A Combined Empirical and Mechanistic Codon Model

Adi Doron-Faigenboim and Tal Pupko

Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, 69978, Israel

The evolutionary selection forces acting on a protein are commonly inferred using evolutionary codon models by contrasting the rate of synonymous to nonsynonymous substitutions. Most widely used models are based on theoretical assumptions and ignore the empirical observation that distinct amino acids differ in their replacement rates. In this paper, we develop a general method that allows assimilation of empirical amino acid replacement probabilities into a codon-substitution matrix. In this way, the resulting codon model takes into account not only the transition–transversion bias and the nonsynonymous/synonymous ratio, but also the different amino acid replacement probabilities as specified in empirical amino acid matrices. Different empirical amino acid replacement matrices, such as secondary structure–specific matrices or organelle-specific matrices (e.g., mitochondria and chloroplasts), can be incorporated into the model, making it context dependent. Using a diverse set of coding DNA sequences, we show that the novel model better fits biological data as compared with either mechanistic or empirical codon models. Using the suggested model, we further analyze human immunodeficiency virus type 1 protease sequences obtained from drug-treated patients and reveal positive selection in sites that are known to confer drug resistance to the virus.

Introduction

A multiple sequence alignment combined with the underlying phylogenetic tree and a model of sequence evolution allows inference of the evolutionary selection forces acting on a protein. While amino acid evolutionary models are restricted to computing the purifying selection acting on each site (e.g., Gaucher et al. 2002; Pupko et al. 2002; Susko et al. 2002; Mayrose et al. 2004), codon evolutionary models can be used to compute both purifying and positive Darwinian selection (Nielsen and Yang 1998; Yang and Nielsen 2002; Yang and Swanson 2002; Massingham and Goldman 2005). This is usually done by contrasting the rate of neutral evolution as estimated by synonymous (silent) substitutions to the rate of nonsynonymous (amino acid altering) substitutions. Formally, the ratio of the nonsynonymous substitutions rate (*K*a) to the synonymous substitutions rate (*K*s) is estimated. Sites showing *K*a/*K*s values significantly lower than 1 are regarded as undergoing purifying selection and therefore may have a functionally or structurally important role. Sites showing *K*a/*K*s values significantly higher than 1 are indicative of positive Darwinian selection, suggesting adaptive evolution (e.g., Sharp 1997; Akashi 1999; Hurst 2002).

Methods used for inferring *K*a/*K*s ratios are constantly being developed (Li et al. 1985; Wong et al. 2004; Tang and Wu 2006). Widely used models take into account factors such as different probabilities for transitions and transversions, codon bias, and among-site rate variation (Goldman and Yang 1994; Muse and Gaut 1994; Nielsen and Yang 1998; Yang et al. 2000). In addition, although previous methods inferred a global *K*a/*K*s value for the entire sequence or for subsequences using a sliding window approach (Fares et al. 2002; Berglund et al. 2005), recent methods estimate the *K*a/*K*s ratio per amino acid site (Yang 2002; Suzuki 2004b). This enables the detection of single sites that undergo positive selection despite a low global *K*a/*K*s value for the entire protein.

Key words: evolutionary models, positive selection, purifying selection, empirical amino acid replacement matrices, Bayesian inference, *K*a/*K*s, codon models.

E-mail: talp@post.tau.ac.il.

Mol. Biol. Evol. 24(2):388–397. 2007
doi:10.1093/molbev/msl175
Advance Access publication November 16, 2006

Goldman and Yang (1994) and Muse and Gaut (1994) developed codon-based evolutionary models for inferring the *K*a/*K*s ratio. These models are mechanistic, that is, they include parameters for the transition–transversion bias, the codon frequencies, and, in the case of the Goldman and Yang (1994) model, also the different replacement probabilities between amino acids based on the Grantham (1974) physicochemical distance matrix. Nielsen and Yang (1998) and Yang et al. (2000) further developed mechanistic Bayesian models that assume a prior distribution of *K*a/*K*s ratios. However, unlike the model of Goldman and Yang (1994), these models ignore the fact that distinct amino acids differ in their replacement rates. Thus, these models will give the same probability for a tryptophan codon changing into a leucine codon (UGG → UUG), as for a phenylalanine codon changing into a leucine codon (UUU → UUG) because they both require 1 transversion. However, according to empirically derived amino acid–based replacement matrices, such as the JTT matrix (Jones et al. 1992), the latter event should be about 5 times more likely than the former. Recently, Sainudiin et al. (2005) and Wong et al. (2006) have developed mechanistic Bayesian codon models in which amino acid physicochemical properties are explicitly taken into account. In these models, codons are partitioned according to a predefined physicochemical property, such as polarity or charge. The difference in this property between 2 codons dictates their substitution probability. In this way, specific physicochemical selective pressures acting on a protein can be modeled. This approach, however, is limited as it only allows a single partition per model.

Empirical amino acid replacement matrices (e.g., Dayhoff et al. 1978; Jones et al. 1992; Adachi and Hasegawa 1996; Whelan and Goldman 2001) are extensively used in various kinds of protein sequence analyses, such as multiple sequence alignment tools (Thompson et al. 1994), homologous searches (Altschul et al. 1997), and phylogeny reconstruction (Kumar et al. 2004). In such empirical matrices, the parameters of the replacement probabilities are estimated from large data sets of protein sequences and then assumed to be fixed when these matrices are applied to a specific protein. However, when one considers codon-based data, estimating empirical codon matrices requires a substantial amount of data and accurate inference of thousands of parameters (because they involve an

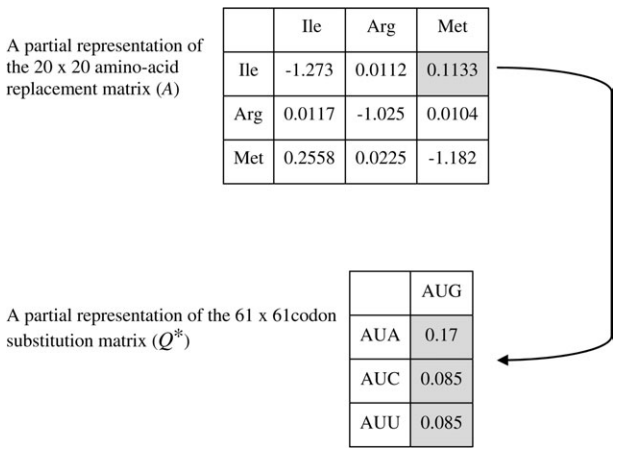

FIG. 1.—A graphic representation of the "mechanistic–empirical combined" model. A 20 × 20–amino acid matrix is expanded to a 61 × 61 codon matrix. Each replacement between a pair of amino acids (marked in gray) from matrix *A* is converted to its corresponding codon-substitution rates (marked in gray).

alphabet size of 61). Recently, Schneider et al. (2005) estimated an empirical codon-substitution matrix from a large number of coding data sets as an alternative to parameterized models. The empirical matrix is constructed from 5 metazoan genomes and thus most accurately describes the evolution of these species. There is trade-off between using parameter-rich models, which better fit the data under study, but can risk overfitting, and nonparameterized models, which have no free parameters but can risk underfitting. The empirical matrix is a conservative approach of the latter type, which assumes 1 general matrix for all genes. This implies 1 *Ka*/*Ks* ratio as well as the same transition–transversion bias for all genes.

