MAXIMUM LIKELIHOOD ANALYSIS OF ADAPTIVE EVOLUTION IN HIV-1 GP120 *ENV* GENE

ZIHENG YANG

Galton Laboratory, Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, England

Every functional protein appears to have some conserved amino acids which are critically important to the basic structure and function of the protein and thus under purifying selection. Some proteins also have variable amino acids which, when changed, offer a selective advantage, and thus undergo adaptive evolution. These amino acids are also important to the structure and function of the protein, although in a different way. It seems that the selective pressure in every protein varies among sites. Maximum likelihood models developed recently for comparison of silent and replacement nucleotide substitution rates allow for variable selective pressures among amino acid sites, and provide a powerful approach to studying the evolutionary process of protein-coding genes. This paper applies the likelihood models to analyze a data set of 186 HIV-1 gp120 *env* gene sequences for comparison with a previous analysis of the same data set. The maximum likelihood analysis identified a number of sites under positive selection, some in the conserved regions of the protein.

1 Introduction

The HIV-1 *env* gene is one of the best-known examples of molecular adaptation (e.g., [1-4]). The envelope glycoprotein interacts with the receptors on target cells in the host, and amino acids involved in such interactions are expected to be constrained by purifying selection to maintain viral infectivity. On the other hand, the same viral protein is involved in immune recognition by the host, and amino acid changes resulting in viral escape from the host immune surveillance will be at a selective advantage. Such evolutionary changes will be promoted by Darwinian selection. Positive selection in the HIV-1 *env* gene has been demonstrated by viral escape experiments [5, 6] and by the observed higher nonsynonymous (amino acid-altering, d_N) than synonymous (silent, d_S) substitution rates [1, 4].

Comparison of synonymous and nonsynonymous substitution rates provides an important means for studying the selective pressure on a protein. The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a sensitive measure of selective pressure on the protein, with values of $\omega = 1$, > 1 and < 1 indicating neutral evolution, positive (diversifying) selection and negative (purifying) selection, respectively. Note that the definition of ω accounts for the fact that more nonsynonymous than synonymous mutations are expected due to the structure of the genetic code. Most previous studies calculated the d_N and d_S rates

(and their ratio ω) by averaging over all sites and ignored the fact that the selective pressure varies among amino acid sites. As many amino acids may be under strong structural or functional constraints with very small d_N or ω and adaptive evolution likely affects only a few amino acids, such analysis rarely found ω ratios >1 or detected positive selection [7]. Some genes showed similar synonymous and nonsynonymous rates and similar rates at the three codon positions and were thus falsely claimed to be completely neutral [8, 9], while the patterns in fact reflect strong diversifying selection at a few sites and purifying selection at others.

Recent efforts have focused on methods that account for variable selective pressures among sites. Fitch *et al.* [10] and Suzuki and Gojobori [11] reconstructed sequences in extinct ancestors using parsimony and counted synonymous and nonsynonymous changes along the phylogeny at each site to determine which sites have undergone excessive nonsynonymous substitutions and may thus be under positive selection. Nielsen and Yang [12] and Yang *et al.* [13] developed probabilistic models of codon substitution that allow for variable ω ratios among sites and implemented them in the likelihood framework. The method uses a statistical distribution to describe the variation of ω among sties; that is, there are several classes of sites in the sequence, which have different ω ratios. We use the likelihood ratio test to determine whether allowing for sites with $\omega > 1$ significantly improves the fit of the model to data. If the ω ratio for any site class is > 1, we use the Bayes theorem to calculate the (posterior) probability that each site, given its data, is from such a site class. Sites with high probabilities are inferred to be under positive selection (see [12, 13] for details of the models). The maximum likelihood (ML) method has several advantages. The substitution model accounts for important features of DNA sequence evolution that are often ignored by other methods, such as biased transition and transversion rates and biased codon usage. The model naturally accounts for the genetic code, and the probability theory underlying the method corrects for multiple hits properly. Applications of the ML method to real data suggest that the method is powerful in detecting adaptive evolution at a few sites in a background of purifying selection [12-14]. For example, in an analysis of HIV-1 *nef* gene, Zanotto *et al.* [15] detected a number of sites under positive selection, many of which corresponded to previously-identified epitopes, while both pairwise comparison and sliding window analysis failed.

