BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i319–i327
doi:10.1093/bioinformatics/btm176

Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates

Itay Mayrose†, Adi Doron-Faigenboim†, Eran Bacharach and Tal Pupko*

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

ABSTRACT

Codon evolutionary models are widely used to infer the selection forces acting on a protein. The non-synonymous to synonymous rate ratio (denoted by Ka/Ks) is used to infer specific positions that are under purifying or positive selection. Current evolutionary models usually assume that only the non-synonymous rates vary among sites while the synonymous substitution rates are constant. This assumption ignores the possibility of selection forces acting at the DNA or mRNA levels. Towards a more realistic description of sequence evolution, we present a model that accounts for among-site-variation of both synonymous and non-synonymous substitution rates. Furthermore, we alleviate the widespread assumption that positions evolve independently of each other. Thus, possible sources of bias caused by random fluctuations in either the synonymous or non-synonymous rate estimations at a single site is removed. Our model is based on two hidden Markov models that operate on the spatial dimension: one describes the dependency between adjacent non-synonymous rates while the other describes the dependency between adjacent synonymous rates. The presented model is applied to study the selection pressure across the HIV-1 genome. The new model better describes the evolution of all HIV-1 genes, as compared to current codon models. Using both simulations and real data analyses, we illustrate that accounting for synonymous rate variability and dependency greatly increases the accuracy of Ka/Ks estimation and in particular of positively selected sites. Finally, we discuss the applicability of the developed model to infer the selection forces in regulatory and overlapping regions of the HIV-1 genome.

Contact: talp@post.tau.ac.il

1 INTRODUCTION

Likelihood methods combined with probabilistic models of sequence evolution are considered the state-of-the-art methods for phylogeny inference and allow robust parameter estimation and vigorous testing of evolutionary hypotheses (Whelan *et al.*, 2001; Yang, 2006). While amino-acid evolutionary models are restricted to computing the degree of purifying selection acting on each site (e.g. Mayrose *et al.*, 2004), codon evolutionary models can be used to detect both the purifying and the positive Darwinian selection forces (reviewed in Yang, 2005). The selection type and intensity are inferred by contrasting the ratio of non-synonymous (amino-acid altering; Ka) to synonymous (silent; Ks) substitution rates, ω. Sites showing ω significantly lower than 1 are regarded as undergoing purifying selection and may have a functionally or structurally important role. Sites with $\omega > 1$ are indicative of positive Darwinian selection, suggesting adaptive evolution.

The most widely used codon evolutionary models assume that the purifying selection acting on protein-coding DNA sequences is the result of selection that operates at the protein level only (Yang *et al.*, 2000). These models assume that synonymous substitutions are free from selection pressure and represent neutral evolution. In such a case, Ks is constant across codon positions, Ka is heterogeneous and the inference of site-specific ω values is based solely on Ka variations. Accumulating lines of evidence, however, suggest that this assumption is biologically unrealistic and that synonymous substitutions are regularly subjected to purifying selection (reviewed in Chamary *et al.*, 2006). This may be the result of, for example, constraints for maintaining mRNA secondary structure or *cis* regulatory motifs (e.g. exonic splicing enhancers) superimposed on the coding sequence of the gene.

Recently, Pond and Muse (2005) presented an evolutionary model that allows for site-to-site variation of both the synonymous and the non-synonymous substitution rates. This model was shown to better fit 9 out of 10 datasets analyzed. In this study it was also illustrated that sites inferred with a significant support to be positively selected under the synonymous constant model may be inferred as being under purifying selection when synonymous rate variation is included in the model (and vice versa). Notwithstanding, a worrying consequence of including site-to-site synonymous rate variation in the model is that now the inference of ω relies on the ratio of two inferred parameters. As such, random fluctuations in Ka and Ks estimates may more easily lead to an erroneous ω inference. Consider for example, a neutrally evolving site with both Ka and Ks equal to 1. Random fluctuations in the underlying sequences or in their sampling can easily shift the inference of Ka and Ks to 1.2 and 0.8, respectively, leading to an inferred positively selected site with $\omega = 1.5$. This shortcoming may be bypassed by including in the model our biological understanding that rates of evolution are correlated between adjacent sites.

A few models were previously developed that relaxed the unrealistic assumption that sites in DNA or protein sequences evolve independently (Felsenstein and Churchill, 1996; Stern and Pupko, 2006; Yang, 1995). Hidden Markov models

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(HMM) were used to account for the correlation between the rates of adjacent positions. These models were shown to better fit DNA and protein data and to improve the accuracy of site-specific evolutionary rate inference. When codon models are considered, dependencies in either Ka, Ks or both may exist. For example, in a functional region of a protein, consecutive positions with low non-synonymous rates are often observed. Similarly, exonic regulatory elements are characterized by a stretch of sites showing a reduction of both Ka and Ks rates (Goren *et al.*, 2006).

Here, we present models in which both the synonymous and non-synonymous rates vary among sites. Models assuming dependencies between either the Ka, the Ks or both are developed and studied. We show that the inference of positively selected sites is greatly influenced by considering synonymous rate variation, and that a model in which rate dependencies are accounted for greatly increases the accuracy of positive selection inference. The models are applied to study human immunodeficiency virus type 1 (HIV-1), a medically important test case for positive selection studies (e.g. de Oliveira *et al.*, 2004; Yang *et al.*, 2003). We demonstrate the usefulness of our novel models in analyzing overlapping viral genes, in detecting viral *cis*-regulatory elements, and in studying patterns of selection in specific viral genes and sites.

2 METHODS

2.1 The basic evolutionary model (KaC–KsC)

Our codon models are represented as a continuous time Markov process, defined by the instantaneous rate matrix Q. In the simplest model, the rate of changing from codon i to codon j (Q_{ij}) is defined as follows:

$$Q_{ij} = \begin{cases} \lambda_s \cdot \kappa \cdot \pi_j & i \text{ and } j \text{ differ by one synonymous transition} \\ \lambda_s \cdot \pi_j & i \text{ and } j \text{ differ by one synonymous transversion} \\ \lambda_a \cdot \kappa \cdot \pi_j & i \text{ and } j \text{ differ by one nonsynonymous transition} \\ \lambda_a \cdot \pi_j & i \text{ and } j \text{ differ by one nonsynonymous transversion} \\ 0 & \text{otherwise} \end{cases}$$

where λ_s denotes the synonymous substitution rate, λ_a denotes the non-synonymous substitution rate, κ denotes the transition versus transversion bias, and π_j is the target codon frequency calculated using the product of the observed nucleotide frequencies at the three codon positions (the F3 × 4 model of Yang *et al.*, 2000).

