J Mol Evol (2002) 55:509–521
DOI: 10.1007/s00239-002-2346-9

JOURNAL OF MOLECULAR EVOLUTION

A Sliding Window-Based Method to Detect Selective Constraints in Protein-Coding Genes and Its Application to RNA Viruses

Mario A. Fares,[1] Santiago F. Elena,[1] Javier Ortiz,[2] Andrés Moya,[1] Eladio Barrio[1]

[1] Institut *Cavanilles* de Biodiversitat i Biologia Evolutiva and Departament de Genètica, Universitat de València, València, Spain
[2] Servei de Bioinformàtica, Universitat de València, València, Spain

Received: 30 January 2002 / Accepted: 10 May 2002

Abstract. Here we present a new sliding window-based method specially designed to detect selective constraints in specific regions of a multiple protein-coding sequence alignment. In contrast to previous window-based procedures, our method is based on a nonarbitrary statistical approach to find the appropriate codon-window size to test deviations of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions from the expectation. The probabilities of d_N and d_S are obtained from simulated data and used to detect significant deviations of d_N and d_S in a specific window region of the real sequence alignment. The nonsynonymous-to-synonymous rate ratio ($\omega = d_N/d_S$) was used to highlight selective constraints in any window wherein d_S or d_N was significantly different from the expectation. In these significant windows, ω and its variance [$V(\omega)$] were calculated and used to test the neutral hypothesis. Computer simulations showed that the method is accurate even for highly divergent sequences. The main advantages of the new method are that it (i) uses a statistically appropriate window size to detect different selective patterns, (ii) is computationally less intensive than maximum likelihood methods, and (iii) detects saturation of synonymous sites, which can give deviations from neutrality. Hence, it allows the analysis of highly divergent sequences and the test of different alternative hypothesis as well. The application of the method to different human immunodeficiency virus type 1 and to foot-and-mouth disease virus genes confirms the action of positive selection on previously described regions as well as on new regions.

Key words: FMDV — HIV-1 — Positive selection — Selective constraints — Structural/functional domains — Window-based method

Correspondence to: Mario A. Fares, Institute of Genetics, Department of Genetics, Trinity College, Dublin 2, Ireland; *email:*

Introduction

One of the interests of evolutionary genetics is to unveil the mechanisms of natural selection whereby proteins evolve. Positive (adaptive) and negative (purifying) selection are the two sides of natural selection that can explain the appearance and maintenance of protein functions, respectively. In the light of the neutral theory of molecular evolution, the fixation of a majority of nonsynonymous mutations is explained not by positive selection but by random fixation of neutral or nearly neutral mutations (Kimura 1983). However, evidence supporting the fixation of nonsynonymous mutations by positive selection has been documented in the mammalian major histocompatibility complex (MHC) (Hughes and Nei 1988, 1989), the abalone sperm lysine (Metz and Palumbi 1996), the envelope proteins of HIV-1 (Seibert et al. 1995; Yamaguchi and Gojobori 1997), and the capsid proteins of FMDV (Fares et al. 2001).

selection most of the time (Li 1997), fixation of amino acid replacements by positive selection being rare and relegated to small regions of the molecule. Therefore, different structural and functional domains are likely to be subjected to distinct selective constraints and, thus, to evolve at different rates. Good examples of this have been provided by the analyses of the proinsulin protein (Kimura 1983), thyroid hormone receptors (Green and Chambor 1986), MHC (Hughes and Nei 1988), and paralogous genes coding for interacting polipeptides, such as the α and β subunits of the ATPase complex (Marín et al. 2001).

The study of these protein domains can be performed by taking single codons as units of selection or by averaging rates of nucleotide substitutions along the entire protein-coding gene. However, the analysis of a specific domain or region of the protein-coding sequence represents a better approach to show nucleotide variability between different structural or functional regions and might provide detailed information on the selective constraints driving the evolution of protein-coding genes. When no knowledge about the functional or structural domains of a protein is available, a promising approach to unveil selective constraints is to devise statistical models to measure the intensity of selection acting on these protein domains. In this respect, comparison of nonsynonymous nucleotide substitutions (causing amino acid replacements) to synonymous (silent) changes constitutes an important tool for studying the mechanisms of DNA evolution (Kimura 1983; Gillespie 1991; Ohta 1993). The intensity of natural selection can be measured simply by estimating the nonsynonymous-to-synonymous nucleotide substitution rate ratio ($\omega = d_N/d_S$). Thus, values of $\omega > 1$, $\omega = 1$, and $\omega < 1$ indicate positive selection, neutrality, and purifying selection, respectively. This is the most accepted and stringent way to detect positively selected changes in protein-coding nucleotide sequences (Sharp 1997; Akashi 1999; Crandall et al. 1999).