The data sets analyzed in most of the studies mentioned above are relatively small. Recently, Yang [16] made some improvements to the ML algorithm, enabling the method to be applied to large data sets of a few hundred sequences. In this paper, I use the ML method to reanalyze the data set of Yamaguchi-Kabata and Gojobori [4] containing 186 HIV-1 gp120 *env* gene sequences. The *env* gene is well characterized and provides a good test example. It is also interesting to compare the ML analysis with the previous analysis [4], which used inferred ancestral sequences to count synonymous and nonsynonymous changes at each site along the tree [11].

2 Data and Methods

The data consisted of 186 HIV-1 *env* gene (gp120) sequences from subtype B, aligned and analyzed by Yamaguchi-Kabata and Gojobori [4]. Since my analysis does not require inference of ancestral sequences, I did not use the outgroup subtype-D sequence those authors used. The reference sequence for numbering amino acid positions is from strain HXBc2 (K03455). Following Yamaguchi-Kabata and Gojobori, sites involving alignment gaps are removed, with 421 codons left in each sequence. Those sites are identified in figure 1, together with the amino acids in the reference HXBc2 sequence. Site 31T is removed in the present analysis, although it was used and found to be under positive selection by Yamaguchi-Kabata and Gojobori [4].

The methods of Nei and Gojobori [17] and Yang and Nielsen [18] are used to to perform pairwise comparison to calculate the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site and the number of nucleotide substitutions per codon (t), resulting in three distance matrices for each method. The neighbor-joining method [19] was used to infer phylogenies from each distance matrix, with the `neighbor` program of the PHYLIP package used [20]. The resulting six trees were evaluated using the `baseml` program in the PAML package [21], with different substitution rates, transition/transversion rate ratios, and base frequencies assumed for the three codon positions. The tree based on the t distance from the method of Yang and Nielsen [18] had the highest likelihood and was used in fitting codon-substitution models later. The estimated substitution rates at the three codon positions are in proportions 1 : 0.87 : 1.28, while the transition/transversion rate ratios for the three positions are $\kappa = 2.98$, 3.26, and 4.68.

Codon-substitution models that allow the selective pressure, indicated by the ω ratio, to vary among sites are fitted to the sequence data by ML [13]. Parameters involved in those models are explained below in the Results section. The `codeml` program in the PAML package [21] is used.

3 Results

3.1 Pairwise Sequence Comparison

Estimates of synonymous (d_S) and nonsynonymous (d_N) substitution rates in pairwise comparisons of the 186 sequences are plotted in figure 1. The method of Yang and Nielsen [18] is used. In some pairs, $d_N > d_S$, while in the majority of comparisons, $d_N < d_S$. As those distances are averages over all sites in the sequence, they are not very informative about the selective pressures exerted on the protein. However, it is noteworthy that the calculated d_N/d_S ratios are much higher than in most other

proteins (*e.g.*, [22]).

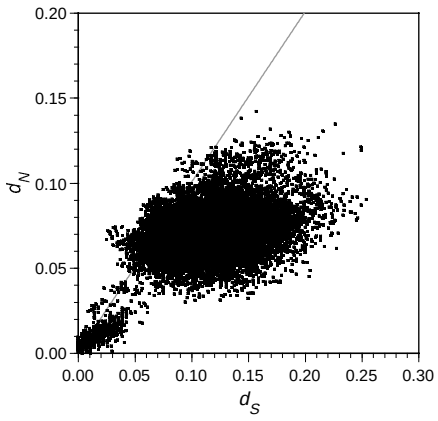

Fig. 1. *Estimates of synonymous (d_S) and nonsynonymous (d_N) substitution rates in pairwise comparisons using the method of Yang and Nielsen (2000). The straight line represents the neutral expectation* $d_N = d_S$

