W506–W511 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm382

Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach

Adi Stern[1], Adi Doron-Faigenboim[1], Elana Erez[1], Eric Martz[2], Eran Bacharach[1] and Tal Pupko[1,*]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel and [2]Department of Microbiology, University of Massachusetts, Amherst, MA 01003, USA

Received January 31, 2007; Revised April 15, 2007; Accepted April 30, 2007

ABSTRACT

Biologically significant sites in a protein may be identified by contrasting the rates of synonymous (K_s) and non-synonymous (K_a) substitutions. This enables the inference of site-specific positive Darwinian selection and purifying selection. We present here Selecton version 2.2 (http://selecton.bioinfo.tau.ac.il), a web server which automatically calculates the ratio between K_a and K_s (ω) at each site of the protein. This ratio is graphically displayed on each site using a color-coding scheme, indicating either positive selection, purifying selection or lack of selection. Selecton implements an assembly of different evolutionary models, which allow for statistical testing of the hypothesis that a protein has undergone positive selection. Specifically, the recently developed mechanistic-empirical model is introduced, which takes into account the physicochemical properties of amino acids. Advanced options were introduced to allow maximal fine tuning of the server to the user's specific needs, including calculation of statistical support of the ω values, an advanced graphic display of the protein's 3-dimensional structure, use of different genetic codes and inputting of a pre-built phylogenetic tree. Selecton version 2.2 is an effective, user-friendly and freely available web server which implements up-to-date methods for computing site-specific selection forces, and the visualization of these forces on the protein's sequence and structure.

INTRODUCTION

Current protein sequences have been shaped by a prolonged and extensive evolutionary process. Thus, studying a protein's evolutionary history may contribute to the inference of the proteins' properties. Evolutionary conserved sites may be indicative of active sites or of protein–protein interaction domains, while highly variable sites may represent sites subjected to positive Darwinian selection (1,2). Such positively selected sites may be interpreted as being a consequence of molecular adaptation, which confers an evolutionary advantage to the organism. Detecting the level of selection operating on a given protein is enabled by computing ω, the ratio between non-synonymous (K_a) and synonymous (K_s) substitutions. Sites showing ω values significantly higher than one are indicative of positive Darwinian selection, while sites showing ω values significantly lower than one are indicative of purifying selection (2,3).

Selecton was first introduced in 2005 (4), and implemented an evolutionary codon model (5) which enabled calculating ω at each codon site using a maximum-likelihood (ML) approach. With the advent of sequenced genomes, the comparison of DNA sequences in order to infer meaningful information has become a basic procedure for many researchers. The study of the selection forces operating on a protein, and specifically the attempt to identify positive selection in proteins has been on the rise, and with this rise the number of users of Selecton has increased dramatically. This has motivated us to upgrade the server to include most recent state-of-the-art methods for calculating positive selection. We have thus implemented five evolutionary models (6–8) that enable studying the selection forces operating

[*]To whom correspondence should be addressed. Tel: 972-3-640-7693; Fax: 972-3-642-2046; Email: talp@post.tau.ac.il

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors on a protein. Specifically, these models enable testing whether positive selection has operated on the gene under study. This is achieved by comparing between a null model assuming no positive selection, and a model which allows positive selection. We note that positive selection is not a common phenomenon and is not apparent in most examined datasets (9,10). In these cases, Selecton can be used to accurately infer sites undergoing purifying selection. Both positive and purifying selections are calculated by inferring ω at each codon site. All calculations explicitly take into account the phylogenetic relations among the sequences and the underlying stochastic process of evolution. The value of ω at each site is then translated to a discrete color scale, and projected onto one of the homologous sequences specified by the user. If a 3-dimensional (3D) structure of the protein is available, the scores will also be projected onto the Van-der-Waals surface of the protein.