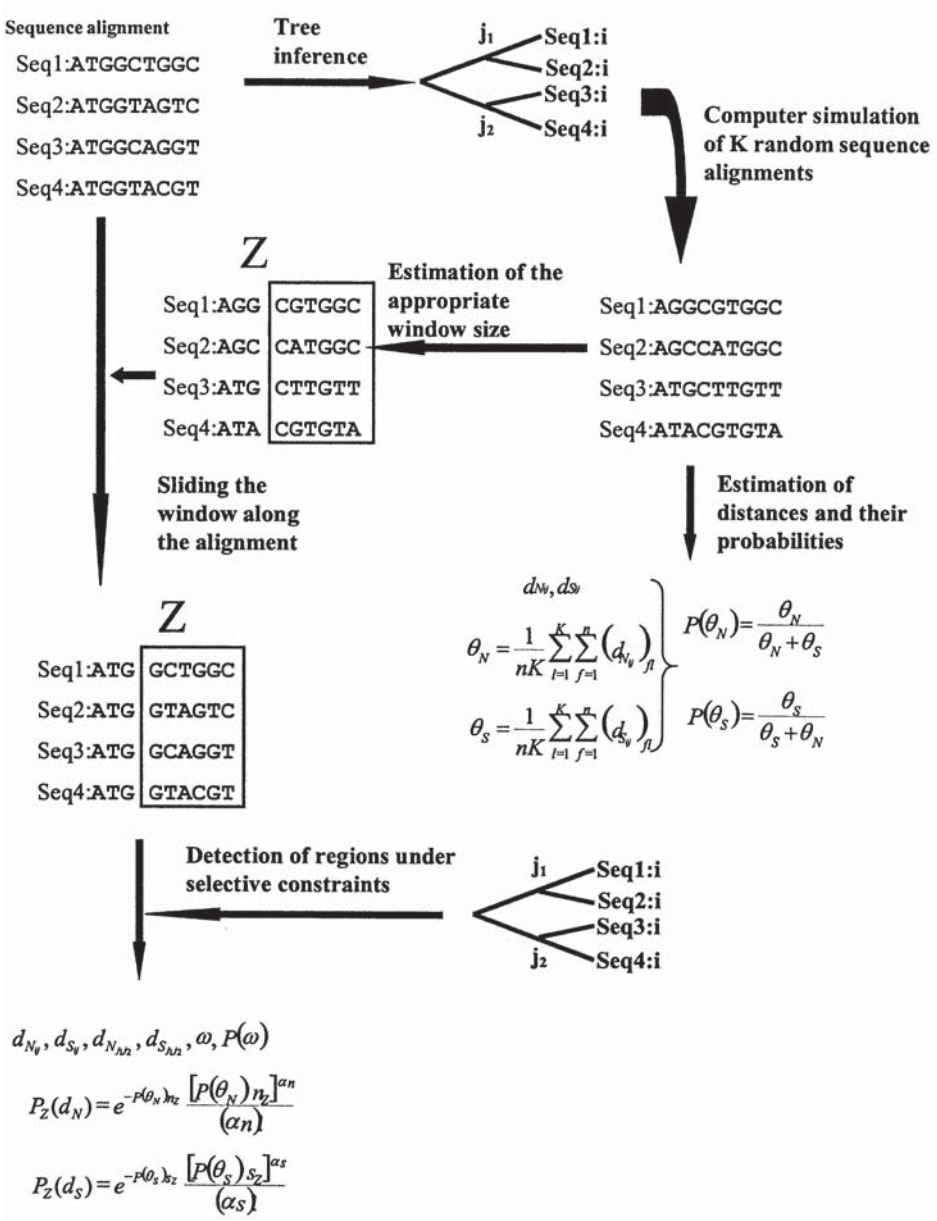
Several methods have been developed to estimate the number of synonymous and nonsynonymous nucleotide substitutions per site (Miyata and Yasunaga 1980; Li et al. 1985; Nei and Gojobori 1986; Li 1993; Pamilo and Bianchi 1993; Goldman and Yang 1994; Muse and Gaut 1994; Comeron 1995; Ina 1995). Nonetheless, these methods require the use of a large number of codons to avoid the effect of variance on estimates of nucleotide distances. Moreover, the likelihood ratio test (LRT) has been extensively applied by several authors to detect positively selected codons under a known phylogeny (e.g., Goldman and Yang 1994; Yang and Nielsen 1998; Yang et al. 2000a; Zanotto et al. 2000). A different approach to show nucleotide substitution rate variation among different genomic regions is to plot difalong a sequence alignment (Tajima 1991). Although, these methods were not originally developed to analyze selective constraints in protein-coding genes, Hughes and Nei (1989) used a sliding window-based method to estimate the differences between the average rate of nonsynonymous and that of synonymous nucleotide substitutions per site and window, detecting selective pressures acting on the MHC. However, these authors did not use any statistical approach to select an appropriate window size to explore protein-coding genes or test the significance of the differences between the two types of nucleotide substitutions, which could lead to misinterpreting results. Although the methods developed by Nielsen and Yang (1998) have been used extensively to detect positive selection, these methods rely on the assumption that all amino acid sites have the same ω value or that there is a limited and specified number of codon categories characterized by their ω values, which seems unrealistic. Suzuki and Gojobori (1999) developed a maximum parsimony method that, given a phylogeny, detects positive selection at single amino acid sites. Here we propose a new sliding window procedure, based on the same principles as those proposed by Suzuki and Gojobori (1999) but using the most appropriate window size to highlight regions within a protein-coding gene subjected to different selective pressures. The central focus of this work is that different protein regions are subjected to different selective constraints that can be underpinned by an appropriate sliding window procedure to detect regions with higher nucleotide substitutions than expected.