3.2 Estimation of Substitution Parameters and Likelihood Ratio Test of Positive Selection

Table 1 shows the log-likelihood values and parameter estimates obtained under different models of variable ωs among sites [12, 13]. Model M0 assumes one ω ratio for all sites; the estimated ratio (0.593) suggests that, on average, purifying selection dominates the evolution of the *env* gene. The neutral model (M1) assumes a class of conserved sites with $\omega_0 = 0$ and another class of neutral sites with $\omega_1 = 1$. The proportion p_0 for the conserved class is estimated by ML, with $p_1 = 1 - p_0$. This model does not allow for sites with $\omega > 1$ and serves as a null model for testing positive selection. Model M2 (selection) adds a third site class with the proportion (p_2) and the ω ratio (ω_2) estimated from the data. The estimates suggest a large proportion of sites (19%) under strong diversifying selection with $\omega_2 = 4.3$ (table 1). M3 (discrete) also assumes three site classes, with both the proportions and ω ratios estimated from data. This model suggests that about 10% of sites are under diversifying selection with $\omega_2 = 3.1$. Model M7 (beta) assumes a beta distribution of ω among sites. The beta distribution with parameters p and q can take different shapes in the interval (0, 1) and is a flexible null model for testing positive selection. The ML estimates under this model (table 1) indicates a U-shaped distribution. Model M8 (beta&ω) adds another site class with the proportion p_1 and the ratio ω estimated from the data. Parameter estimates under this model suggest that 90% of

sites have ω from a U-shaped beta distribution, and about 10% of sites are under positive selection with $\omega = 2.8$. Note that all models that can allow for sites with $\omega > 1$ (M2, M3, and M8) do suggest presence of such sites, although the exact proportion and the ω ratio vary.

Table 1. Log-likelihood values and ML estimates of parameters

Model code	p	ℓ	Estimates of Parameters	d_N/d_S
M0: one-ratio	1	–31,065.09	$\omega = 0.593$	0.593
M1: neutral	1	–30,690.37	$p_0 = 0.189$ ($p_1 = 0.811$)	0.811
M2: selection	3	–29,741.64	$p_0 = 0.182$, $p_1 = 0.628$ **($p_2 = 0.190$)**, **$\omega_2 = 4.295$**	1.443
M3: discrete	5	–29,040.62	$p_0 = 0.618$, $p_1 = 0.283$ **($p_2 = 0.098$)**, $\omega_0 = 0.122$, $\omega_1 = 0.970$, **$\omega_2 = 3.083$**	0.654
M7: beta	2	–29,402.73	$p = 0.322$, $q = 0.485$	0.399
M8: beta&ω	4	–29,022.33	$p_0 = 0.899$, $p = 0.403$, $q = 0.705$ **($p_1 = 0.101$)**, **$\omega = 2.799$**	0.609

Note.—p is the number of parameters in the ω distribution. Estimates of κ are between 3.0 and 3.7 among models. Estimates of parameters indicating positive selection are in bold. ℓ is the log-likelihood value, while d_N/d_S is the average ω over sites.

Table 2. Likelihood ratio test statistics for testing positive selection

Models compared	$2\Delta\ell$	χ^2 significance value (1%)	d.f
M0 vs. M3	4048.94	13.28	4
M7 vs. M8	760.80	9.21	2

Those models can be compared using the likelihood ratio test (table 2). For example, the comparison between models M0 and M3 is a test of the assumption that the selective pressure indicated by the ω ratio is constant among sites. The huge likelihood difference between the two models suggests that ω varies greatly among sites (table 2). While this is more a test of variability than a test of positive selection, parameter estimates under M3 suggest presence of sites under positive selection (with $\omega_2 > 1$). The comparison between M7 and M8 examines directly whether allowing for sites with $\omega > 1$ in M8 leads to significant improvement to M7's fit to data. Again the likelihood improvement is significant (table 2). In sum, those models provide strong statistical evidence for presence of amino acid sites under diversifying selection.