Various other web servers are available for analyzing the evolutionary forces operating on a gene, including servers which perform analyses only on an amino-acid alignment (e.g. 11,12–14), and servers which perform analyses of positive selection in codon-based alignments (15–18). For instance, The SNAP server (15) is based on counting methods which estimate the number of synonymous and non-synonymous substitutions between all pairs of sequences (19,20), and the Data-Monkey server (16) is based on a combination of model hypothesis testing and codon ancestral sequence reconstruction. The advantage of Selecton over these alternative web servers is that it combines the implementation of state-of-the-art methods for detecting positive selection, in a highly user-friendly interface which requires no previous expert knowledge. This enables the vast community of molecular biology researchers to study positive and purifying selection operating on a protein. Furthermore, expert users may use the advanced options to fine-tune the server to their exact requirements.

We provide here a brief review of Selecton, emphasizing the novel features added in version 2.2 of the server. We further present the results of the server when analyzed on the TRIM5α protein, a viral host-defense factor that has recently reached the spotlight of positive selection studies (3,21). We delineate how the use of the Selecton server enables the effortless detection of the primate species-specific retroviral restriction domain of TRIM5α found previously (21,22).

SELECTON VERSION 2 MAJOR INNOVATIONS

Evolutionary models implemented

Several different evolutionary models were implemented in Selecton version 2.2, each emphasizing different biological phenomena:

M7 and M8: These two models were first introduced by Yang *et al.* (6). In brief, the M8 model assumes that ω values come from a mixture of a discrete beta distribution, and an additional category $\omega_s \geqslant 1$ which allows for positive selection. The M8 is the default model for Selecton runs. The M7 model, nested within M8, does not include the additional ω_s category. Since the beta distribution is defined only on the interval [0,1], it thus follows that M7 does not allow for positive selection in the protein.

M8a: This model is a variation on the M8 model. In M8a, the additional category ω_s is set to 1. Thus, this model allows only for purifying and neutral selection (7).

M5: This model assumes a gamma distribution over ω (6).

MEC: This model (8) is the only model here which takes into account the differences between amino-acid replacement rates. In brief, this model expands a 20 by 20 amino-acid replacement rate matrix [such as the commonly used JTT matrix (23)] into a 61 by 61 sense-codon rate matrix. Hence, when the non-synonymous ratio K_a is inferred, the different replacement probabilities between amino acids with distinct properties are taken into account. For instance, all other models in Selecton assume that the evolutionary rates of leucine (UUG) being replaced by either tryptophan (UGG) or phenylalanine (UUU) are equal, since both require one transversion. However, according to the JTT matrix the latter is five times more likely than the former. Thus, under the MEC model, a position with radical replacements will obtain a higher K_a value than a position with more moderate replacements. It should be noted that in the MEC model, ω values are not directly equivalent to the K_a/K_s ratios. ω values are used to calculate these ratios, which are later color-coded onto the results For a more elaborate explanation, refer to the Selecton FAQ section (http://selecton.bioinfo.tau.ac.il/faq.html).

Comparison of these models allows for statistical testing of the hypothesis that there is positive selection operating on the protein (H_1), by contrasting this hypothesis to a null model (H_0). Part of the Selecton output is the likelihood of each model, allowing for comparison using either a likelihood ratio test (LRT) if the models are nested, or by comparing the second order Akaike Information Criterion (AIC_C) (24) scores if they are not. In brief, an LRT consists of comparing twice the log-likelihood difference of both models to a χ^2 table. An alternative approach is to compare the AIC_C scores defined by $-2 \cdot \log L + 2p \cdot (N/N - p - 1)$, where L represents the likelihood of the model given the data, p represents the number of free parameters and N represents the sequence length. The lower the AIC_C score, the better the fit of the model to the data, and hence the model is considered more justified.

We recommend performing one of the following comparisons:

- M8 against M8a (nested models with one degree of freedom). This is the default comparison performed by Selecton.
- M8 against M7 (nested models with two degrees of freedom).
- MEC against M8a (non-nested models: five free parameters and four free parameters, respectively)

For a comparison on the advantages and disadvantages of each model, please refer to the Selecton FAQ section (http://selecton.bioinfo.tau.ac.il/faq.html).