Here we suggest the construction of a combined codon model that, on the one hand, assimilates empirical amino acid replacement probabilities and, on the other hand, takes into account theoretical assumptions (such as the transition–transversion bias, different codon frequencies, and different selection forces acting within and among genes). Different empirical amino acid replacement matrices have been developed. For instance, mitochondrial proteins contain a high proportion of hydrophobic residues (Naylor et al. 1995) and are mostly membranous (Adachi and Hasegawa 1996). Thus, the use of a specific replacement matrix, such as the vertebrate mitochondrial (mtREV) matrix (Adachi and Hasegawa 1996) has been shown to better describe mitochondrial proteins than the more general replacement matrix JTT (Jones et al. 1992). Other context-dependent empirical amino acid probability matrices were developed, for example, secondary structures (alpha helices, beta sheets, and loops) (Koshi and Goldstein 1995) and for transmembrane and nontransmembrane domains (Jones et al. 1994). These matrices as well as other context-dependent empirical amino acid models can be integrated into the combined codon models to create "context-dependent codon models." Such models are more realistic and may have a substantial effect on the accuracy of the *Ka*/*Ks* estimates and on phylogeny.

The JTT (Jones et al. 1992), the mtREV (Adachi and Hasegawa 1996), and the chloroplast (cpREV) (Adachi et al. 2000) empirical amino acid matrices were used to construct 3 such codon-based models. Twenty-seven nuclear, viral, mitochondrial, and chloroplast genes were analyzed to evaluate which of the models (mechanistic, empirical, or our mechanistic–empirical combined model) better fits protein data sets. We show that the suggested combined model is superior to the classical mechanistic model as well as to the empirical model for the vast majority of genes analyzed. Finally, we used our model to analyze sequences of human immunodeficiency virus type 1 (HIV-1) protease obtained from drug-treated patients and to infer the selection forces acting on each codon site in the protein. This analysis revealed specific sites as undergoing positive selection; most of these sites were previously shown to confer drug resistance to the virus.

Theory

Model of Codon Substitution Assuming Selection

Similar to most evolutionary models, the codon model used here assumes a stochastic Markovian process. The states of the model are the 61 sense codons (in the case of the universal genetic code). In this model, we expand a 20×20 empirical amino acid matrix into a 61×61 codon matrix. Each replacement between 2 amino acids, aa_i and aa_j, in the empirical matrix is represented by $c_i \times c_j$ elements in the new codon matrix, where c_i and c_j are the number of codons coding for aa_i and aa_j, respectively (see fig. 1). However, as is evident from figure 1, each substitution should be weighted differently based on 2 theoretical parameters (transition–transversion bias and codon frequencies), as we will now formulate.

Let *A* represent the empirical amino acid replacement matrix and *Q** the derived codon-based matrix. The basic assumption is that the infinitesimal replacement rate between 2 amino acids is the sum of such rates between all the codons coding for these 2 amino acids, weighted by the relative frequencies of those amino acids and codons. This assumption was previously pointed out by Yang et al. (1998) and is represented in the following equation:

$$\psi_i \cdot A_{ij} = \sum_{\{l:\mathrm{aa}_l = i\}} \sum_{\{s:\mathrm{aa}_s = j\}} \pi_l \cdot Q^*_{ls}, \qquad (1)$$

where ψ_i and π_l denote the frequency of amino acid *i* and codon *l*, respectively. *Aij* represents the substitution rate from amino acid *i* to *j*, derived from the empirical amino acid replacement matrix *A*. aa_l and aa_s denote the amino acids coded by codon *l* and *s*, respectively, and Q^*_{ls} represents the substitution rate from codon *l* to *s*. Q^*_{ls} is given by the following equation (for $l \neq s$):

$$Q^*_{ls} = \begin{cases} t_r \cdot \pi_s \cdot x_{(\mathrm{aa}_l,\mathrm{aa}_s)} & l \text{ and } s \text{ differ by 1 transition} \\ t_v \cdot \pi_s \cdot x_{(\mathrm{aa}_l,\mathrm{aa}_s)} & l \text{ and } s \text{ differ by 1 transversion} \\ t_{\pi\pi} \cdot \pi_s \cdot x_{(\mathrm{aa}_l,\mathrm{aa}_s)} & l \text{ and } s \text{ differ by 2 transitions} \\ t_{vv} \cdot \pi_s \cdot x_{(\mathrm{aa}_l,\mathrm{aa}_s)} & l \text{ and } s \text{ differ by 2 transversions} \\ t_{rv} \cdot \pi_s \cdot x_{(\mathrm{aa}_l,\mathrm{aa}_s)} & l \text{ and } s \text{ differ by 1 transition and 1 transversion} \\ t_{\mathrm{sub}} \cdot \pi_s \cdot x_{(\mathrm{aa}_l,\mathrm{aa}_s)} & l \text{ and } s \text{ differ by 3 substitutions} \end{cases}, \qquad (2)$$

where t_r and t_v are weight parameters representing 1 transition and 1 transversion, respectively. $t_{\pi\pi}$, t_{vv}, and t_{rv} are weight parameters representing 2 transitions, 2 transversions, and 1 transition and 1 transversion, respectively. t_{sub}

is a weight parameter representing 3 substitutions of any kind. The codon frequencies π_s are calculated here using the products of the observed nucleotide frequencies at each of the 3 codon positions (Yang et al. 2000) (denoted as F3 × 4 method). $x_{(\mathrm{aa}_l,\mathrm{aa}_s)}$ is a specific factor in the codon matrix that is used to differentiate between the substitution rates among codons coding for different amino acids. These factors are not free parameters of the model. Rather, each pair of amino acids defines one such x factor and thus one specific case of equation (1). Thus, these factors represent the empirical amino acid replacement probabilities that are integrated into the codon model. Once the transition–transversion parameters, codon and amino acid frequencies, and Aij are known, each $x_{(\mathrm{aa}_l,\mathrm{aa}_s)}$ is determined by solving equation (1). For synonymous substitutions (i.e., $\mathrm{aa}_l = \mathrm{aa}_s$), $x_{(\mathrm{aa}_l,\mathrm{aa}_s)}$ is arbitrarily set to one.