3.3 Inference of Sites Under Positive Selection

Models M3 (discrete) and M8 (beta&ω) are found to produce very similar results, and below I use M3 as an example. Parameter estimates under the model suggest three site classes in the proportions $p_0 = 0.618$, $p_1 = 0.283$, $p_2 = 0.098$ with $\omega_0 = 0.122$, $\omega_1 = 0.970$, and $\omega_2 = 3.083$ (table 1). Those proportions are the prior probabilities for each site. We use the Bayes theorem to calculate the (posterior) probabilities that each site, given the data at that site, is from the three site classes. If the posterior probability for the ω_2 class, *P*, is large, say >95%, we may conclude that the site is under positive selection. For example, at site 25, the posterior probabilities are 0.000, 0.095, and 0.905, so that there is *P* = 90.5% chance that the site is from the ω_2 class. The posterior probability distributions for sites in the *env* gene are shown in figure 2.

Table 3. List of positively selected sites

Sites	Location	Sites	Location	Sites	Location
19T	Signal sequence	275V	C2	**343K**	**C3**
33K	C1	279D	C2	344Q	C3
85V	**C1**	**283T**	**C2**	346A	C3
87V	**C1**	**291S**	**C2**	347S	C3
161I	V2	**306R**	**V3**	360I	C3
164S	V2	**308R**	**V3**	363Q	C3
169V	V2	317F	V3	**(387S)**	**V4**
200V	**C2**	**333I**	**C3**	389Q	V4
209S	C2	**336A**	**C3**	**440S**	**C4**
219A	C2	**337K**	**C3**	**(442Q)**	**C4**
232T	C2	339N	C3	467I	V5
240T	**C2**	**340N**	**C3**	(500K)	C5

Note.— *Amino acid sites with the posterior probability of coming from the ω_2 class P >99% or >95% (in parentheses). Those in bold were also identified by Yamaguchi-Kabata and Gojobori* [4]. *Location designations are according to ref.* [4] *(see also [23, 24]).*