Because of the confounding effects between evolutionary rates and divergence times (Felsenstein, 1981), we can arbitrarily set $\lambda_s = 1$. The ratio of non-synonymous to synonymous rates, $\omega = \lambda_a/\lambda_s$, provides an indication of the magnitude of selective pressure. This model is similar to the codon models suggested by Nielsen and Yang (1998) and Muse and Gaut (1994). We refer to this model as KaC–KsC as both the Ka and Ks rates are constant across the gene (i.e. do not vary).

2.2 Variable Ka with constant Ks (KaV–KsC)

The KaV–KsC model expands the basic model above by allowing Ka to vary across sites. Thus, λ_s is shared by all sites and equals 1, while λ_a is treated as a random variable drawn independently for each site according to a pre-defined distribution (e.g. Gamma). This is the most commonly used model for detecting positive selection (Nielsen and Yang, 1998; Yang *et al.*, 2000).

2.3 Variable Ka and variable Ks (KaV–KsV)

In this model, both the synonymous and non-synonymous substitution rates are allowed to vary across sites. Thus, both the λ_a and λ_s parameters are treated as random variables sampled independently for each site from two given rate distributions, with the distribution of λ_s restricted to have mean 1. Here, we use the Gamma distribution $\Gamma(\alpha_a, \beta_a)$ to model the non-synonymous rate variation. The synonymous rate variation is modeled by the unit mean Gamma distribution $\Gamma(\alpha_s, \alpha_s)$. The Gamma distributions are approximated by C_a and C_s equally probable discrete categories (Yang, 1994), representing the non-synonymous and synonymous distributions, respectively. This model was previously suggested by Pond and Muse (2005).

2.4 Modeling site-dependencies (KaD–KsD)

In this model, rate dependencies between codon sites are assumed. The non-synonymous substitution rate λ_a^i at site i is dependent on the non-synonymous substitution rate λ_a^{i-1} at site $i-1$. Similarly, we also assume that the synonymous substitution rate λ_s^i and λ_s^{i-1} are dependent. Once the sites are assigned to their rates, each position evolves independently. In each site there are C_a possible non-synonymous rate categories and C_s synonymous rate categories. The dependency between rates at adjacent positions is modeled using two Markov chains that operate on the spatial dimension: one describes the dependency between non-synonymous rates while the other describes the dependency between synonymous rates. The two Markov chains are assumed to be independent of each other. Site dependency within each chain is modeled using the parameters $\rho_a, \rho_s \in (0, 1)$ for the non-synonymous and synonymous chains, respectively. Larger ρ values indicate higher correlation between adjacent sites.

Practically, instead of modeling two independent chains with C_a and C_s states we can use a combined Markov chain with $C_a \times C_s$ states. The state S_{jk} denotes a site with a non-synonymous rate category j and synonymous rate category k.

The data are then analyzed using an HMM, in which the rates at each position are the hidden states (Durbin *et al.*, 1998). The HMM is characterized by the transition probabilities between the states of the combined Markov chain, by the initial probabilities of the hidden states and by the emission probabilities, which represent the probability of the observations given assignments to the hidden states.

Let T represent the transition matrix between any two states. The transition probability between the states S_{jk} and S_{lp} is

$$(T^{jk \to lp} | \alpha_a, \beta_a, \alpha_s, \rho_a, \rho_s) = (T^{j \to l} | \alpha_a, \beta_a, \rho_a) \times (T^{k \to p} | \alpha_s, \rho_s) \quad (1)$$

where $(T^{j \to l} | \alpha_a, \beta_a, \rho_a)$ is the transition probability between the non-synonymous states j and l. $(T^{k \to p} | \alpha_s, \rho_s)$ is computed similarly for the synonymous transition. These probabilities are calculated using the correlated bivariate gamma distribution (Stern and Pupko, 2006; Yang, 1995).

The initial probabilities of the hidden states (the rates) in the combined chain are computed using the probabilities obtained by the discrete gamma approximation. Thus, the initial probability of being at the non-synonymous states j and the synonymous state k is

$$\pi_{jk} = P(S_j | \alpha_a, \beta_a) \times P(S_k | \alpha_s) \quad (2)$$

where $P(S_j \mid \alpha_a, \beta_a)$ and $P(S_k \mid \alpha_s)$ are the initial probabilities of the non-synonymous category j and the synonymous category k, respectively. Since each rate category is given equal probability (Yang, 1994) the initial probabilities are simply $1/(C_a \times C_s)$.

The emission probabilities are the likelihoods of the data at each position. Specifically, $P(d_i | S_{jk})$ is the probability of observing the data at position i, given a certain rate assignment S_{jk}. This probability is computed using standard approaches (Felsenstein, 1981). Note that all

computations are done assuming the phylogenetic tree topology and its branch lengths are given, but for simplicity, we omit them from the equations.

To summarize, the parameters of the KaD–KsD model are $\theta = \{\alpha_a,\ \beta_a,\ \alpha_s,\ \rho_a,\ \rho_s\}$. Two variants of this model exist: the KaV–KsD model which assumes dependencies between synonymous rates only, and the KaD–KsV model which assumes dependencies between non-synonymous rates only. These two models can be obtained from the KaD–KsD model by constraining $\rho_a = 0$, or $\rho_s = 0$, respectively. The parameters of the model are estimated by maximizing the likelihood function, $P(d|\theta)$. The likelihood can be calculated using the backward dynamic programming algorithm (Durbin *et al.*, 1998). Let $b_{jk}(i) = P(d_{i+1}, \ldots, d_n|S^i_{jk})$ be the probability of observing the partial data from sites $i+1$ through n, given that the combined non-synonymous and synonymous state at site i is S_{jk}. where n is the sequence length. Then

$$b_{jk}(i) = \sum_{l=1}^{C_a}\sum_{p=1}^{C_s} T^{jk\to lp} b_{lp}(i+1) P(d_{i+1}|S^{i+1}_{lp}) \tag{3}$$

with $b_{jk}(n) = 1$. The likelihood is thus

$$L = P(d_1, \ldots, d_n|\theta) = \sum_{j=1}^{C_a}\sum_{k=1}^{C_s} \pi_{jk} b_{jk}(1) P(d_1|S^1_{jk}) \tag{4}$$

