

Supplementary material

A comparative study of adaptive molecular evolution in different HIV groups and subtypes

Marc Choisy¹, Christopher H. Woelk², Jean-François Guégan¹,
and David L. Robertson^{3*}

¹CEPM, UMR CNRS-IRD 9926, Montpellier, France

²University of California San Diego, Department of Pathology, 9500 Gilman Dr., La Jolla, CA, 92093, USA

³School of Biological Sciences, University of Manchester, Manchester, UK

Running title: Detection and quantification of positively selected sites in HIV gene sequence alignments

*Corresponding author. Mailing address: University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester, M13 9PT. Phone: +44 (0)161 275 5089. Fax: 0161 275 5082.

E-mail: david.robertson@man.ac.uk.

Appendix A. Models M0, M1, M2, M3, M7, and M8.

Yang and coworkers (4) originally proposed 14 models (M0 through M14) for their ML analysis but it became evident from the analysis of biological sequence data that a subset of these models (M0, M1, M2, M3, M7 and M8) was sufficient for detecting positive selection. The M0 (one-ratio) model assumes a single ω for all sites. M1 (neutral) assumes a proportion p_0 of conserved sites with $\omega_0 = 0$ and a proportion $p_1 = 1 - p_0$ of neutral sites with $\omega_1 = 1$. M2 (selection) adds an additional class of sites to M1 ($p_2 = 1 - p_0 - p_1$) for which ω_2 can be estimated from the data. M3 (discrete) estimates ω for a predetermined number of classes (in this case three). Model M7 (beta) uses a discrete beta distribution with ten categories to model different ω ratios (between 0 and 1) among sites. The shape of this beta distribution is governed by the two parameters p and q . Model M8 (beta& ω) adds an additional class of sites to model M7 whereby a proportion of sites (p_1) can have an ω_1 above 1. These models are fully described in the literature (1, 2, 4). M2, M3 and M8 are able to account for positive selection whereas M0, M1 and M7 are not. M0 and M1 are both nested with M2 and M3, M2 is nested with M3, and M7 is nested with M8. Thus the following LRTs were performed in this paper: M0 *vs* M2, M1 *vs* M2, M0 *vs* M3, M1 *vs* M3, M2 *vs* M3 and M7 *vs* M8. Models were implemented using the CODEML program of the PAML package, version 3.1(3).

1. **Goldman, N., and Z. Yang.** 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**:725-736.
2. **Nielsen, R., and Z. Yang.** 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929-936.
3. **Yang, Z. H.** 1997. PAML: a program package for the phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**:555-556.
4. **Yang, Z. H., R. Nielsen, N. Goldman, and A. M. K. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431-449.

Appendix B. Likelihood and parameter estimates for selection analysis.

The first column lists the data sets and the models used. The second and third columns show the log likelihood of the model ($\ln L$) and the average ω ratio (d_N/d_S) respectively. The last column contains the parameter estimates.

Data set/model	$\ln L$	d_N/d_S	Parameter estimates
HIV-1 M A1			
M0	-8834.570	0.551	$\omega = 0.5508$
M1	-8444.831	0.421	$p_0 = 0.57860, p_1 = 0.42140$
M2	-8262.656	0.857	$p_0 = 0.55626, p_1 = 0.35631, p_2 = 0.08743$ $\omega_2 = 5.72956$
M3	-8239.691	0.725	$p_0 = 0.69914, p_1 = 0.22231, p_2 = 0.07854$ $\omega_0 = 0.06691, \omega_1 = 1.18505, \omega_2 = 5.27749$
M7	-8414.049	0.417	$p = 0.05740, q = 0.08560$
M8	-8244.451	0.690	$p = 0.13812, q = 0.34700$ $p_0 = 0.90827, p_1 = 0.09173, \omega_1 = 4.70204$
HIV-1 M B			
M0	-13679.987	0.543	$\omega = 0.5433$
M1	-13164.540	0.514	$p_0 = 0.48620, p_1 = 0.51380$
M2	-12822.132	0.938	$p_0 = 0.46914, p_1 = 0.43930, p_2 = 0.09156$ $\omega_2 = 5.44728$
M3	-12720.594	0.664	$p_0 = 0.70048, p_1 = 0.23205, p_2 = 0.06747$ $\omega_0 = 0.09516, \omega_1 = 1.15949, \omega_2 = 4.86597$
M7	-12979.832	0.323	$p = 0.16979, q = 0.35580$
M8	-12726.897	0.623	$p = 0.22378, q = 0.53990$ $p_0 = 0.91119, p_1 = 0.08881, \omega_1 = 4.00858$
HIV-1 M C			
M0	-13672.725	0.539	$\omega = 0.5389$
M1	-13175.912	0.528	$p_0 = 0.47250, p_1 = 0.52750$
M2	-12781.804	0.991	$p_0 = 0.45177, p_1 = 0.45985, p_2 = 0.08837$ $\omega_2 = 6.01076$
M3	-12655.155	0.694	$p_0 = 0.74515, p_1 = 0.20237, p_2 = 0.05248$ $\omega_0 = 0.12213, \omega_1 = 1.37975, \omega_2 = 6.16796$
M7	-12952.180	0.313	$p = 0.18750, q = 0.41193$
M8	-12668.849	0.610	$p = 0.24681, q = 0.58871$ $p_0 = 0.92450, p_1 = 0.07550, \omega_1 = 4.46262$
HIV-1 M D			
M0	-7977.420	0.462	$\omega = 0.4620$