To account for different selection strengths (Ka/Ks), the nonsynonymous substitutions in Q^* are multiplied by a factor ω, which determines the intensity of selection. The resulting matrix Q' represents the infinitesimal codon-substitution rate and is given by the following equations.

For $l \neq$ s:

$$Q'_{ls} = \begin{cases} \omega \cdot Q^*_{ls} \text{ for a nonsynonymous substitution} \\ Q^*_{ls} \text{ for a synonymous substitution} \end{cases}, \quad (3)$$

and for the diagonal elements:

$$Q'_{ll} = -\sum_{\{s:s\neq l\}} Q'_{ls}. \quad (4)$$

Selection + Neutral Model

Because the model described above is based on an empirical amino acid replacement model, different pairs of codons obtain different replacement probabilities depending on the amino acids they code for. As a result, the model implicitly assumes selection. In other words, in a neutral model all the x factors in equation (2) should be equal for all pairs of amino acids. This cannot be reached regardless of the value of ω. In order to create a model that allows also neutral evolution, we combine the above "selection model" with a model that assumes no selection, that is, a model in which all the x factors are set to one. Under the latter model, the substitution is specified by:

$$Q^0_{ls} = \begin{cases} t_r \cdot \pi_s & l \text{ and } s \text{ differ by 1 transition} \\ t_v \cdot \pi_s & l \text{ and } s \text{ differ by 1 transversion} \\ t_{\mathrm{rr}} \cdot \pi_s & l \text{ and } s \text{ differ by 2 transition} \\ t_{\mathrm{vv}} \cdot \pi_s & l \text{ and } s \text{ differ by 2 transversion} \\ t_{\mathrm{rv}} \cdot \pi_s & l \text{ and } s \text{ differ by 1 transition and 1 transversion} \\ t_{\mathrm{sub}} \cdot \pi_s & l \text{ and } s \text{ differ by more than 2 substitutions} \end{cases}, \quad (5)$$

where t_r, t_v, t_{rr}, t_{vv}, t_{rv}, t_{sub}, and π_s are the parameters as before.

This selection + neutral model assumes a probability f for the selection matrix (Q') and a probability $1-f$ for the neutral matrix (Q^0). Thus, the instantaneous rate matrix for the process (Q) is $f \cdot Q' + (1-f) \cdot Q^0$. We assume that the parameters t_r, t_v, t_{rr}, t_{vv}, t_{rv}, t_{sub}, and π_s are the same in Q^0 and Q'. We hereby refer to this novel model as the "MEC" model, which stands for mechanistic–empirical combination, as opposed to the mechanistic model described by Nielsen and Yang (1998), which we denote as the M model. We refer to the empirical model (Schneider et al. 2005) as the E model.

Empirical Bayesian Estimation of Ka/Ks

Common to other mechanistic models, the free parameters of the model are estimated from the data being analyzed. Here, the evolutionary times (branch lengths), the transition–transversion parameters (t_r, t_v, t_{rr}, t_{rv}, t_{vv}, and t_{sub}), and the parameter f are assumed to be identical over all sites and are estimated using the maximum likelihood (ML) paradigm. The parameter ω is assumed to vary among sites, and thus a prior statistical distribution accounting for heterogeneous ω values among sites is used. The parameters of this distribution are also estimated using the ML methodology. We note that any modification of a free parameter necessitates recomputing the x factors so that equation (1) holds.

Here, 2 different prior distributions over ω are assumed, either a gamma distribution or a beta + ω distribution (which assumes that the ω values for a proportion p_0 of the sites is distributed beta[p,q], whereas the remaining proportion $1-p_0$ is assigned an ω value higher than 1). These distributions, known as the M5 and M8 models, respectively, were suggested by Yang et al. (2000). After all the parameters are estimated, an empirical Bayesian approach can be used to infer ω for each site. Here, $K = 10$ discrete categories are used to approximate the continuous gamma or beta distributions, where all categories have equal prior probabilities $1/K$. The posterior probability that a specific site is from category k is

$$P(\omega_k|\mathrm{data},T,\theta) \cong \frac{P(\mathrm{data}|\omega_k,T,\theta)\cdot P(\omega_k)}{\sum_{i=1}^{K} P(\mathrm{data}|\omega_i,T,\theta)\cdot P(\omega_i)}, \quad (6)$$

where T denotes the tree topology and branch lengths, θ denotes all the free parameters (t_r, t_v, t_{rr}, t_{rv}, t_{vv}, t_{sub}, f, and the parameters of the prior ω distribution). $P(\mathrm{data}|\omega_i,T,\theta)$ is calculated from the phylogenetic tree and branch lengths using Felsenstein's pruning algorithm (Felsenstein 1981), and $P(\omega_i)$is the prior probability of ω_i.

The Ka/Ks ratio is a function of ω. Once ω is specified, Q is defined and the Ka/Ks ratio can be calculated as described by Goldman and Yang (1994). Ka is calculated by summing $\pi_l Q_{ls}$ over all codon pairs, l and s, coding for different amino acids, and dividing by the same summation under the neutral model. Similarly, Ks is calculated by summing $\pi_l Q_{ls}$ over all codon pairs, l and s ($l \neq s$), coding for the same amino acid, and dividing by the same summation under the neutral model. This allows the calculation of Ka/Ks for each discrete category (ω_i), which we denote by (Ka/Ks)(ω_i). The estimated Ka/Ks in each site is its posterior expectation computed by

$$E((Ka/Ks)|\mathrm{data},T,\theta) \cong \sum_{i=1}^{k} P(\omega_i|\mathrm{data},T,\theta)\cdot(Ka/Ks)(\omega_i) = \frac{\sum_{i=1}^{k} P(\mathrm{data}|\omega_i,T,\theta)\cdot P(\omega_i)\cdot(Ka/Ks)(\omega_i)}{\sum_{i=1}^{k} P(\mathrm{data}|\omega_i,T,\theta)\cdot P(\omega_i)}. \quad (7)$$

Sites for which the expected values are larger than 1 and the posterior probability of $Ka/Ks > 1$ is higher than 95% are considered as undergoing positive selection.

Model Comparison

All data sets conducted in this study were analyzed with the MEC, M (M8 and M5), and E codon models. For the MEC model, we considered 3 empirical replacement amino acid probabilities matrices depending on the data analyzed: JTT (Jones et al. 1992) for nuclear and viral proteins, mtREV (Adachi and Hasegawa 1996) for mitochondrial proteins, and cpREV (Adachi et al. 2000) for chloroplast proteins, denoted by MECjtt, MECmt, and MECcp, respectively.