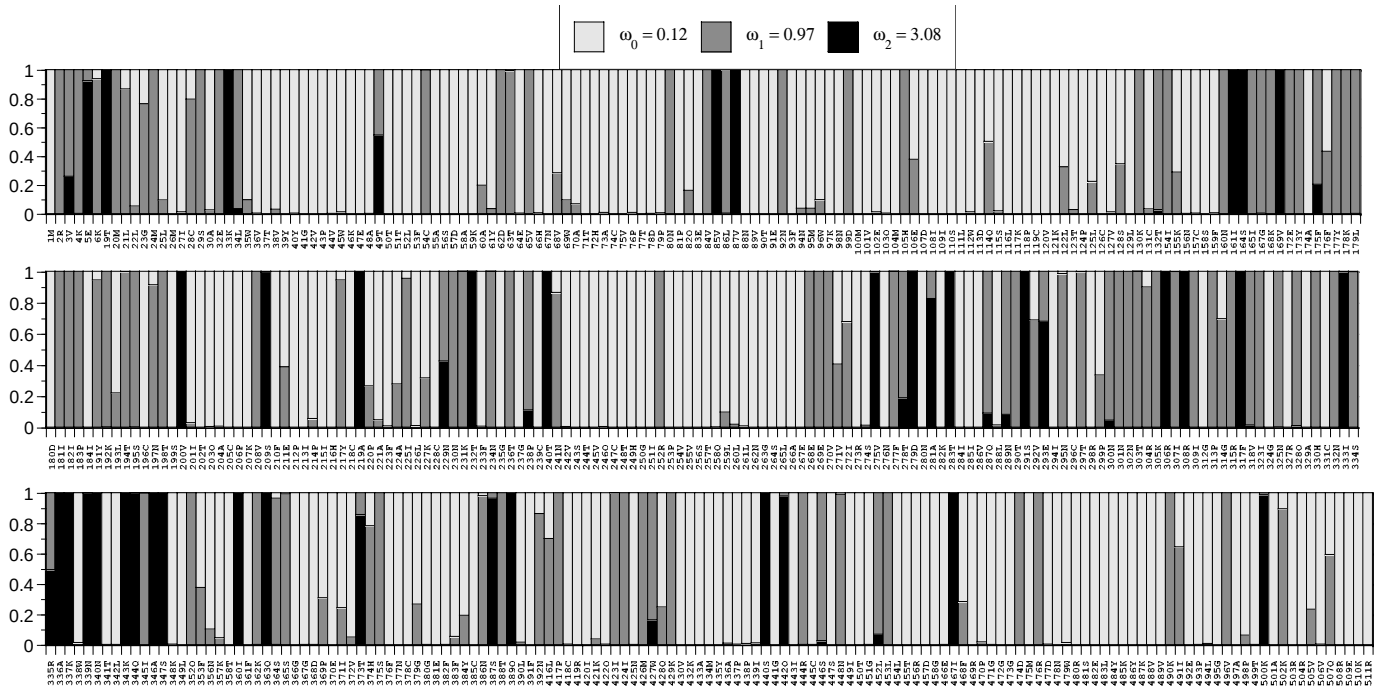

Fig. 2. *Posterior distribution of* **ω** *for each site calculated under M3 (discrete). The protein consists of five "variable" loop regions and five "constant" regions [4, 23, 24]. These are C1: 29S-125L, V1: 126C-157C, V2: 158S-196C, C2: 197N-295N, V3: 296C-330H, C3: 331C-384Y; V4: 385C-417P; C4: 418C-460N; V5: 461S-471G; C5: 472G-511R. Note that some variable sites involving alignment gaps are removed, with only 421 codons analyzed.*

Thirty-six sites had posterior probabilities >95%; those sites are listed in table 3. Models M3 and M8 produced the same list of sites at this significance level. Twenty-six of the sites are in the conserved regions ("C"), and nine are in the variable loops ("V"). Eighteen of the sites were also identified by Yamaguchi-Kabata and Gojobori [4] to be under positive selection, while the remaining 18 sites are not. Also 15 sites identified by those authors to be under positive selection are not in the list of the present analysis (fig. 2). These are 31T, 130K, 132T, 172E, 178K, 195S, 281A, 293E, 303T, 319T, 322K, 335R, 429K, 444R, and 446S. Among them, sites 31, 319, 322 are not included in the present analysis, and sites 281, 293, and 335 showed evidence, although not very strong, of positive selection in the present analysis as well. The other nine sites (130, 132, 172, 178, 195, 303, 429, 444, and 446), however, have low probabilities for the ω_2 class (see fig. 2).

To examine whether the differences are caused by the use of different tree topologies, the tree used by Yamaguchi-Kabata and Gojobori [4] is applied to fit the discrete model (M3). The log likelihood for this tree is –29052.77, slightly worse than the tree used to produce table 1. The estimates of parameters are $p_0 = 0.614$, $p_1 = 0.288$ ($p_2 = 0.098$), $\omega_0 = 0.121$, $\omega_1 = 0.985$, $\omega_2 = 3.102$, very similar to estimates in table 1. ML analysis using this tree suggested 36 sites under positive selection, identifying one new site (3V) and missing one old site (387S). The effect of the tree topology is thus minimal. The differences in the two studies thus appear to be due to the methods of analysis.

4 Discussions

4.1 Inference of Sites Under Positive Selection

The Bayes calculation of posterior probabilities to identify sites under positive selection used estimated parameter values and do not account for the sampling errors in the estimates. The approach is known as Empirical Bayes. An alternative is the hierarchical Bayes approach, which uses a prior distribution of parameters to account for their uncertainties. However, such an approach is computationally much more demanding than the approach adopted here. The posterior distribution might be expected to be more spread out if random errors in parameters were incorporated, suggesting that our method might overestimate the confidence of inference. Furthermore, it should be noted that many inferences are made when all sites in the sequence are assigned to classes. As a result, it is very unlikely for all of them to be correct. Nevertheless, statistical identification of sites under positive selection might be very useful for generating hypotheses for laboratory-based investigation.