To enable easier use of Selecton, at the end of a run with a model which allows for positive selection, the user may click on a button labeled ‘Test Statistical Significance’. This will run Selecton with the appropriate null model, perform the likelihood comparison (LRT or AIC_c) and output the significance level of this comparison. It should be noted that in the first two nested comparisons, nesting is achieved by fixing one of the parameters on the boundary of the parameter space (6,25), requiring caution when comparing the models with an LRT. Nevertheless, it has been previously shown that using a χ_1^2 to approximate the distribution of the LRT when comparing the M8 versus M8a leads to a conservative approach (25). This approach was adopted by Selecton.

Parameter estimation

Parameters common to all models used in Selecton are codon equilibrium frequencies π_i, the transition transversion ratio κ and the phylogenetic tree branch lengths. π_i are calculated as in (6,26) using the products of the observed nucleotide frequencies (also known as F3X4). κ and branch lengths are all ML estimates. We use an expectation maximization approach (27) to solve the problem of multivariate optimization in the case of branch lengths [a similar approach is described in detail in (28,29)].

An empirical Bayesian method for calculating ω values

The heart of the Selecton server is the calculation of the ω values at each codon position. In the previous version of the server, the ML method (4) was implemented as the sole method of calculation. Recently, we have shown that an empirical Bayesian method can significantly improve the accuracy of inference of conservation scores (30). While the ML method was found to have a relatively high level of false positives, the Bayesian method showed an improved specificity that reduced the level of false positives. The empirical Bayesian method is particularly superior to the ML method when the number of homologous sequences analyzed is small (30). Thus, an empirical Bayesian method of calculating ω values was implemented in the server (6). Following Yang (31), the distributions are approximated using eight discrete categories (the user may define a different number if desired) and the ω values are computed by calculating the expectation of the posterior ω distribution. It should be noted that although more reliable than the ML method, the empirical Bayesian method also suffers from inaccuracy in small data sets, mostly due to sampling errors in the estimation of parameters (such as the distribution shape parameters). Recently two alternatives have been proposed: the full Bayesian estimation (32) and the hierarchical Bayesian estimation (33). However, both alternatives are much more computationally intensive, and hence were not implemented in Selecton.

Reliability of the ω inferences

The reliability of the ω values estimates depends on several factors, including the number of gaps in the MSA, the number of homologous sequences used and their divergence. Thus, an essential improvement implemented in version 2.2 of Selecton is the inclusion of a measure of confidence of the inference. A confidence interval around each ω estimate is defined by the 5th and 95th percentiles of the posterior distribution inferred for each position (see FAQ section of Selecton, http://selecton.bioinfo.tau.ac.il/faq.html). For positions with an inferred $\omega > 1$, if the lower bound of the confidence interval is >1, the inference of positive selection at this position is considered reliable. Selecton furthermore outputs the distribution of the posterior probabilities of ω at each site of the protein.

Visualization of results

The new version of Selecton enables the projection of the ω inferences also onto the primary sequence of the protein. Selecton uses a seven-color scale for representing the different types of selection. Shades of yellow (colors 1 and 2) indicate $\omega > 1$, with dark-yellow standing for sites where reliable positive selection was inferred, and light-yellow standing for positive selection that is not statistically significant. Shades of white through magenta (colors 3 through 7) indicate various level of $\omega \leqslant 1$.

A powerful new 3D visualization tool, FirstGlance in Jmol (FGiJ, http://firstglance.jmol.org), was also implemented in Selecton. FGiJ displays the ω values on the 3D structure of the protein. Its power lies in its easy handling and it needs no installation. It allows preparing presentations with Selecton results snapshots, saving a PDB file containing the Selecton color-coding scheme and allows easy manipulation of different properties of the colored 3D molecule. FGiJ works in all popular web browsers and computer platforms.

Genetic code

Eleven different genetic codes were implemented, including four different nuclear code variants and seven different mitochondrial code variants. This allows the analysis of genes from organisms and organelles which use non-standard genetic codes.