2.5 Estimating site-specific synonymous and non-synonymous rates

Selective pressure at individual sites can be inferred by calculating the Ka/Ks ratio at each site and the posterior probability of a site to evolve under positive selection pressure. The posterior probability of site i belonging to state S_{jk} is

$$P^i_{jk} = \frac{b_{jk}(i) f_{jk}(i)}{P(d)} \tag{5}$$

where $f_{jk}(i) = P(d_1, \ldots, d_i|S^i_{jk})$ is the joint probability of observing sites 1 through i given that the combined state at site i is S_{jk}. $f_{jk}(i)$ can be calculated using the forward dynamic algorithm (Durbin *et al.*, 1998)

$$f_{jk}(i) = P(d_i|S_{jk}) \sum_{l=1}^{C_a}\sum_{p=1}^{C_s} T^{lp\to jk} f_{lp}(i-1) \tag{6}$$

with $f_{jk}\ (1) = P(d_1,\ S_{jk}) = \pi_{jk} P(d_1 \mid S_{jk})$

Point estimates of $\hat{k}^i_a$ and $\hat{k}^i_s$ can then be obtained by calculating the expectation over the empirical posterior distribution. $\hat{k}^i_a$ is obtained by

$$\hat{k}^i_a = E(\lambda^i_a|d) = \sum_{j=1}^{C_a}\sum_{k=1}^{C_s} \lambda_a(S_{jk}) \cdot P^i_{jk} \tag{7}$$

where $\lambda_a(S_{jk})$ is the non-synonymous rate assignment in state S_{jk}. $\hat{k}^i_s$ is similarly obtained.

The posterior probability of site i to evolve under positive selection pressure is then the sum of posterior probabilities of states, in which the non-synonymous rate is higher than the synonymous rate

$$PS(i) = \sum_{j=1}^{C_a}\sum_{k=1}^{C_s} P^i_{jk} \cdot 1\{\lambda_a(S_{jk}) > \lambda_s(S_{jk})\} \tag{8}$$

2.6 Model comparison

The likelihood ratio test (LRT) can be used in order to determine which model best fits the data. The LRT is applicable since the models are nested (except for KaD–KsV and KaV–KsD): when $\rho_a,\ \rho_s = 0$, KaD–KsD collapses to KaV–KsV. KaV–KsV collapses to KaV–KsC when α_s approaches infinity. The differences in log-likelihood between the models are compared to a χ^2 distribution to obtain a P-value (Yang, 2006). Hence, the additional parameters are statistically justified if the log-likelihood improvement between KaD–KsD and KaV–KsV is at least 4.6 or 3.32 between KaV–KsV and KaV–KsC; P-value <0.01 according to the χ^2_2 and χ^2_1 asymptotic distributions, respectively. Model comparisons using the 2nd order Akaike Information Criterion (AIC_c) (Burnham and Anderson, 2002) were also performed and gave essentially identical results (not shown).

2.7 HIV-1 datasets

To test the utility of the proposed model we analyzed the nine coding genes of HIV-1. Aligned nucleotide sequences for each gene were retrieved from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/). The multiple sequence alignment (MSA) of each gene was then modified according to the following steps: (i) A reference sequence for the gene, taken from the HIV-1 complete genome (accession number AF033819) was added to the alignment using the profile to profile alignment option of CLUSTALW (Thompson *et al.*, 1994); (ii) Sequences containing missing data and stop codons inside the reading frame, or sequences not starting with ATG, were removed from the analysis. This last criterion was not applied to the *pol* dataset since it naturally does not start with the ATG codon; (iii) In order to eliminate the number of gaps in the alignment, sequences that opened insertion positions that are shared by less than 25% of the sequences were removed and (iv) In order to save computational time only the 80 most divergent sequences of the remaining alignment were used. The most divergent sequences were chosen using the following procedure: first, all pairwise distances were calculated. Next, the sequence with the maximal distance to one of the sequences already selected was iteratively added. The phylogenetic tree for each gene was reconstructed using the neighbor-joining algorithm (Saitou and Nei, 1987) with pairwise distances estimated using the Jukes and Cantor distance for codons. The parameters of each model were then optimized using the maximum likelihood (ML) paradigm. The gamma distribution was approximated using three categories. In models that include both Ka and Ks variation, three categories are assumed for Ka and three for Ks, resulting in nine possible rate states. The approximation of 3 categories was used to determine the ML estimates of the model parameters. In order to increase accuracy when computing site-specific $\hat{k}_a$ and $\hat{k}_s$, the number of discrete rate categories was increased from three to eight using the same ML parameter estimates.

2.8 Simulation study

Simulations were conducted in order to examine the accuracy of site-specific $\hat{k}_a$ and $\hat{k}_s$ under the different models. We simulated each site with a specific 'true' rate. An MSA was thus generated based on a vector of true rates. Subsequently, $\hat{k}_a$ and $\hat{k}_s$ rates for each column were inferred using the different models. For the simulations, one must provide a true rate for each site. In order to obtain true rates that are biologically relevant, characteristic rates were computed based on the empirically inferred rates of the *gag* and *vif* datasets (Table 1). For each dataset, 30 and 15 sequences were simulated. The model tree for each dataset was obtained using the 30 (15) most divergent sequences of the original dataset. For each dataset, three vectors of true Ka and Ks rates were obtained: those inferred with KaV–KsC, KaV–KsV or KaD–KsD. Simulating with true rates inferred with a specific model may bias the results towards this model. Thus, for each simulated dataset, we conducted three separate analyses, corresponding to the three vectors of true rates. In total, 12 simulation scenarios were performed (*gag* and *vif* genes, 15 and 30 sequences, each with three vectors of true

Table 1. Log-likelihood (LL) values for nine HIV-1 coding genes under the five models analyzed