Alignment gaps in the variable regions of the protein are removed in the

analysis (table 1). To see the effect of alignment gaps, the data were also analyzed with alignment gaps included and treated as ambiguity data [16]. The sequence then contained 511 codons (1,533 nucleotides). The one-ratio model leads to the estimate $\omega = 0.723$, higher than that in table 1. Similar to results of table 1, all models that allow for sites under positive selection suggest presence of such sites. For example, parameter estimates under M8 (beta&ω) are $p_0 = 0.964$, $p = 0.304$, $q = 0.619$, ($p_1 = 0.036$), $\omega = 4.311$, suggesting about 3.6% of sites under positive selection with $\omega_2 = 4.3$. Many more sites are identified to be under positive selection, including a number of variable sites that were removed in the analysis of table 1. These include most of sites 133D-150E in region V1, sites 395W-413T in V4, and sites 460N-465S in V5. Alignment clearly has a major impact on identification of sites under positive selection, and sites in mis-aligned regions are expected to have a high chance of being identified as under selection. In the HIV-1 *env* gene, however, insertions and deletions are common and appear to play a key role in immune escape and viral variation. It is thus possible that these sites are indeed under positive selection.

The likelihood models assume a constant synonymous rate over sites and only the nonsynonymous rate varies due to selection on the protein. As the selective pressure to maintain the structure of the RNA genome appears to reduce the synonymous rate in certain regions of the gene, the assumption of a constant synonymous rate may be unrealistic. However, selective constraint on the RNA reduces both synonymous and nonsynonymous rates in such regions, so that the rate ratio ω will not be elevated. We thus expect it to be unlikely for the model to generate false positives. Simulations may be needed to find out whether this is the case. Recombination might also affect the analysis, although Yamaguchi-Kabata and Gojobori [4] discussed that it does not appear important for these data. A detailed discussion of the strength and weakness of the ML methods used here is provided by Yang and Bielawski [25].

4.2 Computational Requirement

The ML iteration algorithm cycles through two phases. Phase I optimizes branch lengths, one at a time, with parameters in the substitution model (such as κ and ω). Phase II optimizes substitution parameters, with branch lengths in the tree fixed (Fig. 3). As computation is saved when only one branch length is optimized at one time, this algorithm makes analysis of large data set possible [16]. The old algorithm, which updates all parameters including the branch lengths simultaneously, and calculating first derivatives by the difference approximation, is not feasible for data sets of this size.

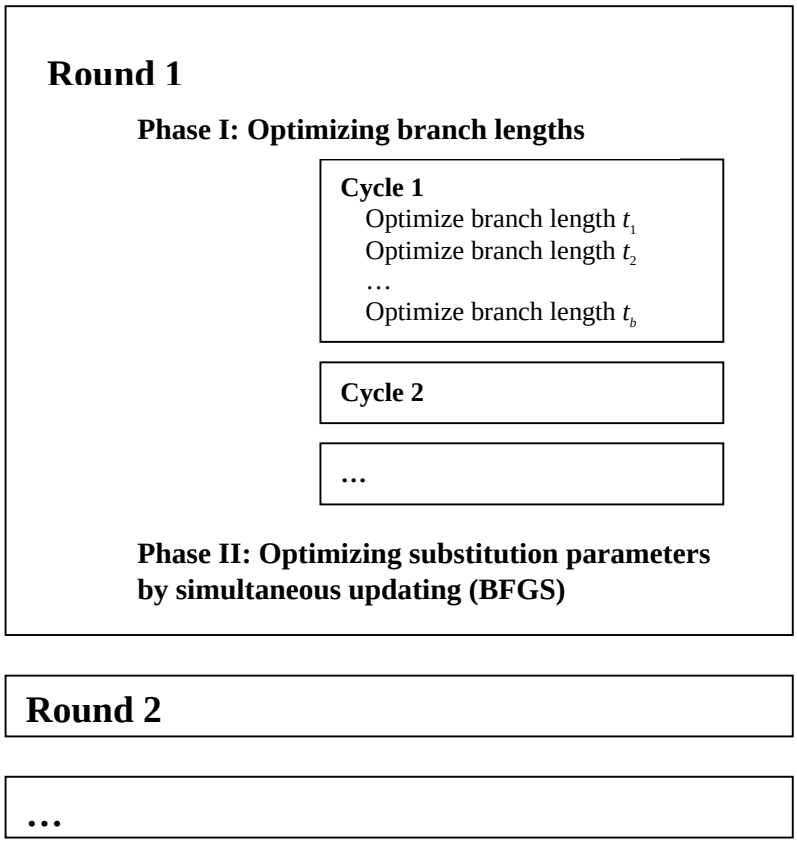

Fig. 3 *Iteration algorithm for obtaining ML parameter estimates (Yang 2000)*

The ML analysis of the present data set is feasible on today's workstations, although nearly 500 Megabytes of memory was required for the computationally most-intensive model M8 (beta&ω). The computational time ranges from several hours for the simple model M0 (one-ratio) to several days for M8 (beta&ω).