Phylogenetic tree

By default Selecton runs are carried out using phylogenetic trees that the server computes using the neighbor-joining algorithm (34). As input for the neighbor-joining algorithm, pairwise distances are computed applying the ML criterion under a codon model (35) which assumes no selection ($\omega = 1$ for all sites and $\kappa = 2$). We note that in general, more accurate tree topologies lead to better estimate of parameters (36,37). However, it was previously shown that the detection of positive selection is in general robust to tree topology inaccuracies (6,38,39). Hence, the strategy we adopted in Selecton was to avoid the computationally intense search for the precise tree topology, yet to allow the user to provide a pre-computed phylogenetic-tree as an additional input. This new feature enables users to supply a more accurate tree, if available. Additionally, users can supply the tree phylogeny and have the server optimize its branch lengths.

Precision level

The precision level of the computations is defined by setting the cutoff (ε), which defines when two likelihood values have converged. Selecton allows the user to choose between three levels of precision, which also directly affect the speed of calculation: low (ε = 1), intermediate (ε = 0.1) and high (ε = 0.01). The default level of precision for Selecton runs is intermediate.

BIOLOGICAL EXAMPLE

We illustrate the power of Selecton to detect site-specific selection forces by analyzing the evolution of the TRIM5α protein, a protein that has recently been shown to have undergone extensive positive selection during the course of primate evolution (21,22). Furthermore, positively selected regions were found to correlate with the species-specificity determinants of the protein. Here, we wish to exemplify the ease with which Selecton enables detecting the species-specific viral restriction domains of TRIM5α.

Study of TRIM5α

TRIM5 is a member of the large tripartite motif family in primate genomes, characterized by having RING finger, B-box and coiled-coil domains, as well as an additional SPRY domain found in the α isoform (40). TRIM5α was found to account for HIV-1 resistance observed in rhesus cells (41,42). It is not yet known how TRIM5α mediates viral restriction, although a shorter, alternate transcript of the TRIM5 gene has been shown to be a ubiquitin ligase (43). TRIM5α restriction probably acts on the viral capsid (44), although direct physical interaction between TRIM5α and the capsid proteins has not yet been demonstrated.

TRIM5α variants from humans, rhesus monkeys and African green monkeys (AGM) display different but overlapping restriction specificities, which all have the following common property: each TRIM5α is unable to restrict retroviruses isolated from the same species, yet is able to restrict most retroviruses from other species (41). This indicates that TRIM5α is an important natural barrier to cross-species retrovirus transmission.

This type of interaction between a host protein and a parasite protein leads to genetic conflict between the two proteins. Such a conflict may lead to rapid fixation of mutations that alter amino acids at the protein–protein interface, which is the hallmark of positive selection (6). Thus, it has been hypothesized that TRIM5α is in an antagonistic conflict with the retroviral capsid proteins. Sawyer *et al.* (21) analyzed the selection forces acting on TRIM5α and identified a patch of positively selected residues in the SPRY domain. This patch was identified as the species-specific determinant, which is sufficient and necessary for HIV restriction in rhesus monkey cells. Substitution of this patch from the human TRIM5α with the rhesus patch, and vice versa, conferred or abolished HIV-1 restriction, respectively (21). In fact, the region determining the species-specificity of the HIV-1 restriction was eventually mapped to two alternative positions in the rhesus SPRY domains (21). A single arginine to proline replacement at residue 332 of the human TRIM5α, or conversely the exchange of the six residues at positions 335–340 for the eight residues of the rhesus sequence, conferred the human TRIM5α an enhanced ability to restrict HIV-1 (22).

To test the use of Selecton, 20 primate TRIM5α sequences (21) were used as input for the Selecton server. The server was run with the MEC model (log-likelihood = –6716; AIC_C score = 13 442) and compared with the M8a null model (log-likelihood = –6779; AIC_C score = 13 564). Since the AIC_C score of the MEC model is lower, we assume that the MEC model which allows for positives selection indeed fits the TRIM5α data better than a model which does not. The results of the MEC analysis were projected by the server onto the primary sequence of the human TRIM5α (Figure 1). The full results of the run are available in the Gallery section of Selecton (http://selecton.bioinfo.tau.ac.il/gallery.html). The results show an abundance of yellow-colored sites, indicating that TRIM5α has undergone extensive positive selection. Specifically, the two specific determinants conferring HIV-1 species-specific restriction showed exceptionally high levels of positive selection (Figure 1; positions boxed in black), indicating that these sites have undergone excess amino-acid fixations during the course of primate evolution. In fact, the entire SPRY domain (sites 281–493) displays extensive positive selection, as opposed to the RING finger domain (sites 15–59), the B-box domain (sites 90–132) and the coiled-coil domains (sites 130–241), which display mostly purifying selection with some dispersed positively selected sites.