Gene	SL[a]	LL scores	LL differences[b]			
		KaV−KsC	KaV−KsV	KaV−KsD	KaD−KsV	KaD−KsD
env	883	−64 826.8	914.9	1044.3 (129.4)	947.1 (32.2)	1079.9 (165)
gag	500	−25 669.6	362.4	395.8 (33.4)	375.2 (12.8)	408.6 (46.2)
nef	210	−19 421.5	339.3	377.5 (38.2)	341.8 (2.5)	380.1 (40.8)
pol	1004	−56 194.1	1346	1507 (160.4)	1394.6 (48)	1565.1 (218.5)
rev	123	−12 265.5	227.8	232 (4.2)	243.5 (15.7)	247.5 (19.7)
tat	101	−9 252	213.8	222.8 (8)	219.3 (5.5)	228.2 (14.3)
vif	192	−15 138.5	239.4	273.8 (34.4)	244.1 (4.7)	278.8 (39.4)
vpr	98	−8004.3	130.4	153.5 (23.2)	130.1 (0)	154.2 (23.7)
vpu	82	−9194	187.7	194.3 (6.6)	190.7 (3)	197.1 (9.3)

[a]Sequence length.
[b]Log-likelihood difference compared to the KaV–KsC model.
The log-likelihood differences compared to the KaV–KsV model are given in parentheses.

rates). For each simulated scenario 10 independent and identical runs were conducted. The accuracy of inference was measured by the mean absolute deviation (MAD) distance between the simulated and inferred Ka/Ks values

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |\text{estimated } \omega^i - \text{true } \omega^i| \quad (9)$$

where the estimated ω^i is $\hat{k}_a^i/\hat{k}_s^i$.

3 RESULTS

3.1 HIV-1 dataset analysis

To test the utility of the proposed model the nine coding genes of HIV-1 were analyzed. For all genes, models that account for the synonymous rate variation were significantly superior to the commonly used KaV–KsC model, which assumes that all positions evolve with the same Ks rate (Table 1). The minimal log-likelihood difference between the KaV–KsC model and the models accounting for Ks rate variation was 130, and the maximal difference was more than 1500. The hypothesis of constant synonymous rate was thus rejected at $P \ll 0.0001$ for all HIV-1 genes. The coefficient of variation (CV), measured by the SD divided by the mean, was used to compare the dispersion of the estimated Ka and Ks distributions. The CV of the non-synonymous rate distribution was found to be in the same order of magnitude as that of the synonymous rate distribution (Table 2), further supporting the essentiality of correctly modelling Ks variation.

A strong pattern of rate dependency among adjacent positions was found for all HIV-1 genes. This was true for both the Ka and the Ks rates. When comparing a model that accounts for dependencies of both Ka and Ks rates (the KaD–KsD model) with a model that does not (KaV–KsV), a significant increase in log-likelihood was found for all genes analyzed ($P < 0.001$ for all genes). As expected, the difference in log-likelihood is correlated with the sequence length (Table 1). For example, for *pol*, the longest gene, the difference was in the order of hundreds. We next tested what factor contributed most to the superiority of KaD–KsD over KaV–KsV: the dependency among synonymous rates or the dependency among non-synonymous rates. Dependency of Ka rates corresponds to regional selection pressure at the protein level, while Ks dependency reflects regional selection on the coding DNA or mRNA. Taking into account Ka dependency only (the KaD–KsV model) significantly increased the log-likelihood as compared to the independent model (KaV–KsV) in six out of nine HIV-1 genes (Table 1). This result is in agreement with previously published works emphasizing the importance of taking into account dependencies at the protein level (e.g. Stern and Pupko, 2006). Taking into account dependency among Ks rates only (comparing the KaV–KsV and the KaV–KsD models), resulted in a significant increase in log-likelihood in all datasets. Interestingly, the contribution of Ks rate dependency was higher than the contribution of Ka rate dependency for all but the *rev* dataset. The higher Ks rate dependency was also evident in the comparison of the models' correlation parameters: the inferred correlation between Ks rates was higher than the Ka rate correlation in all genes aside from *rev* (Table 2). Thus, our results strongly indicate that Ks rates vary in a spatial manner along the genes.

Table 2. The correlation between adjacent Ka rates (ρ_a) and adjacent Ks rates (ρ_s) and the coefficient variation (CV) of the Ka and Ks distributions inferred by the KaD–KsD model for the nine HIV-1 genes

Gene	ρ_a	ρ_s	CV(Ka)	CV(Ks)
env	0.4	0.88	1.24	0.64
gag	0.35	0.76	1.47	0.57
nef	0.2	0.8	1.11	0.59
pol	0.54	0.84	1.64	0.65
rev	0.64	0.42	0.9	0.82
tat	0.4	0.55	1	0.9
vif	0.29	0.81	1.02	0.64
vpr	0	0.84	1.21	0.52
vpu	0.37	0.59	0.92	0.75

Table 3. Simulation results: accuracy of site-specific ω inference based on the different models under different simulation scenarios

Simulation scenario	Gene	True rates[b]	NS[c]	Mean MAD[a]			
				KaV–KsC[d]	KaV–KsV[e]	KaV–KsD	KaD–KsD
1	*vif*	KaV–KsV	15	0.86 ($P<10^{-5}$)	0.84 ($P<10^{-3}$)	**0.75**	0.78
2	*vif*	KaD–KsD	15	0.70 ($P<10^{-5}$)	0.70 ($P<10^{-4}$)	**0.62**	0.63
3	*vif*	KaV–KsC	15	0.48 ($P<10^{-5}$)	0.50 ($P<10^{-5}$)	0.45	**0.35**
4	*gag*	KaV–KsV	15	0.50 ($P<10^{-3}$)	0.49 ($P<0.01$)	0.51	**0.43**
5	*gag*	KaV–KsC	15	0.55 ($P<10^{-6}$)	0.55 ($P<10^{-4}$)	0.53	**0.38**
6	*gag*	KaD–KsD	15	0.41 ($P<0.001$)	0.41 ($P<0.01$)	0.42	**0.36**
7	*vif*	KaV–KsV	30	0.80 ($P=0.08$)	0.80 ($P=0.13$)	**0.76**	0.77
8	*vif*	KaV–KsC	30	0.43 ($P<0.01$)	0.44 ($P<0.01$)	0.43	**0.37**
9	*vif*	KaD–KsD	30	0.65 ($P<10^{-4}$)	0.64 ($P<0.01$)	**0.54**	0.58
10	*gag*	KaV–KsV	30	0.45 ($P<0.01$)	0.45 ($P<0.01$)	0.44	**0.42**
11	*gag*	KaV–KsC	30	0.44 ($P<0.01$)	0.44 ($P<0.01$)	0.42	**0.37**
12	*gag*	KaD–KsD	30	0.36 ($P<10^{-4}$)	0.37 ($P<10^{-5}$)	**0.33**	0.34

[a]MAD values are the average over 10 identical and independent runs. Values are shown in bold type for the model that achieved the best accuracy in each simulated scenario.
[b]True rates were inferred with the given model using the *gag* or *vif* datasets in Table 1.
[c]Number of sequences.
[d]*P*-values between KaV–KsC and KaD–KsD.
[e]*P*-values between KaV–KsV and KaD–KsD.