Materials and Methods

Our method identifies regions from a protein-coding gene that have accumulated a different number of nucleotide substitutions than expected by chance, hence rejecting neutrality as the best explanation for the evolution of these specific regions or domains. For this purpose, our method implements different steps (Fig. 1) that can be summarized as follows.

Estimation of the Probability of Synonymous and Nonsynonymous Nucleotide Substitutions in Specific Window Regions Along the Sequence Alignment

The first step of the method is the estimation of the expected probability of nonsynonymous, $P(d_N)$, and synonymous, $P(d_S)$, nucleotide substitutions per codon site in each sliding window. To estimate these probabilities, let us first suppose that the Xth nonsynonymous nucleotide substitution and the Yth synonymous nucleotide substitution are variables with probability $1/L$ of occurring in the sequence under study, where L is the total number of codon sites in the sequence. If we assume that nonsynonymous and synonymous substitutions are discrete and take the values x_i and y_i, respectively, then the expectations of X^r- and Y^r- are the rth



$$d_{N_{ij}}, d_{S_{ij}}, d_{N_{Anc}}, d_{S_{Anc}}, \omega, P(\omega)$$

$$P_Z(d_N) = e^{-P(\theta_N)n_Z}\frac{[P(\theta_N)n_Z]^{\alpha n}}{(\alpha n)}$$

$$P_Z(d_S) = e^{-P(\theta_S)s_Z}\frac{[P(\theta_S)s_Z]^{\alpha s}}{(\alpha s)}$$

Fig. 1. Diagram of the sliding window procedure to detect selective constraints in specific regions of protein-coding genes. Z is the window size in codons and n is the total length of the sequence alignment after adding the random sequence pieces to the beginning and end of the sequence alignment.

X_N and synonymous change variable Y_S. These moments are defined as

$$\theta_N^r = E(X_N) = \frac{1}{n}\sum_i x_i^r, \quad \theta_S^r = E(Y_S) = \frac{1}{s}\sum_i y_i^r \qquad (1)$$

where n and s are the total number of nonsynonymous and synonymous nucleotide sites, respectively.

The first moments, $r = 1$, are the expectations of X_N and Y_S and are the mean of the variables that describes nonsynonymous and synonymous nucleotide substitutions on an alignment of sequences, respectively.