The MEC model presented here differs from the M model in that it allows instantaneous substitutions between pairs of codons that differ at 2 or 3 codon positions and in its ability to take into account the different replacement probabilities between amino acids. In order to evaluate the specific contribution of allowing instantaneous substitutions between codons differing in more than 1 nucleotide, we also compare the MEC, E, and M models with a variant of the M model, which allows such substitutions. Formally, this model is represented by the Q^0 matrix as in equation (5), with the inclusion of ω as in equation (3). We refer to this model as the M+ model.

The log-likelihood values obtained for each data set under the different models can be compared to test which model (MEC, M, M+, or E) best explains the data. The likelihood ratio test (LRT) is commonly used in order to test whether a certain model fits a particular data set significantly better than another model. However, the LRT is applicable only when 2 models are nested, which is not the case here. We thus used the second-order Akaike information criterion (AIC_c) (Akaike 1974), defined as

$$AIC_C = -2\log L + 2p\frac{N}{N-p-1}, \quad (8)$$

where L is the likelihood, p represents the number of free parameters, and N represents the sequence length.

Data Sets

Nuclear and Viral Data Sets

Thirteen data sets of protein-coding genes were analyzed. Nine multiple sequence alignments and tree topologies were taken from Yang et al. (2000) (referred to as D2, D3, D4, D6, D7, D8, D9, and D10 in this paper). Here, we renamed these data sets as D1–D8, respectively. Two data sets from Yang and Swanson (2002) were analyzed: the human class I major histocompatibility complex (MHC) and the abalone sperm lysine genes, denoted here as D9 and D10, respectively. Three additional data sets of the protein phosphatase 2C (PP2C) superfamily were analyzed (Stern et al. Forthcoming). The PP2C proteins are Mg^{2+}/Mn^{2+}-dependent serine/threonine phosphatases, which are essential for regulation of cell cycle and stress-signaling pathways in cells (Sun and Tonks 1994; Hanada et al. 2001). Each of the 3 data sets represents a pair of paralogous PP2C genes that were chosen because they are believed to be the result of a relatively recent duplication event (Stern et al. Forthcoming). Following is a description of these 3 PP2C data sets:

PP2Cα and PP2Cβ, denoted by D11, include 2 paralogous groups from different organisms: 6 sequences of PP2Cα (rat, human, mouse, bovine, chimpanzee, and rabbit) and 5 sequences of PP2Cβ (rat, human, mouse, bovine, and chimpanzee).

PP2Cζ and PPM1H, denoted by D12, include 2 paralogous groups from different organisms: 8 sequences of PP2Cζ (human, chimpanzee, mouse, rat, dog, tetraodon, fugu, and cow) and 9 sequences of PPM1H (human, chimpanzee, chicken, mouse, rat, dog, zebrafish, frog, and tetraodon).

POPX-1/FEM2 and POPX-2, denoted by D13, include 9 sequences of POPX-1/FEM2 (human, chimpanzee, dog, chicken, mouse, rat, fugu, frog, and zebra fish) and 10 POPX-2 sequences (human, chimpanzee, dog, chicken, mouse, rat, zebrafish, frog, tetraodon, and fugu).

Mitochondrial Data Sets

Twelve mitochondrial protein–coding genes (*cox1*, *cox2*, *cox3*, *cytb*, *nd1*, *nd2*, *nd3*, *nd4*, *nd4l*, *nd5*, *atp6*, and *atp8*) from 20 organisms were analyzed. These data sets as well as the tree topology were taken from a previous study by Cao et al. (1998).

Chloroplast Data Sets

Two chloroplast genes, *rbcL* and *matK*, were analyzed. These data sets are a subset of the data analyzed by Kato et al. (2003) (the archaeal sequences were excluded).

HIV Protease

A data set of HIV-1 protease sequences was used to demonstrate the ability of the new model to infer site-specific selection forces following drug treatment. Twenty-two sequences of patients that were treated with Amprenavir (APV) were extracted from the Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu/). The multiple sequence alignment is available as Supplementary Material online.

Results

Comparisons of Nuclear and Viral Genes

When comparing the fit of the different models to biological data sets, 12 out of the 13 tested data sets showed a significant improvement in the log likelihood under the MECjtt model as compared with the M model for the beta + ω prior distribution (table 1). Similar results were obtained when the gamma distribution was assumed, but in this case, MECjtt was significantly higher only in 11 out of the 13 data sets (table 2). For all 13 data sets both the MECjtt and the M models showed significantly higher maximum log-likelihood values compared with the E model (table 1).

The M+ model is a variant of the M model, in which instantaneous substitutions between pairs of codons that differ at 2 or 3 codon positions are allowed. Comparing the M with the M+ models shows that allowing such substitutions significantly improves the likelihood in 9 out of

Table 1
The AIC_c Scores and Maximum Log-Likelihood Values for the Analysis of 13 Data Sets under the M, M+, E, and MECjtt Models. The Beta + ω Prior Distribution Was Assumed for the M and MECjtt Models

			AIC_c Score (log likelihood)			
Data Set	SL[a]	NS[b]	M	M+	E	MECjtt
D1	144	17	7,469.16 (−3,686.1)	7,364.62 (−3,624.43)	9,844 (−4,882.32)	**7,110.82 (−3,495.53)**
D2	254	23	9,440.51 (−4,663.8)	9,398.02 (−4,631.7)	12,291 (−6,093.34)	**9,281.25 (−4,571.7)**
D3	490	22	19,611.73 (−9,755.1)	19,659.2 (−9,772.63)	27,044 (−13,477.3)	**19,417.48 (−9,650.52)**
D4	192	29	6,917.2 (−3,370.7)	6,927.03 (−3,364.47)	10,199 (−5,021.89)	**6,890.17 (−3,343.7)**
D5	947	23	18,833 (−9,365.9)	18,824.53 (−9,356.06)	26,535 (−13,222.6)	**18,592.26 (−9,238)**
D6	500	23	**13,797.19** (−6,845.4)	13,800.11 **(−6,840.64)**	19,100 (−9,503)	13,823.99 (−6,851.3)
D7	342	18	18,463.44 (−9,188.8)	18,347.05 (−9,124.18)	26,097 (−13,011.7)	**18,104.45 (−9,001.6)**
D8	91	13	2,294.97 (−1,106.4)	2,307.76 (−1,101.2)	3,353 (−1,645.42)	**2,283.66 (−1,086.6)**
D9	135	25	9,100.46 (−4,464.62)	8,979.09 (−4,389.61)	11,352 (−5,603.25)	**8,719.53 (−4,256.74)**
D10	270	192	12,687.95 (−7,234.74)	**12,411.42 (−7,071.04)**	17,402 (−9,619.33)	12,492.64 (−7,106.81)
D11	480	11	8,949.55 (−4,449.45)	8,912.94 (−4,425.54)	11,314 (−5,637.28)	**8,855.78 (−4,395. 82)**
D12	777	17	26,187.31 (−13,055.85)	25,860.88 (−12,887.1)	32,713 (−16,324)	**25,467.32 (−12,689.2)**
D13	832	19	39,599.32 (−19,757.59)	39,288.26 (−19,596.5)	48,253 (−24,089.8)	**38,924.91 (−19,413.7)**

NOTE.—Values are shown in bold for the model with the lowest AIC_c score and for the model with the highest log-likelihood value.
[a] Sequence length.
[b] Number of sequences.