3.2 Accuracy of rate estimation: a simulation study

Simulations were used to compare the inference accuracy of site-specific ω values under the various models. This comparison tested the accuracy of inference across the whole ω range (i.e. both positive and purifying selection) and did not concentrate only on inferring positive selection. As shown in Table 3, the inference obtained using the KaD–KsD model was constantly more accurate compared to the inference obtained using either the KaV–KsV or KaV–KsC models. These results were significant for 11 out of 12 simulated scenarios ($P<0.01$; paired *t*-test). The comparison between KaD–KsD and KaV–KsD was inconclusive as none of the two was consistently more accurate than the other across the 12 simulated scenarios tested.

We next concentrated on the success of various models to specifically infer sites that are under positive selection. The receiver operating characteristic (ROC) curve was used in order to compare the precision and sensitivity of inference under the different models. The closer the curve to the upper left corner, the more successful the prediction is. An example of two ROC curves for simulation scenarios 1 and 2 (Table 3) is shown in Figure 1. These results demonstrate that ω values inferred with the KaV–KsC model are consistently the least accurate. The accuracy of the KaD–KsD was greater than that inferred with KaV–KsV when the simulated rates were taken from an empirical rate vector inferred with KaD–KsD (see methods for simulation procedures), while the two models achieved comparable accuracy when simulating with KaV–KsV. In both simulated scenarios, however, the highest precision and accuracy was achieved with the KaV–KsD model. The superiority of the KaV–KsD model over KaD–KsD seen in Figure 1 can be attributed to the small number of sequences (15) simulated. The parameter-rich KaD–KsD model may

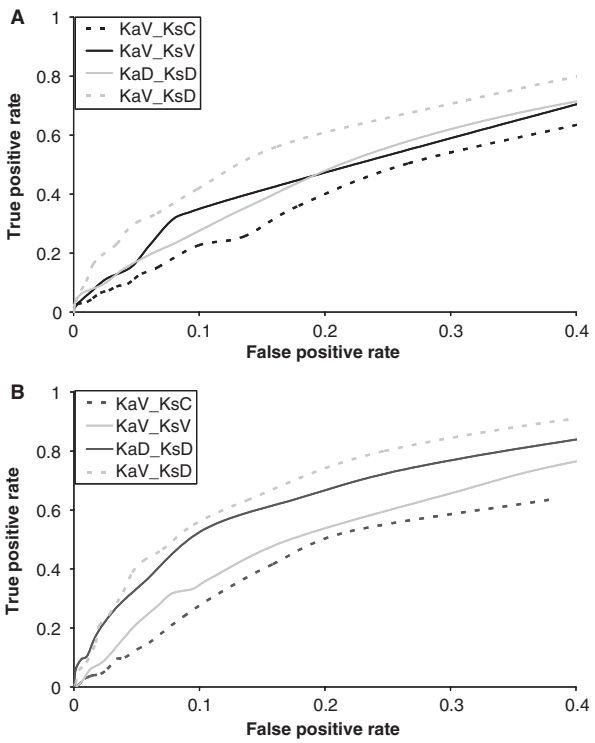

Fig. 1. ROC curves for predicting positively selected sites. **(A)** ROC curve under simulation scenario 1 and **(B)** ROC curve under simulation scenario 2. Positively selected sites are regarded as those that were simulated with $\omega>1.5$ and purifying selected sites (true negatives) are those that were simulated with $\omega<0.9$. Sites simulated with $0.9<\omega<1.5$ are on the boundary between positive, purifying and neutral evolution and were excluded from the analysis. Other cutoff values gave similar results.

Table 4. The number of inferred positively selected sites for each HIV-1 gene under the different models[a]

Gene	KaV–KsC	KaV–KsV	KaV–KsD	KaD–KsV	KaD–KsD
env	149 (17%)	108 (12%)	54 (6%)	108 (12%)	52 (6%)
gag	65 (13%)	33 (7%)	27 (5%)	33 (7%)	25 (5%)
nef	37 (18%)	19 (9%)	17 (8%)	20 (10%)	16 (8%)
pol	89 (9%)	38 (4%)	18 (2%)	33 (3%)	23 (2%)
rev	52 (43%)	30 (24%)	30 (24%)	30 (24%)	28 (23%)
tat	41 (41%)	28 (27%)	27 (27%)	29 (29%)	27 (27%)
vif	24 (13%)	23 (11%)	19 (10%)	24 (12%)	19 (10%)
vpr	16 (16%)	4 (4%)	4 (4%)	4 (4%)	5 (5%)
vpu	18 (22%)	12 (15%)	11 (13%)	12 (15%)	11 (13%)

[a]Posterior probability > 0.9. The percentage of positively selected sites out of the whole gene are given in parentheses.

overfit such a small dataset. Indeed, when the number of sequences is large (the 80 HIV-1 sequences, see below) no significant differences in the inferences of the two models are observed.

3.3 Inferring positive selection across the HIV-1 genome

The number of positively selected sites for each HIV-1 gene analyzed is given in Table 4. In agreement with previous studies (de Oliveira *et al*., 2004; Yang *et al*., 2003) positively selected sites were abundantly found in all nine coding genes under all models. However, the number of positively selected sites was found to be highly sensitive to the model applied. Accounting for heterogeneous synonymous rates substantially reduced the number of inferred positively selected sites (Table 4). Accounting for dependency among Ks rates further reduced the number of positively selected sites inferred (Table 4). Taken together, more than twice as many positive sites are inferred by KaV–KsC compared to that by KaD–KsD.

