could also be addressed by this method concerns the dating of frameshift mutations. A frameshift changes the rates of evolutions between different reading frames, and this could be rephrased to estimate the occurrence of the frameshift. Most likely no other type of mutation has this property. However, the situation can become very complex if several frameshifts have occurred and several reading frames are involved, and the question will not be treated here.

## Conclusion

A method was presented that models the evolution of genomic DNA and describes its evolution by a set of rates. A hierarchy of hypothesis allows tests of assumptions about the evolution of genomic DNA. The main advantage of the method is its generality. It can handle any combination of noncoding, singly coding, and multiply coding regions and answer questions about equality of selection coefficients and substitution rates. It should be useful to give a distance measure between genomic structures and to rank the degree of selection on different genes.

Its main weakness is shared by most methods that analyze coding regions—it handles the irregular restrictions imposed by the genetic code unsatisfactorily; it ignores codon bias; and, lastly, it treats each sequence or region as a homogenous string, ignoring the uneven distribution of functional constraint along the molecule. A proper statistical model would have to take these factors into account.

*Acknowledgments.* Dr. Hans Siegismund and Anne-Mette Krabbe Pedersen are thanked for critical discussions of the manuscript. J.H. was supported by the Carlsberg Foundation and the Danish Research Council, grants 11-8916-1 and 11-9639-1. J.S. was supported by the Danish Research Council, grants 11-8916-1 and 11-9639-1, and the Carlsberg Foundation.

## References

Edwards AWF (1972) Likelihood. Cambridge University Press, Cambridge

Hein JJ, Støvlbæk J (1994) Genomic alignment. J Mol Evol 38:310–316

Jukes TH, Cantor C (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

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

Li W-S, Wu C-I, Luo C-C (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 2:2:150–174

Li W-S (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Myiata T, Yasanunga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from nucleotide sequences and its applications. J Mol Evol 16:23–36

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–26

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C. Cambridge University Press, Cambridge