13 data sets for the beta + ω prior distribution (table 1), and in 11 out of 13 data sets for the gamma prior distribution (table 2). In some cases, the M and M+ models differ in more than 100 points of likelihood, showing that taking into account multiple instantaneous substitutions between pairs of codons is not an artifact of overparameterization, but rather reflects the substitution pattern in the data.

The MECjtt is superior to the M+ model in 11 of the 13 tested data sets for the beta + ω prior distribution (table 1) and in 10 of the 13 data sets for the gamma prior distribution (table 2). The difference in log-likelihood scores between these model was sometimes larger than 100. Thus, integrating the empirical amino acid replacement probabilities into codon models significantly increases the fit of the model to the data.

Context-Dependent Models

The mitochondrial and chloroplast genomes are known to evolve under different selection pressures than nuclear genes, as indicated by the observed differences between the empirical mitochondrial (Adachi and Hasegawa 1996) and chloroplast (Adachi et al. 2000) matrices compared with the standard empirical matrix (JTT; Jones et al. 1992). Moreover, the nuclear and mitochondrial genomes use a different genetic code. We thus expect codon models derived from the organelle empirical amino acid matrices to better describe the evolution of mitochondrial and chloroplast genomes compared with the standard codon models. Therefore, we compared the M, M+, and the MECjtt models with a mitochondrial codon model, MECmt, or chloroplast codon model, MECcp, for mitochondrial and chloroplast data sets, respectively.

Mitochondrial Data Sets

We compared the fit of MECmt, MECjtt, M, and M+ models with 12 mitochondrial data sets. Table 3 contains log-likelihood and AIC_c values for these 12 data sets assuming the beta + ω distribution. For 11 data sets, the

Table 2
The AIC_c Scores and Maximum Log-Likelihood Values for the Analysis of 13 Data Sets under the M, M+, and MECjtt Models Assuming a Gamma Prior Distribution

			AIC_c Score (log likelihood)		
Data Set	SL[a]	NS[b]	M	M+	MECjtt
D1	144	17	7,471.47 (−3690.8)	7,359.8 (−3,625.9)	**7,102.68 (−3495.4)**
D2	254	23	9,440.45 (−4,663.78)	9,386.95 (−4,629.35)	**9,266.30 (−4,567.44)**
D3	490	22	19,814.67 (−9,858.9)	19,638.82 (−9,764.92)	**19,426 (−9,657.3)**
D4	192	29	6,906.99 (−3,369.77)	6,912.7 (−3,361.85)	**6,881.99 (−3,344.24)**
D5	947	23	18,865.44 (−9,384.3)	18,818.22 (−9,355.15)	**18,616.25 (−9,253)**
D6	500	23	13,797.55 (−6,848)	**13,796.03 (−6,841.1)**	13,820.55 (−6,852.1)
D7	342	18	18,458.07 (−9,188.7)	18,331.64 (−9,119.08)	**18,089.22 (−8,996.6)**
D8	91	13	2,285.74 (−1,105.9)	2,297.64 (−1,101.01)	**2,274.13 (−1,086.9)**
D9	135	25	9,091.42 (−4,465.35)	8,969.91 (−4390.97)	**8,702.61 (−4,257.32)**
D10	270	192	12,628.77 (−7215.95)	**12,385.86 (−7,068.18)**	12,474.76 (−7,112.63)
D11	480	11	8,948.64 (−4,451.21)	8,906.86 (−4,424.76)	**8,845.16 (−4,392.78)**
D12	777	17	26,184.67 (−13,056.7)	25,849.03 (−12,883.4)	**25,466.86 (−12,691.2)**
D13	832	19	39,605.05 (−19,762.7)	**39,281 (−19,595.1)**	39,611.43 (−19,759.2)

NOTE.—Values are shown in bold for the model with the lowest AIC_c score and for the model with the highest log-likelihood value.
[a] Sequence length.
[b] Number of sequences.

Table 3
AIC_c Scores and Maximum Log-Likelihood Values for the Analysis of 12 Mitochondrial Data Sets under the 4 Models: M, M+, MECmt, and MECjtt

Data Set	SL[a]	NS[b]	AIC_c Score (log likelihood)			
			M	M+	MECmt	MECjtt
atp6	226	20	12,837 (−6,366.63)	12,817.36 (−6,349.01)	**12,693.44 (−6,285.43)**	12,703.94 (−6,290.68)
atp8	71	20	**4,975.2** (−2,381.1)	4,995.59 (−2,352.71)	4,993.9 (−2,342.04)	4,983.32 **(−2,336.75)**
cox1	515	20	22,998.59 (−11,453.5)	23,007.26 (−11,451.8)	**22,813.09 (−11,353.5)**	22,851.29 (−11,372.6)
cox2	235	20	10,871.45 (−5,384.32)	10,882.9 (−5,382.39)	**10,749.59 (−5,314.15)**	10,771.79 (−5,325.25)
cox3	281	20	12,791.02 (−6,345.92)	12,779.46 (−6,333.05)	**12,562.62 (−6,223.17)**	12,643.88 (−6,263.8)
nd1	318	20	16,841.73 (−8,372.3)	16,815.55 (−8,352.42)	**16,518.09 (−8,202.3)**	16,601.47 (−8,243.99)
nd2	348	20	22,617.68 (−11,260.9)	22,545.84 (−11,218.4)	**22,044.53 (−10,966.4)**	22,091.53 (−10,989.9)
nd3	119	20	6,756.11 (−3,307.29)	6,763.6 (−3,303.03)	**6,642.1 (−3,239.45)**	6,684.44 (−3,260.62)
nd4	474	20	26,996.88 (−13,452.3)	26,938.39 (−13,416.9)	**26,358.67 (−13,125.8)**	26,474.67 (−13,183.8)
nd4l	98	20	5,845.11 (−2,847.72)	5,844.06 (−2,829.91)	**5,734.54 (−2,771.27)**	5,769.12 (−2,788.56)
nd5	616	20	37,734.15 (−18,821.9)	37,768.74 (−18,833.4)	**37,129.9 (−18,512.8)**	37,245.1 (−18,570.4)
cytb	392	20	19,835.97 (−9,870.81)	19,806.15 (−9,849.52)	**19,488.59 (−9,689.44)**	19,600.95 (−9,745.62)

NOTE.—Values are shown in bold for the model with the lowest AIC_c score and for the model with the highest log-likelihood value.
[a] Sequence length.
[b] Number of sequences.

highest log-likelihood and AIC_c values were obtained under MECmt. The second highest values were obtained under MECjtt. The 1 data set (*atp8*) that did not support the use of MEC models comprises relatively short sequences, containing only 71 sites. Thus, we hypothesize that there were not enough data in this protein to support the additional free parameters used in the MEC models.