It is expected that the positively selected sites inferred using the KaD–KsD model would be a subset of the sites inferred under the KaV–KsC model. This, however, was not generally the case. The three models together inferred 623 distinct positively selected sites. Only 135 sites (22% of the total) were shared among all three models (Figure 2). The KaD–KsD model exhibited the lowest proportion of uniquely inferred sites and the highest proportion of sites that are shared by all three models. Furthermore, the agreement between KaV–KsC and the other two models was the lowest. Hence, sites inferred under the KaD–KsD model seem to be the most consistent with sites inferred using the other two models. Of course, this consistency does not ensure correct results, but its absence in the other models is alarming. The high sensitivity of the KaV–KsC and KaV–KsV models questions the validity of their assumptions for the data at hand.

Across all HIV-1 genes, the results obtained with KaV–KsD were highly similar to those obtained with KaD–KsD. The results obtained with KaD–KsV highly overlapped those obtained with KaV–KsV (Table 4). This further indicated that accounting for the dependence among Ks rates is more significant than the dependence among Ka rates.

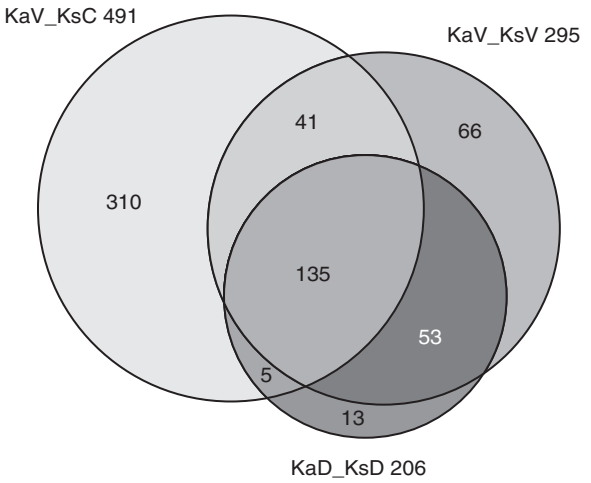

Fig. 2. The total number of positively inferred sites (Posterior probability > 0.9) across the nine genes of HIV-1 for the KaV–KsC, KaV–KsV and KaD–KsD models.

3.4 Positive selection in overlapping regions

The existence of overlapping genes is a widespread phenomenon in the genomes of small viruses (Pavesi, 2006). In overlapping regions, we expect constraints over synonymous and non-synonymous substitutions that are different from those in non-overlapping regions. In fact, overlapping regions are expected to display low and correlated Ks and Ka rates. Our KaD–KsD is especially useful for analyzing a gene that overlaps with another gene, as both Ks variability and spatial dependence is expected.

Here we illustrate that the KaD–KsD model accounts for the selection forces acting on the overlapping region between the *vif* and *pol* genes. The overlapping region spans 18 codons (sites 1–18 of *vif* with sites 986–1004 of *pol*). Figure 3 shows that this region (the 5′ of *vif* and the 3′ of *pol*) exhibits a substantial reduction of Ks rates as compared to the rest of the gene. Site 12 of the *vif* gene shows a peculiar selection pattern. Although part of the overlapping region, it exhibits a very high Ks rate and low Ka rate (Figure 3A). What can explain such a high Ks variation in this site? The explanation is provided in Figure 3B. Site 12 of *vif* corresponds to site 999 in the *pol* gene. Focusing on this site in *pol*, the high Ka/Ks ratio ($\omega = 11.4$) suggests intensive positive selection acting on this site. Thus, the positive selection in *pol* drives the high Ks variability of the corresponding site in *vif*. The purifying selection at the protein level, extracted on this position in *vif* does not permit any non-synonymous changes, thus increasing the Ks rate only. Position number 8, which overlaps position 994 in *pol* that has the second highest ω value does not exhibit a very large peak in its Ks rate. Alternatively, both its Ka and Ks rates are elevated. The purifying selection extracted on position number 8 is probably less significant than that extracted on position number 12, thus allowing for both substitution types. Finally, it is important to note that these observations cannot be obtained when analyzing the data with the commonly used KaV–KsC model. Under this model, the Ks peak at site 12 of *vif* cannot be detected. Furthermore, the assumption of a

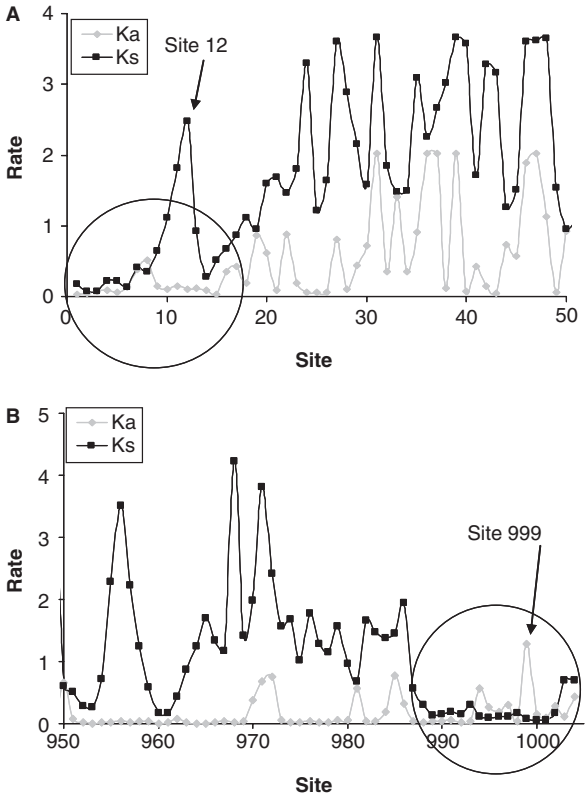

Fig. 3. Ka and Ks values for a region of 50 codon sites of **(A)** *vif* and **(B)** *pol* as inferred by the KaD–KsD model. The overlapping regions are marked by circles. The Ka and Ks rates are shown by gray and black lines, respectively.