M1	-7732.625	0.437	$p_0 = 0.56301, p_1 = 0.43699$
M2	-7614.904	0.775	$p_0 = 0.54055, p_1 = 0.39073, p_2 = 0.06872$ $\omega_2 = 5.58635$
M3	-7580.839	0.611	$p_0 = 0.81027, p_1 = 0.16476, p_2 = 0.02497$ $\omega_0 = 0.13383, \omega_1 = 1.84093, \omega_2 = 7.96518$
M7	-7694.235	0.315	$p = 0.14035, q = 0.30567$
M8	-7583.151	0.568	$p = 0.30053, q = 0.91637$ $p_0 = 0.90998, p_1 = 0.09002, \omega_1 = 3.82134$

HIV-1 O

M0	-20702.779	0.494	$\omega = 0.4939$
M1	-19784.103	0.526	$p_0 = 0.47404, p_1 = 0.52596$
M2	-19352.047	0.854	$p_0 = 0.46756, p_1 = 0.45527, p_2 = 0.07717$ $\omega_2 = 5.16582$
M3	-19245.742	0.626	$p_0 = 0.57335, p_1 = 0.34941, p_2 = 0.07724$ $\omega_0 = 0.03775, \omega_1 = 0.83717, \omega_2 = 4.04237$
M7	-19528.141	0.341	$p = 0.15265, q = 0.29462$
M8	-19220.479	0.590	$p = 0.16001, q = 0.32942$ $p_0 = 0.92833, p_1 = 0.07167, \omega_1 = 3.99248$

HIV-2 A1

M0	-14763.981	0.364	$\omega = 0.3644$
M1	-14091.437	0.433	$p_0 = 0.56751, p_1 = 0.43249$
M2	-13882.169	0.676	$p_0 = 0.56062, p_1 = 0.37798, p_2 = 0.06140$ $\omega_2 = 4.84529$
M3	-13768.586	0.463	$p_0 = 0.70010, p_1 = 0.23928, p_2 = 0.06062$ $\omega_0 = 0.04271, \omega_1 = 0.86680, \omega_2 = 3.71720$
M7	-13924.707	0.276	$p = 0.12446, q = 0.32596$
M8	-13765.788	0.444	$p = 0.14980, q = 0.47151$ $p_0 = 0.97065, p_1 = 0.02935, \omega_1 = 3.56379$

Appendix C. Likelihood ratio test (LRTs) between models to test the significance of results obtained through selection analysis.

LRTs are performed by taking twice the difference in log likelihood between two models and comparing the value obtained with a χ^2 distribution (degrees of freedom equal to the difference in the number of parameters between the models). *p*-values in bold indicate comparisons where the null hypothesis (no positive selection) can be rejected in favour of the alternative hypothesis (positive selection) such that the model on the left is rejected in favour of the one on the right.

LRT	M0 vs M2		M1 vs M2		M0 vs M3		M1 vs M3		M2 vs M3		M7 vs M8	
df*	2		2		4		4		2		2	
	χ^2	<i>p</i> -value	χ^2	<i>p</i> -value	χ^2	<i>p</i> -value	χ^2	<i>p</i> -value	χ^2	<i>p</i> -value	χ^2	<i>p</i> -value
HIV-1 M A1	1143.828	<0.001	364.349	<0.001	1189.758	<0.001	410.279	<0.001	45.930	<0.001	339.197	<0.001
HIV-1 M B	1715.710	<0.001	684.818	<0.001	1918.785	<0.001	887.892	<0.001	203.075	<0.001	505.870	<0.001
HIV-1 M C	1781.843	<0.001	788.216	<0.001	2035.141	<0.001	1041.514	<0.001	253.298	<0.001	566.663	<0.001
HIV-1 M D	725.032	<0.001	235.441	<0.001	793.162	<0.001	303.571	<0.001	68.130	<0.001	222.168	<0.001
HIV-1 O	2701.464	<0.001	864.112	<0.001	2914.074	<0.001	1076.722	<0.001	212.610	<0.001	615.324	<0.001
HIV-2 A1	1763.624	<0.001	418.536	<0.001	1990.790	<0.001	645.702	<0.001	227.166	<0.001	317.837	<0.001

*df, degree of freedom between the respective models.

Appendix D. Paired Wilcoxon ranked sum test to determine differences in the strength of positive selection between different HIV data sets. Z refers to the statistic and N refers to the number of sites with posterior probabilities of being in the positively selected class of M8 above the 0.95 level. Significant differences ($P < 0.05$) are indicated in bold. A continuity correction was applied to the normal approximation for the P-values.

HIV data set	HIV-1 M:A	HIV-1 M:B	HIV-1 M:C	HIV-1 M:D	HIV-1 O
HIV-1 M:B	Z =2.1567 N = 6 P = 0.0310				
HIV-1 M:C	Z = -2.772 N = 10 P = 0.0056	Z = -2.1567 N = 6 P = 0.0310			
HIV-1 M:D	Z = -2.772 N = 4 P = 0.1003	Z = -1.6432 N = 4 P = 0.1003	Z = -1.3363 N = 3 P = 0.1814		
HIV-1 O	Z = 1.9799 N = 5 P = 0.0477	Z =2.2222 N = 6 P = 0.0263	Z =2.2222 N = 6 P = 0.0263	Z = 1.3363 N = 3 P = 0.1814	
HIV-2 A	Z = 1.3363 N = 3 P = 0.1814	Z = 0 N = 1 P = 1.000	Z =1.3363 N = 3 P = 0.1814	Z = 0.8944 N = 2 P = 0.3711	Z =0.8944 N = 2 P = 0.3711