Chloroplast Data Sets

We further analyzed 2 chloroplast data sets with the MECjtt, MECcp, M, and M+ models. In both data sets, the maximum log-likelihood values under the MECcp model were significantly higher as compared with the M model. However, for 1 data set a significant highest score was obtained under the M+ model. Surprisingly, for both data sets the maximum log-likelihood values obtained under the MECjtt model were higher compared with those obtained under the MECcp model (table 4). This point is further addressed in the discussion.

Site-Specific *K*a/*K*s of the HIV-1 Protease

To illustrate the potential of the MEC model, we focused on the inference of site-specific selection forces of the HIV-1 protease. HIV-1 protease is an essential enzyme for viral replication and is the target for design of antiviral drugs (Peng et al. 1989; Flexner 1998). The enzyme is an aspartic protease composed of 2 identical 99–amino acid monomers. Specific and well-characterized patterns of drug resistance mutations are associated with a variety of protease inhibitors (Condra et al. 1996; Molla et al. 1996; Craig et al. 1998; Patick et al. 1998). Such mutations are divided into 2 categories: primary resistance mutations, located at the substrate cleft, and secondary mutations, located elsewhere in the enzyme. Primary mutations generally reduce the inhibitor binding, whereas secondary mutations may compensate for the decreased kinetics inflicted by the primary mutation or confer resistance by altering enzyme catalysis, dimer stability, or inhibitor-binding kinetics or by reshaping the active site through long-range structural perturbations (Ho et al. 1994; Molla et al. 1996; Erickson et al. 1999; Barbour et al. 2002; Muzammil et al. 2003). It is assumed that such drug resistance–conferring sites are highly correlated with sites showing *K*a/*K*s values significantly greater than 1 (e.g., Chen et al. 2004).

We used the MEC model to study drug resistance of HIV-1 protease to a specific protease inhibitor, APV. Twenty-two sequences of protease were analyzed (see data set). The ω parameter was assumed to follow a gamma distribution. Tree topology was constructed by the Neighbor-Joining algorithm (Saitou and Nei 1987). As input for the Neighbor-Joining algorithm, pairwise distances were computed applying the ML criterion under the M model and assuming $\omega = 1$ for all sites and transition–transversion ratio $= 2$. Fixing the tree topology, all branch lengths as

Table 4
AIC_c Scores and Maximum Log-Likelihood Values for the Analysis of 2 Chloroplast Data Sets under the 4 Models: M, M+, MECjtt, and MECcp

Data Set	SL[a]	NS[b]	AIC_c Score (log likelihood)			
			M	M+	MECjtt	MECcp
matK	509	28	19,652.36 (−9,760.95)	**19,531.71 (−9,694.24)**	19,581.74 (−9,717.96)	19,596.76 (−9,725.47)
rbcL	394	64	16,760.15 (−8,185.32)	16,724.2 (−8,155.94)	**16,547.34 (−8,065.17)**	16,552.2 (−8,069.1)

NOTE.—Values are shown in bold type for the model with the lowest AIC_c score and for the model with the highest log-likelihood value.
[a] Sequence length.
[b] Number of sequences.

well as the 9 parameters of the model were estimated by ML. Visualization of site-specific *K*a/*K*s estimations under the MEC model was obtained by translating the scores to a discrete color scale and their projection onto the 3-dimensional structure (Protein Data Bank ID: 1T7J; Surleraux et al. 2005) using the RasMol program (Sayle and Milner-White 1995) (fig. 2). Positive selection was evident in 5 sites (10, 37, 54, 63, and 82). Each of these sites had a posterior expectation of *K*a/*K*s higher than 1 with a posterior probability of at least 0.95 (see Theory). Two additional sites (35 and 50) belong to categories that yield *K*a/*K*s > 1, with a posterior probability of at least 0.85. Of the predicted 7 sites, 5 (10, 50, 54, 63, and 82) are known to confer drug resistance (Shafer et al. 2000). Sites 10 and 63 contribute to resistance and belong to the secondary mutation category. Site 54 in the flap region (fig. 2) confers intermediate resistance. Site 50 within the cleft region (fig. 2) is known to confer high-level resistance. This site is capable by itself of reducing susceptibility to the APV drug and thus belongs to the primary resistance mutation category. Site 82 (cleft region; fig. 2) was detected as a positively selected site with a posterior probability of 0.98. However, in this site only valine to isoleucine (V82I) replacements are observed. Although site 82 is reported to be responsible for APV drug resistance, V to I mutations are not cited as APV drug resistance related (Shafer et al. 2000). Thus, the positive selection in this site might be explained by an adaptive response to other factors, such as the immune system, rather than a specific response to APV.

Five sites that are reported in the literature as drug resistance–conferring sites were not detected as undergoing positive selection. These sites (84, 46, 47, 32, and 90) did not show *K*a/*K*s values significantly >1. For some of these sites, clearly the data do not support positive selection. For example, in site 90 the sequence alignment contains only leucine and hence, the estimated *K*a/*K*s is very low (0.124). Site 84 belongs to the primary resistance category and results in a *K*a/*K*s score of only 0.38. However, according to the alignment, only 2 replacements are observed in that site (i.e., all sequences have isoleucine except 2 sequences that contain valine). In other sites (47 and 32) known as conferring low or intermediate resistance or contributing to resistance, the *K*a/*K*s values may indicate a weak purifying selection close to neutral selection (with *K*a/*K*s values around 0.85). Another explanation is that these sites show a mixture of strong purifying selection and positive selection in a lineage-specific manner. In site 46 (fig. 2), positive selection was suggested, with a *K*a/*K*s $= 1.3$; however, it was not statistically significant (posterior probability of 0.84).

In addition to predicting positive selection forces for sites that are known to confer drug resistance, we also predicted 2 sites (35 and 37) that were not previously reported as such. Site 37 is identified as having undergone positive selection with a *K*a/*K*s posterior expectation value equal to 2.75 and posterior probabilities of the positive selection categories equal to 0.97. Site 35 obtained a *K*a/*K*s estimate of 1.75 with a posterior probability of 0.85. Thus, we may predict that these sites may contribute to viral replication in the presence of APV. As can be seen in figure 2, these sites are structurally remote from the active site and are hence predicted to belong to the second category.