θ_S^1 and θ_N^1 are estimated using K random sequence alignments simulated by maximum likelihood according to a known phylogenetic tree using the EVOLVER program from the PAML package, v3.0 (Yang 2000). This program generates a codon and evolves the sequence along the branches of the given phylogeny using specified branch lengths and substitution parameters (Yang et al. 2000b). Simulations were performed using as parameters the branch lengths estimated by the modified Nei and Gojobori method (Zhang et al. 1998) as well as codon frequencies estimated from the real data set. All simulations were carried out assuming neutral evolution ($\omega = 1$), which is the null hypothesis against which the ω value estimated in each window region of the real sequence alignment is compared. Once sequences are simulated, nonsynonymous substitutions per nonsynonymous site ($d_{N_{ij}}$) and synonymous substitutions per synonymous site ($d_{S_{ij}}$), between sequence i and its inferred ancestral sequence j, are estimated by the unbiased method of Li (1993). Given the large distances among the sequences under study, ancestral sequences were inferred by maximum likelihood (Yang et al. 1995; Koshi and Goldstein 1996; Schultz et al. 1996; Zhang and Nei 1997).

enough, both maximum likelihood and maximum parsimony methods give similarly reliable sequence inferences (Yang et al. 1995; Zhang and Nei 1997). The simulations allow us to avoid the effect of the nucleotide compositional bias in third codon positions on the codon usage, the estimation of d_N and d_S under neutrality, and the buffering of the regional codon-composition effect on the estimates of d_N and d_S.

The second step consists in the estimation of the probabilities of nucleotide substitutions as

$$P(\theta_N) = \frac{\theta_N}{\theta_N + \theta_S}, \quad P(\theta_S) = \frac{\theta_S}{\theta_S + \theta_N} \tag{2}$$

where θ_N and θ_S are the mean number of synonymous and nonsynonymous substitutions estimated from the K random sequence alignments using the expressions

$$\theta_N = \frac{1}{NK}\sum_{l=1}^{K}\sum_{f=1}^{N}(d_{N_{ij}})_{fl}, \quad \theta_S = \frac{1}{NK}\sum_{l=1}^{K}\sum_{f=1}^{N}(d_{S_{ij}})_{fl} \tag{3}$$

Here N stands for the total number of pairwise comparisons between the random simulated sequence i and its inferred ancestral sequence j.

If the sequence alignment is large enough (>300 codons) we can assume that both nonsynonymous (d_N) and synonymous (d_S) nucleotide substitutions follow a Poisson distribution with parameters

$$\lambda_N = nP(\theta_N), \quad \lambda_S = sP(\theta_S) \tag{4}$$

Therefore, the probabilities of observing $d_N = \alpha$ and $d_S = \beta$ nucleotide changes in a specific window Z of the real sequence alignment are, respectively,

$$P_Z(X_{N_{ij}^Z}|\lambda_N) = e^{-\lambda_{N_{ij}^Z}}\frac{\lambda_{N_{ij}^Z}^{X_{N_{ij}^Z}}}{X_{N_{ij}^Z}!} \quad \text{and}$$

$$P_Z(Y_{S_{ij}^Z}|\lambda_S) = e^{-\lambda_{S_{ij}^Z}}\frac{\lambda_{S_{ij}^Z}^{Y_{S_{ij}^Z}}}{Y_{S_{ij}^Z}!} \tag{5}$$

where $X_{N_{ij}^Z}$ and $Y_{S_{ij}^Z}$ are the observed number of nonsynonymous and synonymous nucleotide substitutions, respectively, between sequence i and its inferred ancestral sequence j in window Z and are calculated as

$$X_{N_{ij}^Z} = n\alpha, \quad Y_{S_{ij}^Z} = s\beta \tag{6}$$

This procedure of estimating $P_Z(X_{N_{ij}^Z}|\lambda_N)$ and $P_Z(Y_{S_{ij}^Z}|\lambda_S)$ is repeated in every window and the possible action of selection in each region is also tested. Once those window regions with a number of substitutions significantly different than expected are detected, an analogy of selection intensity is performed estimating the nonsynonymous-to-synonymous rate ratio and its variance.