CONCLUSIONS

We describe here Selecton version 2.2, a web-based bioinformatics tool for the identification of site-specific positive selection and purifying selection in a protein. The minimal input for the server consists of a file of homologous coding sequences. The server performs a codon-based alignment of the sequences, calculates ω values at each site, translates these ratios into selection scores and projects them onto the primary or tertiary sequence of the protein, allowing visual identification of blocks or patches of sites with similar ω values. Advanced options of the server include choosing the method of calculation, inputting a phylogenetic tree of the homologous sequences and choosing from amongst a number of evolutionary models implemented in the server. To demonstrate the effectiveness of Selecton, the server was run on a dataset of homologous TRIM5α primate sequences. Selecton correctly identified the species-specific restriction determinants of the protein. Thus, this analysis emphasizes the power of Selecton to accurately identify sites undergoing positive selection, and to present these results in a clear and user-friendly way.

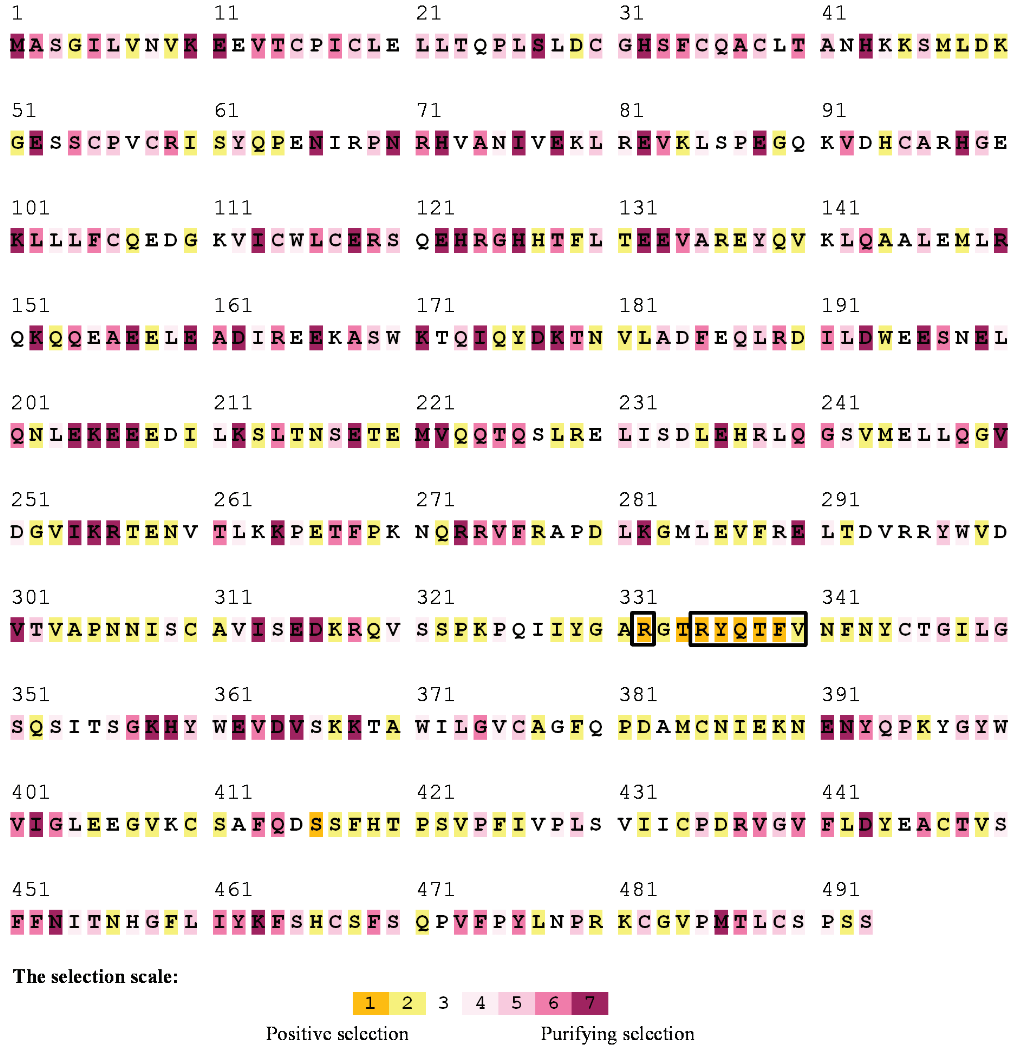
1 MASGILVNVK 11 EEVTCPICLE 21 LLTQPLSLDC 31 GHSFCQACLT 41 ANHKKSMLDK
51 GESSCPVCRI 61 SYQPENIRPN 71 RHVANIVEKL 81 REVKLSPEGQ 91 KVDHCARHGE
101 KLLLFCQEDG 111 KVICWLCERS 121 QEHRGHHTFL 131 TEEVAREYQV 141 KLQAALEMLR
151 QKQQEAEELE 161 ADIREEKASW 171 KTQIQYDKTN 181 VLADFEQLRD 191 ILDWEESNEL
201 QNLEKEEEDI 211 LKSLTNSETE 221 MVQQTQSLRE 231 LISDLEHRLQ 241 GSVMELLQGV
251 DGVIKRTENV 261 TLKKPETFPK 271 NQRRVFRAPD 281 LKGMLEVFRE 291 LTDVRRYWVD
301 VTVAPNNISC 311 AVISEDKRQV 321 SSPKPQIIYG 331 ARGTRYQTFV 341 NFNYCTGILG
351 SQSITSGKHY 361 WEVDVSKKTA 371 WILGVCAGFQ 381 PDAMCNIEKN 391 ENYQPKYGYW
401 VIGLEEGVKC 411 SAFQDSSFHT 421 PSVPFIVPLS 431 VIICPDRVGV 441 FLDYEACTVS
451 FFNITNHGFL 461 IYKFSHCSFS 471 QPVFPYLNPR 481 KCGVPMTLCS 491 PSS
The selection scale:
1 2 3 4 5 6 7
Positive selection
Purifying selection