Discussion

To date, the majority of codon models are based either on theoretical assumptions or on empirical data. The model that we have developed here combines between these 2 approaches. Analysis of a wide variety of data sets shows that using a combined model has a large impact on the likelihood. On average, there was a difference of 122 points between the log likelihood under the MEC models and the log-likelihood under the M model with some data sets showing a difference of as many as 300 points. Furthermore, in comparison with the E model, the MEC model was shown to significantly better fit all analyzed data sets with an average log-likelihood difference of around 2,500 points. This result suggests that a strictly empirical codon model can be significantly improved if parameters that are highly variable among data sets are integrated within the model.

It is widely accepted that different proteins that belong to different organelles as well as different regions within a protein (such as transmembrane and nontransmembrane domains or different secondary structure elements) evolve under different evolutionary constraints. The context-dependent codon models described here directly take this variation into account. Noticeably, the majority of mitochondrial genes show an improvement in the log likelihood under a combined model derived specifically for mitochondrial genes (MECmt), as compared with the general combined model (MECjtt). This indicates the importance of accounting for the different replacement probabilities between amino acids evolving under different contexts. However, analysis of the chloroplast data does not show an advantage to the MECcp model over the MECjtt model. To test whether this result is due to a specific limitation of our MEC model, we compared the likelihood scores computed using JTT and cpREV models for the same data, this time using amino acids data instead of codons. The same trend was observed in this analysis as well, with JTT better fitting the data as compared with cpREV. Repeating these analyses with additional chloroplast data may eliminate this inconsistency.

The majority of the mechanistic codon models disallow instantaneous substitutions between codons that differ at 2 or 3 codon positions. The underlying assumption is that the probability for more than 1 codon position substitution in a small time interval may be negligible. Thus, for example, the rate of change between codon ATG (codes for methionine) and TTT or TTC (both code for phenylalanine) equals zero. However, empirical amino acid models such as JTT allow the instantaneous change between amino acid methionine and phenylalanine, which requires 2 substitutions in 2 codon positions. The superior performance of the M+ over the M model for the majority of the data sets suggests that the process at the DNA level may cause interdependence of substitutions at the 3 codon positions. This observation was previously pointed out by Yang et al. (1998).

Our model assumes that each site evolves independently of the other sites. However, this simplifying assumption is clearly not the case. Models that take into account dependencies among sites often assume dependencies only among adjacent positions (Yang 1995; Felsenstein and

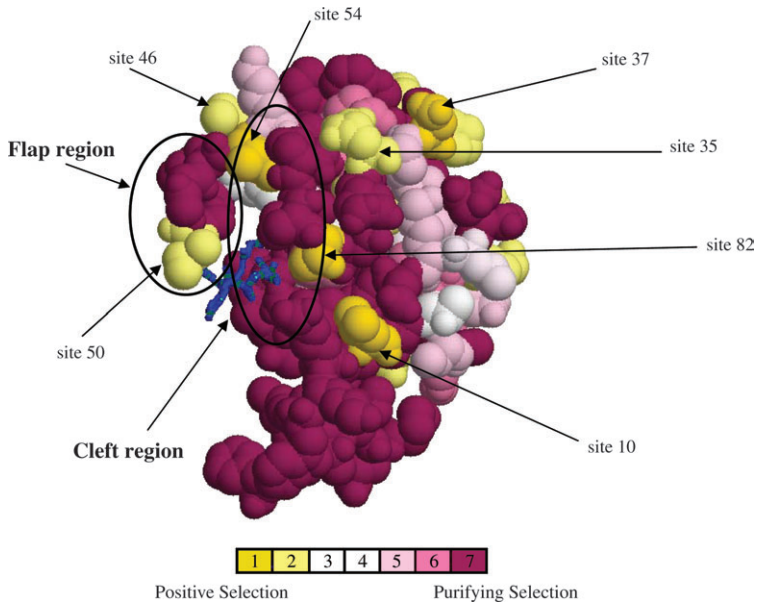

FIG. 2.—The selection pattern for HIV-1 protease chain A as inferred using the MECjtt model. The protein is represented as a space-fill model, where the *K*a/*K*s scores are color coded onto its van der Waals surface. The inhibitor APV is shown in blue as a backbone model. The color-coding bar shows the coloring scheme: burgundy corresponds to purifying selection, white corresponds to neutral selection, and dark yellow to positive selection.

Churchill 1996; Stern and Pupko 2006). Effort is also directed to integrate structural information into the evolutionary model, thus introducing relationship between nonsynonymous substitutions and protein structure (e.g., Robinson et al. 2003). One approach toward this goal involves using a 3-dimensional window for detecting the selection forces acting on the protein (Suzuki 2004a; Berglund et al. 2005). In another approach, the substitution model is constructed so that the tertiary structure is taken into account by using the empirical energy functions (or statistical potentials) (Parisi and Echave 2001; Robinson et al. 2003; Rodrigue et al. 2005; Rastogi et al. 2006). Using these functions, nonsynonymous substitutions rate depends on its effect on protein stability. Models that take into account 3-dimensional structural information as well as context-dependent models accounting, for example, for proteins' secondary structures, are not often used. It is hoped that the large increase in available protein structural information and the development of efficient algorithms for integrating such information into evolutionary models will boost the utility of such models in any phylogenetic analysis of protein-coding sequences.

A few methods were developed to detect positive selection operating on a specific lineage along a phylogenetic tree (Fares et al. 2002; Yang and Nielsen 2002; Berglund et al. 2005; Pond and Frost 2005). Because positive selection operates only on a few sites in short period of evolutionary time (Siltberg and Liberles 2002), methods that allow *K*a/*K*s ratio to vary both among sites and among lineages have better power in detecting positive selection. Relaxing the assumption of homogenous selection pressure among lineages can be easily accommodated in our suggested models by allowing ω to vary among branches.

Codon models are important not only for the inference of selection, but should also be applied for other phylogenetic based application. Absurdly, although most phylogenetic trees are reconstructed based on coding DNA sequences, the most realistic codon-based models are rarely used. This is also the case for ancestral sequence reconstruction, molecular dating, and the construction of multiple sequence alignments. In this sense, the codon models suggested here, which explicitly take into account variation of substitution rates between different amino acids, may be more suitable for these tasks. Furthermore, with the advent of more sophisticated algorithms for constructing amino acid replacement models (e.g., Muller and Vingron 2000; Muller et al. 2002), our approach becomes more feasible for computing a codon-based model for a specific kind of data.

Supplementary Material

The multiple sequence alignment is available available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Itay Mayrose, Adi Stern, Nimrod Rubinstein, Eyal Privman, Osnat Penn, Ofir Cohen, Gina Cannarozzi, Adrian Schneider, David Liberles, Herve Philippe, and 1 anonymous referee for their insightful comments. A.D.F. is an Israeli Ministry of Science Eshkol fellow. T.P. is a Yeshaia Horvitz fellow and is supported by grants from the Israel Science Foundation, from the Israeli Ministry of Science and Technology, and by the German–Israeli Foundation.