Finding an Appropriate Window Size to Detect Selective Constraints

The sliding window-based method requires a previous estimation of the most appropriate window size to analyze the different regions of the alignment. We have to stress caution when the window size is chosen randomly because there is a strong effect of the codon number and composition of the window size used on the results obtained (see Results and Discussion). Furthermore, the use of a specifically optimized window size strongly depends on the data under analysis. Thus, assuming absence of saturation of synonymous conserved sequences require smaller window sizes to detect selective constraints than more variable sequences do. The estimation of the window size in our method is a trade-off between avoiding false significant results and still getting as much biological information as possible. To achieve this goal, we generate random window sizes (δ_i) in each alignment of the K randomly simulated data sets and slide each ith window of δ_i size along the random sequence alignment. For every window (l) the average numbers of nonsynonymous ($\bar{d}_{N_l}$) and synonymous ($\bar{d}_{S_l}$) nucleotide substitutions for all pairwise sequence comparisons, and given the phylogenetic tree, are estimated. Thereafter, the probability of having d_{N_l} for each lth window is calculated using Eq. (5). By repeating this operation, a probability distribution along the sequence alignment, with mean $\bar{d}_N$, is obtained. Before starting the window-sliding procedure, two sequence alignment pieces of size ($\delta - 1$) are randomized and joined to the beginning and end of the sequence alignment to avoid undercounting the first and last $\delta - 1$ codons in the first and last windows, respectively. Therefore, every codon site is counted δ times in all the sliding steps. The generation of random pieces of the sequence alignment is repeated during the ($\delta - 2$) first and last sliding steps, avoiding nucleotide composition effects of the random pieces on the calculations performed for each window. Thereafter, we obtain the distribution of probabilities of $\bar{d}_{N_l}$ in the K random alignments for the different window sizes randomly chosen. Finally, we plot the mean $P(d_N)$ values (95% confidence interval) for each window size against the window size and choose, as the appropriate window size, the largest window having a 5% lower probability higher than 0.05 (Fig. 3).

Once the appropriate window size is determined, we slide windows of this size along the real data set, in the same way as done for the random data sets, and calculate the probabilities of the estimated $d_{N_{ij_Z}}$ and $d_{S_{ij_Z}}$ in each Zth window to test against chance using Eq. (5), as described in the previous section.

This method allows us, by direct comparison of the expected and observed nucleotide substitutions, to discriminate between different hypotheses that, in addition to neutral evolution, positive selection, or purifying selection, can explain different mutational dynamics (summarized in Table 1).

Nonsynonymous-to-Synonymous Rate Ratio and Its Variance as an Estimator of Selection Intensity

The final complementary step to our method, in those windows with $d_{N_{ij}^Z}$ or $d_{S_{ij}^Z}$ values different than expected by chance, consists in the analysis of the type and intensity of selection by estimating the nonsynonymous-to-synonymous rate ratio ($\omega_Z = d_{N_{ij}^Z}/d_{S_{ij}^Z}$) for comparison of sequence i and its inferred ancestral sequence j. As mentioned in the Introduction, a $\omega \neq 1$ is a good indication of the action of selection. However, to avoid obtaining biased values due to saturation of synonymous sites, especially in regions with multiple hits, those values of $\omega_Z \neq 1$ have to be tested for significance. Consequently, the variance of $\omega[V(\omega)]$ needs to be estimated from the data and used to test the significance of ω against neutrality. An estimator of $V(\omega)$ was obtained by means of Fisher's δ method (Weir 1996):

$$\begin{aligned}\hat{V}\omega = \frac{1}{\bar{d}_S^2}\Bigg\{ & \hat{V}(A_0) + \frac{L_2^2\hat{V}(B_2) + L_0^2\hat{V}(B_0)}{(L_0+L_2)^2} \\ & + \omega^2\left[\hat{V}(B_4) + \frac{L_4^2\hat{V}(A_4) + L_2^2\hat{V}(A_2)}{(L_2+L_4)^2}\right] \\ & + \frac{L_0}{L_0+L_2}\mathrm{Cov}(A_0, B_0) \\ & + \frac{\omega L_2^2}{(L_0+L_2)(L_2+L_4)}\mathrm{Cov}(A_2, B_2) \\ & + \frac{\omega^2 L_4}{\quad}\Bigg\}\end{aligned} \tag{7}$$

Table 1. List of alternative hypotheses that explain values of nonsynonymous-to-synonymous rate ratios (ω), d_N, and d_S, significantly different than expected under the neutral model[a]

	d_N	d_S	Hypothesis accepted
ω > 1	>	>	1
	>	=	1
	>	<	1, 3, 5
	=	>	1, 3
	=	=	1
	=	<	3, 5
	<	<	3, 4
ω = 1	>	>	6
	>	=	6
	=	=	0
	=	<	0, 3
	<	>	4, 6
	<	=	0, 4
	<	<	3, 4
ω < 1	>	>	6
	>	=	2, 6
	>	<	3, 6
	=	>	2, 6
	=	=	2
	=	<	2, 4
	<	>	4, 6
	<	=	2, 4
	<	<	3, 4

[a] A value of d_N or d_S significantly higher or lower than the mean values for the sequence alignments is indicated by > or <, respectively. Hypotheses: 0, neutrality; 1, positive selection; 2, purifying selection; 3, saturation of synonymous sites; 4, saturation of nonsynonymous sites; 5, translational selection; and 6, high mutation rates.