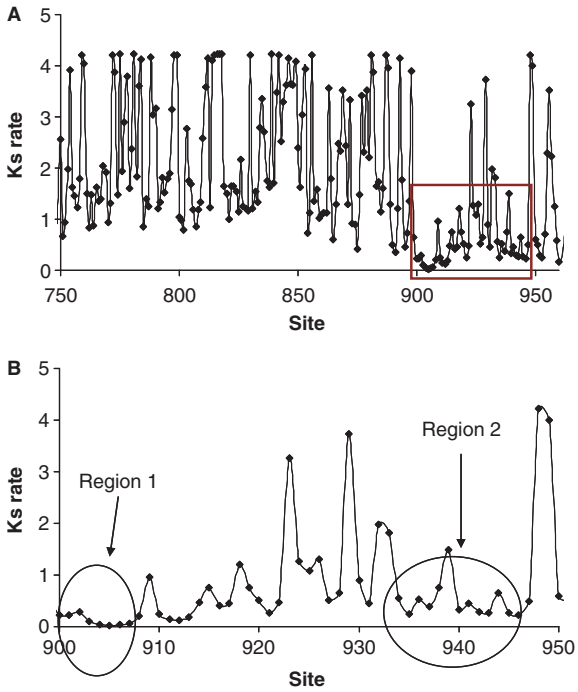

Fig. 4. **(A)** Ks variability over the *pol* gene. The segment of the *pol* gene in which the Ks rate is substantially reduced is rectangled. **(B)** A 'zoom-in' on positions 900–950 showing the locations of the ultraconserved Ks regions. Region 1 contains the cPPT and T-rich elements. Region 2 contains the CTS element.

constant Ks rate across the entire *pol* coding region would have obscured the signal for positive selection in site 999 ($\omega = 1.06$, posterior probability $= 0.81$ in KaV–KsC versus $\omega = 11.4$ and posterior probability > 0.99 in KaD–KsD).

3.5 Identifying regulatory elements in coding regions

The KaD–KsD model allows the identification of coding regions, in which a spatial purifying selection operates at the nucleic acid level. Using this model, we searched for stretches with low synonymous rates across the *pol* sequence. In addition to the overlapping region described above, a second region with a low synonymous rate was revealed (codon positions 900–947, Figure 4A). Part of this sequence (positions 900–907, Figure 4B) exhibits exceptional low rate of Ks (as well as a low Ka rate), suggesting that the primary sequence of the DNA, or the RNA, has a functional role. This segment is enriched with purines. Indeed, HIV and other lentiviruses have two polypurine tracts (PPTs) with a known functional role: both PPTs serve as RNA primers for the reverse transcriptase in the synthesis of the plus-strand DNA. One is located at the 3′ untranslated region of the viral genome. The second is located at the center of the genome, hence called the central PPT (cPPT) (Charneau and Clavel, 1991). This cPPT locus is mapped exactly to the ultraconserved Ks region in positions 902–906 of *pol*. A short stretch of T-rich motif (TTTT) immediately upstream of the cPPT is correlated with the ultraconservation found in positions 900–901. These codon positions were experimentally shown to be important for reverse transcription (Ilyinskii and Desrosiers, 1998).

Among the 48 conserved codons, a second region (positions 934–947) exhibits exceptionally high level of Ks conservation (Figure 4B). This region nicely overlaps a *cis*-acting signal named the central termination sequence (CTS), located at positions 930–939. During reverse transcription, a DNA structure, called DNA flap, is formed adjacent to this site (Charneau *et al.*, 1994). This structure has been shown to be involved in the nuclear import of the HIV genome (Zennou *et al.*, 2000). Altogether, our results demonstrate that functionally important *cis*-acting signals can be identified using our model.

4 DISCUSSIONS

Estimating the selection pattern at the codon level is at the heart of evolutionary research. Various methods and techniques were developed to increase the accuracy of Ka/Ks inference. Earlier methods were based on counting (e.g. Nei and Gojobori, 1986). In these methods, the number of synonymous and non-synonymous substitutions in each codon is inferred and normalized by the number of synonymous and non-synonymous site. While usually very fast, these counting techniques are statistically problematic as no less than four parameters are estimated for each site (the numbers of synonymous and non-synonymous substitutions and the numbers of synonymous and non-synonymous sites).

Muse (1996) further demonstrated that in these methods the estimation of non-synonymous and synonymous rates is not independent of each other. The alternative to these counting methods is to use an explicit model of codon substitution. The inference of Ka/Ks ratio is a by-product of these models. These models tend to be computationally intensive as they involved exponentiating a large (61 by 61) rate matrix. Consequently, current models use oversimplified assumptions such as fixed Ks variation and that each site evolves independently of each other. Thus, when the goal is to estimate Ks variability, one has to resort to counting techniques. In this study, we showed how Ks variability and dependency can be incorporated into probabilistic evolutionary models. We have demonstrated that these models better fit biological data and can significantly increase the accuracy of Ka/Ks estimates.

Yang (2006) and Hurst (2002) have discussed the difficulty of a reliable inference of ω per site. This difficulty, which results in a high number of false positive predictions, is due to the large number of inferences being made at a single analysis, and to the inference of ω as a ratio of two estimated parameters (Anisimova *et al*., 2002). An extensive research was aimed at reducing the number of false positive predictions. Specifically, effort was concentrated on model comparisons, on accurate estimation of model parameters, and on finding the best prior distribution of the non-synonymous rates (Anisimova *et al*., 2001, 2002; Swanson *et al*., 2003; Wong *et al*., 2004; Yang *et al*., 2000, 2005). For example, Anisimova *et al*. (2001, 2002) recommended to search for positively selected sites using several alternative models and choose those sites that are common to all, thus ensuring the robustness of the results. Noteworthy, all these competing models differ only in the assumed distribution of the ω parameter. Recently, the Bayes–Empirical–Bayes approach was developed (Yang *et al*., 2005), taking into account the sampling errors of the ML estimates of model parameters. This approach was shown to reduce the number of falsely inferred positively selected sites, particularly when small datasets are analyzed. Here we show that the number of falsely inferred positive sites can also be sharply decreased by incorporating the dependence between adjacent synonymous and non-synonymous rates. Using such models as KaD–KsD has the effect of eliminating random noise in site-specific Ka and Ks estimates.