5 Acknowledgments

I thank Dr Yumi Yamaguchi-Kabata for providing the sequence data and four anonymous reviewers for comments. This study was supported by grant 31/G10434 from Biotechnology and Biological Sciences Research Council (UK).

6 References

1. S. Bonhoeffer, E.C. Holmes and M.A. Nowak, "Causes of HIV diversity" *Nature* **376**, 125 (1995)
2. D.P. Mindell, "Positive selection and rates of evolution in immunodeficiency viruses from humans and himpanzees" *Proc. Natl. Acad. Sci. U.S.A.* **93**, 3284-3288 (1996)
3. Y. Yamaguchi and T. Gojobori, "Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts" *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1264-1269 (1997)
4. Y. Yamaguchi-Kabata and T. Gojobori, "Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes" *J. Virol.* **74**, 4335-4350 (2000)
5. P. Borrow, *et al.*, "Anti-viral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus" *Nature Med.* **3**, 205-211 (1997)
6. C.C. Wilson, *et al.*, "Frequent detection of escape from cytotoxic T-lymphocyte recognition in perinatal human immunodeficiency virus (HIV) type 1 transmission: the ariel project for the prevention of transmission of HIV from mother to infant" *J. Virol.* **73**, 3975-3985 (1999)
7. K.A. Crandall, *et al.*, "Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection" *Mol. Biol. Evol.* **16**, 372-382 (1999)
8. A.J. Leigh Brown, "Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population" *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1862-1865 (1997)
9. A.I. Agulnik, *et al.*, "Evolution of the DAZ gene family suggests that Y-linked DAZ plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations" *Hum. Mol. Genet.* **7**, 1371-1377 (1998)
10. W.M. Fitch, *et al.*, "Long term trends in the evolution of H(3) HA1 human influenza type A" *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7712-7718 (1997)
11. Y. Suzuki and T. Gojobori, "A method for detecting positive selection at single amino acid sites" *Mol. Biol. Evol.* **16**, 1315–1328 (1999)
12. R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene" *Genetics* **148**, 929-936 (1998)
13. Z. Yang, *et al.*, "Codon-substitution models for heterogeneous selection pressure at amino acid sites" *Genetics* **155**, 431-449 (2000)
14. J.G. Bishop, A.M. Dean and T. Mitchell-Olds, "Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution" *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5322-5327 (2000)
15. P.M. Zanotto, *et al.*, "Genealogical evidence for positive selection in the *nef* gene of HIV-1" *Genetics* **153**, 1077-1089 (1999)
16. Z. Yang, "Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A" *J. Mol. Evol.*in press (2000)
17. M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions" *Mol. Biol. Evol.* **3**, 418-426 (1986)
18. Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models" *Mol. Biol. Evol.* **17**, 32-43 (2000)

19. N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees" *Mol. Biol. Evol.* **4**, 406-425 (1987)
20. J. Felsenstein, *Phylip: Phylogenetic inference program, Version 3*, . University of Washington: Seattle (1993)
21. Z. Yang, *Phylogenetic analysis by maximum likelihood (PAML), Version 3.0 (http://abacus.gene.ucl.ac.uk/software/paml.html)*. London: University College London (2000)
22. J. Bielawski, K. Dunn and Z. Yang, "Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions" *Genetics* in press (2000)
23. S. Modrow, *et al.*, "Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions" *J. Virol.* **61**, 570-578 (1987)
24. I.J. Lauder, *et al.*, "Kernel density analysis of variable and conserved regions of the envelope proteins of human immunodeficiency virus type 1 and associated epitopes" *AIDS Res Hum Retroviruses* **12**, 91-97 (1996)
25. Z. Yang and B. Bielawski, "Statistical methods for detecting molecular adaptation" *Trends Ecol. Evol.*in press (2000)