where A_i and B_i are the transition and transversion rates in the ith degenerated site and their variances $[\hat{V}(A_i), \hat{V}(B_i)]$ are given by Eqs. (3) and (4) in Li (1993), L_i is the number of the ith degenerated sites in the region analyzed, and $\mathrm{Cov}(A_i, B_i)$ is the covariance of the transition and transversion rates in the ith degenerated site, given by Eq. (A8) of Ina (1998). It should be noted that this variance is calculated for ω values along the sequence alignment comparing sequence i and its immediate simulated ancestral sequence j, thus there is no need to include covariances of pairwise comparisons in the estimation of $\hat{V}(\omega)$.

To test the reliability of the estimator of $\hat{V}(\omega)$, we simulated 200 alignment data sets with the EVOLVER program, calculated their empirical $V(\omega)$, and compared them with the estimated variance obtained with Eq. (7). Empirical variance of ω was estimated from the simulated data as

$$V(\omega) = \frac{1}{N}\sum_{i=1}^{N}(\omega_i - \bar{\omega})^2 \qquad (8)$$

Here ω_i is the estimation of ω in the ith window and $\bar{\omega}$ is the mean of ω along the sequence alignment.

Data Sets Analyzed by the Sliding Window-Based Method

To examine the usefulness of this method, we used sequences of the *env* and *gag* genes from HIV-1 (Seibert et al. 1995) and the VP1 to VP4 capsid genes from FMDV (Fares et al. 2001), where positive

HIV-1 sequence data from Seibert et al. (1995) were analyzed to test the selective pressure acting on different genomic regions of this retrovirus. These authors examined the action of selection on the *gag* (structural proteins), *env* (envelope region, including both region *gp120* and region *gp41*), and *pol* (polymerase protein) genes and they observed that the pattern of nucleotide substitution in the case of the *env* gene differed significantly from that of either *pol* or *gag*. Furthermore, they showed that, when the *env* gene was used, significantly higher nonsynonymous than synonymous substitutions were observed for sequence families B, C, and E in the V2 region, for sequence families B and C in the V3 region, and for family C in the V4 region. [Each family included those sequences identified by Seibert et al. (1995) as being phylogenetically closely related.] When overall means were computed for the five families, the average d_N value was significantly higher than the d_S value estimated in V2 and V3 from the *gp120* belonging to the *env* gene. In the case of *gp41*, this region showed significantly higher d_S than d_N values for families A–D.

In our study, a subset of the data analyzed by Seibert et al. (1995) of the *env* gene (23 sequences) and *gag* gene (17 sequences) from the five families were used to test the validity of the method. The sequences are identified by their accession numbers in the phylogenetic trees shown in Fig. 2.

Sequences of FMDV from the capsid region (available upon request) with a known experimental history were also used to test the validity of the method. Fares et al. (2001) studied 31 sequences belonging to the capsid region of FMDV. These sequences were obtained from virus isolated under different experimental procedures that emulated several environmental and epidemiological conditions such as genetic drift, massive infections, immune pressure, and persistent infections. This study showed that several amino acid changes were fixed by positive selection in the main antigenic site (A) and in the 5′ end of the variable capsid protein VP3. Here we use the same 31 sequences to confirm the previously obtained results and to test the possible presence of additional regions under selective pressures.

Phylogenetic Reconstruction

Alignments of sequences were obtained with CLUSTAL-X (Thompson et al. 1994) and adjusted manually when necessary. Phylogenetic analysis of the aligned sequences was done using different methods: neighbor-joining (NJ) (Saitou and Nei 1987), maximum likelihood (ML) (Felsenstein 1981), and maximum parsimony (MP) (Fitch 1971). The MEGA v2.01 program (Kumar et al. 2001) was used to obtain NJ trees and the bootstrap support values for 1000 pseudoreplicates. MP and ML trees were reconstructed using DNPARS and DNAML, respectively, from the PHYLIP v3.5 package (Felsenstein 1993).

Results

Positive Selection in HIV-1 Regions Coding for the env *Protein*

Phylogenetic trees for the *env* gene showed the same topology despite the phylogenetic reconstruction method used (Fig. 2A). The $P(\theta_N)$ value estimated from random alignments of the *env* region was 0.497. Analysis of the suitable window size generated the distribution probabilities depicted in Fig. 3. According to this figure, only windows including more than

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.