Funding to pay the Open Access publication charges for this article was provided by the Israeli Ministry of Science and Technology.

Literature Cited

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol. 42:459–468.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol. 50: 348–358.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automatic Control. 119:716–723.

Akashi H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene. 238:39–51.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Barbour JD, Wrin T, Grant RM, Martin JN, Segal MR, Petropoulos CJ, Deeks SG. 2002. Evolution of phenotypic drug susceptibility and viral replication capacity during long-term virologic failure of protease inhibitor therapy in human immunodeficiency virus-infected adults. J Virol. 76:11104–11112.

Berglund AC, Wallner B, Elofsson A, Liberles DA. 2005. Tertiary windowing to detect positive diversifying selection. J Mol Evol. 60:499–504.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol. 47:307–322.

Chen L, Perlina A, Lee CJ. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. J Virol. 78:3722–3732.

Condra JH, Holder DJ, Schleif WA, et al. (23 co-authors). 1996. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. J Virol. 70:8270–8276.

Craig C, Race E, Sheldon J, Whittaker L, Gilbert S, Moffatt A, Rose J, Dissanayeke S, Chirn GW, Duncan IB, Cammack N. 1998. HIV protease genotype and viral sensitivity to HIV protease inhibitors following saquinavir therapy. AIDS. 12:1611–1618.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Vol. 5 (Suppl 3). Washington (DC): National Biomedical Research Foundation. p. 345–352.

Erickson JW, Gulnik SV, Markowitz M. 1999. Protease inhibitors: resistance, cross-resistance, fitness and the choice of initial and salvage therapies. AIDS. 13 (Suppl A):S189–S204.

Fares MA, Elena SF, Ortiz J, Moya A, Barrio E. 2002. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol. 55:509–521.

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

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol. 13:93–104.

Flexner C. 1998. HIV-protease inhibitors. N Engl J Med. 338:1281–1292.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci. 27:315–321.

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

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185:862–864.

Hanada M, Ninomiya-Tsuji J, Komaki K, Ohnishi M, Katsura K, Kanamaru R, Matsumoto K, Tamura S. 2001. Regulation of the TAK1 signaling pathway by protein phosphatase 2C. J Biol Chem. 276:5753–5759.

Ho DD, Toyoshima T, Mo H, Kempf DJ, Norbeck D, Chen CM, Wideburg NE, Burt SK, Erickson JW, Singh MK. 1994. Characterization of human immunodeficiency virus type 1 variants with increased resistance to a C2-symmetric protease inhibitor. J Virol. 68:2016–2020.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18:486.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Jones DT, Taylor WR, Thornton JM. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. 339:269–275.

Kato Y, Aioi K, Omori Y, Takahata N, Satta Y. 2003. Phylogenetic analyses of Zostera species based on rbcL and matK nucleotide sequences: implications for the origin and diversification of seagrasses in Japanese waters. Genes Genet Syst. 78:329–342.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Eng. 8:641–645.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2:150–174.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics. 169:1753–1762.

Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol. 21:1781–1791.

Molla A, Korneyeva M, Gao Q, et al. (17 co-authors). 1996. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nat Med. 2:760–766.

Muller T, Spang R, Vingron M. 2002. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol. 19:8–13.

Muller T, Vingron M. 2000. Modeling amino acid replacement. J Comput Biol. 7:761–776.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715–724.

Muzammil S, Ross P, Freire E. 2003. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. Biochemistry. 42:631–638.

Naylor GJ, Collins TM, Brown WM. 1995. Hydrophobicity and phylogeny. Nature. 373:565–566.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol. 18: 750–756.

Patick AK, Duran M, Cao Y, Shugarts D, Keller MR, Mazabel E, Knowles M, Chapman S, Kuritzkes DR, Markowitz M. 1998. Genotypic and phenotypic characterization of human immunodeficiency virus type 1 variants isolated from patients treated

with the protease inhibitor nelfinavir. Antimicrob Agents Chemother. 42:2637–2644.
Peng C, Ho BK, Chang TW, Chang NT. 1989. Role of human immunodeficiency virus type 1-specific protease in core protein maturation and viral infectivity. J Virol. 63:2550–2556.
Pond SL, Frost SD. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol. 22:478–485.
Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4-Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics. 18 (Suppl 1): S71–S77.
Rastogi S, Reuter N, Liberles DA. 2006. Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys Chem. 124:134–144.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 20:1692–1704.
Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347:207–217.
Sainudiin R, Wong WS, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. J Mol Evol. 60:315–326.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Sayle RA, Milner-White EJ. 1995. RASMOL: biomolecular graphics for all. Trends Biochem Sci. 20:374.
Schneider A, Cannarozzi GM, Gonnet GH. 2005. Empirical codon substitution matrix. BMC Bioinformatics. 6:134.
Shafer RW, Kantor R, Gonzales MJ. 2000. The genetic basis of HIV-1 resistance to reverse transcriptase and protease inhibitors. AIDS Rev. 2:211–228.
Sharp PM. 1997. In search of molecular darwinism. Nature. 385:111–112.
Siltberg J, Liberles DA. 2002. A simple covarion-based approach to analyse nucleotide substitution rates. J Evol Biol. 15: 588–594.
Stern A, Privman E, Rasis M, Lavi S, Pupko T. Forthcoming. Evolution of the metazoan protein phosphatase 2C superfamily. J Mol Evol.
Stern A, Pupko T. 2006. An evolutionary space-time model with varying among-site dependencies. Mol Biol Evol. 23:392–400.
Sun H, Tonks NK. 1994. The coordinated action of protein tyrosine phosphatases and kinases in cell signaling. Trends Biochem Sci. 19:480–485.
Surleraux DL, Tahri A, Verschueren WG, et al. (15 co-authors). 2005. Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. J Med Chem. 48:1813–1822.
Susko E, Inagaki Y, Field C, Holder ME, Roger AJ. 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol Biol Evol. 19:1514–1523.
Suzuki Y. 2004a. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol Biol Evol. 21:2352–2359.
Suzuki Y. 2004b. New methods for detecting positive selection at single amino acid sites. J Mol Evol. 59:11–19.
Tang H, Wu CI. 2006. A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. Mol Biol Evol. 23:372–379.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18: 691–699.
Wong WS, Sainudiin R, Nielsen R. 2006. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinformatics. 7:148.
Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–1051.
Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics. 139:993–1005.
Yang Z. 2002. Inference of selection from multiple species alignments. Curr Opin Genet Dev. 12:688–694.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.
Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 15:1600–1611.
Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 19:49–57.

Herve Philippe, Associate Editor

Accepted November 2, 2006