Characterizing the selection forces acting on viral genomes at the DNA level is of great interest. For example, there are well-known RNA secondary structures along the HIV-1 genome, such as the Rev responsive element, that play important functional roles during the virus life cycle (Malim *et al*., 1990). Another obvious example is the existence of overlapping genes, in which a strong selection on both synonymous and non-synonymous substitutions is expected. However, since the commonly used codon models fail to account for Ks variation, such regions are usually excluded from the analysis, despite their being of high medical interest. For example, in order to avoid overlapping regions, Yang *et al*. (2003) excluded the whole *tat*, *rev* and *vpu* genes when analyzing selection forces in HIV-1. These three genes, however, are the ones exhibiting the largest fraction of sites that are positively selected (Table 4). Furthermore, assuming a constant synonymous rate for the whole gene is likely to overestimate the actual synonymous rate at the overlapping region and to underestimate the rate at the non-overlapping region. Consequently, the number of positively selected sites is expected to be too low for the overlapping region and too high for the non-overlapping region. The models suggested in this study provide a statistically robust approach to study selection at such overlapping and regulatory elements.

Our analysis revealed a 50-codon-long sequence in the *pol* open reading frame with exceptionally low synonymous and non-synonymous rates. This sequence roughly starts and ends with sequences known to be functionally important, namely the TTTT-rich box, the cPPT and the CTS signals, all act in *cis* during the process of HIV-1 reverse transcription (Charneau and Clavel, 1991; Charneau *et al*., 1994; Ilyinskii and Desrosiers, 1998). This further indicates that our model has the ability to identify important regulatory sequences embedded in open reading frames. Not all conserved segments in this region have a known function. Yet, one can speculate that such segments serve as specific binding sites for cellular and/or viral factors. Support for this hypothesis comes from the fact that these sequences form a DNA flap structure that was suggested to mediate the transport of the HIV-1 genome into the nucleus (Zennou *et al*., 2000).

Positively selected positions are classically defined as those in which the Ka/Ks ratio is significantly larger than 1. This definition, fundamental to evolutionary research, is based on the assumption that the synonymous rates reflect neutral rate of evolution. However, others and we have shown that synonymous rates vary considerably over sites. This indicates that purifying or maybe even positive, rather than neutral evolution, characterizes the synonymous substitutions to a certain extent. If, for example, synonymous sites are under purifying selection, average Ks values are an underestimate to the neutral rate of evolution, and hence, the $Ka/Ks>1$ criterion is too liberal. If, however, positive selection is acting on synonymous sites, then Ks is an overestimate of the neutral rate, and the Ka/Ks criterion might be too conservative. Furthermore, when selection operates at the DNA or mRNA level, the Ka and Ks values are expected to be correlated. Altogether, the variability of Ks questions the validity of the standard Ka/Ks tests for detecting positively selected sites. This Ks variability might vary among different proteins and organisms. In non-viral sequences, an alternative approach is to compare the non-synonymous rate with the evolutionary rate of intronic regions or with the evolutionary rates of pseudogenes. In viral sequences that lack such neutral sequences, how to correct the test for positive selection in a way that accounts for the synonymous variability is an open question that calls for additional research.

ACKNOWLEDGEMENTS

We thank Adi Stern, Julien Dutheil and Eyal Privman for critically reading the manuscript. This study is supported by a grant to E.B. and T.P. from the Israeli Ministry of Science. T.P. and E.B. were also supported by the Israeli Science Foundation grants number 1208/04 and 1184/05. A.D.F. is an Israeli Ministry of Science Eshkol fellow.

Conflict of Interest: none declared.

REFERENCES

Anisimova,M. *et al*. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol*., **18**, 1585–1592.

Anisimova,M. *et al*. (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol*., **19**, 950–958.

Burnham,K.P. and Anderson,D.R. (2002) *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*. Springer-Verlag, New York, USA.

Chamary,J.V. *et al*. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet*., **7**, 98–108.

Charneau,P. and Clavel,F. (1991) A single-stranded gap in human immunodeficiency virus unintegrated linear DNA defined by a central copy of the polypurine tract. *J. Virol*., **65**, 2415–2421.

Charneau,P. (1994) HIV-1 reverse transcription. A termination step at the center of the genome. *J. Mol. Biol*., **241**, 651–662.

de Oliveira,T. *et al*. (2004) Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? *Genetics*, **167**, 1047–1058.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

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

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

Goren,A. *et al*. (2006) Comparative analysis identifies exonic splicing regulatory sequences – the complex definition of enhancers and silencers. *Mol. Cell*, **22**, 769–781.

Hurst,L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet*., **18**, 486.

Ilyinskii,P.O. and Desrosiers,R.C. (1998) Identification of a sequence element immediately upstream of the polypurine tract that is essential for replication of simian immunodeficiency virus. *EMBO J*., **17**, 3766–3774.

Malim,M.H. *et al*. (1990) HIV-1 structural gene expression requires binding of the Rev trans-activator to its RNA target sequence. *Cell*, **60**, 675–683.

Mayrose,I. *et al*. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol*., **21**, 1781–1791.

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

Muse, S.V. (1996) Estimating synonymous and nonsynonymous substitution rates, *Mol Biol Evol*., **13**, 105–114.

Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol*., **3**, 418–426.

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

Pavesi,A. (2006) Origin and evolution of overlapping genes in the family Microviridae. *J. Gen. Virol*., **87**, 1013–1017.

Pond,S.K. and Muse,S.V. (2005) Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol*., **22**, 2375–2385.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4**, 406–425.

Stern,A. and Pupko,T. (2006) An evolutionary space-time model with varying among-site dependencies. *Mol. Biol. Evol*., **23**, 392–400.

Swanson,W.J. *et al*. (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol*., **20**, 18–20.

Thompson,J.D. *et al*. (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. *et al*. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*., **17**, 262–272.

Wong,W.S. *et al*. (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,W. *et al*. (2003) Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J. Mol. Evol*., **57**, 212–221.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol*., **39**, 306–314.

Yang,Z. (1995) A space-time process model for the evolution of DNA sequences. *Genetics*, **139**, 993–1005.

Yang,Z. (2005) The power of phylogenetic comparison in revealing protein function. *Proc. Natl Acad. Sci. USA*, **102**, 3179–3180.

Yang,Z. (2006) *Computational Molecular Evolution*. Oxford University Press, Oxford.

Yang,Z. *et al*. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yang,Z. *et al*. (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol*., **22**, 1107–1118.

Zennou,V. *et al*. (2000) HIV-1 genome nuclear import is mediated by a central DNA flap. *Cell*, **101**, 173–185.