Figure 1. Selecton results for TRIM5α run on 20 primate sequences (21) with the MEC model (8). Positive selection is colored in shades of yellow, and purifying selection is colored in shades of magenta. The two species-specific restriction determinants are indicated in boxes. Replacement of these positions with their rhesus equivalent positions leads to a reversal of restriction characteristics. Both determinants show a significantly high level of positive selection.

ACKNOWLEDGEMENTS

We would like to thank Clemens Lakner, Fredrik Ronquist and two anonymous referees for their helpful comments. We would also like to thank Nimrod Rubinstein, Itay Mayrose and Eyal Privman for their assistance in proofreading. A.S. is a fellow of the Complexity Science Scholarship program and supported by an Israeli Ministry of Science fellowship. A.D.F. is an Israeli Ministry of Science Eshkol fellow. E.M. was supported by a grant from the US National Science Foundation, Division of Undergraduate Education. T.P. is supported by a grant from the Israel Science Foundation, number 1208/04 and a young GIF grant. T.P. and E.B. are supported by a grant from the Israeli Ministry of Science and Technology. Funding to pay the Open Access publication charges for this article was provided by the Israeli Ministry of Science and Technology.

Conflict of interest statement. None declared.

REFERENCES

1. Graur,D. and Li,W.H. (2000) *Fundamentals of molecular evolution, 2nd edn*. Sinauer Press, Sunderland, MA.
2. Miyata,T. and Yasunaga,T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.*, **16**, 23–36.
3. Yang,Z. (2005) The power of phylogenetic comparison in revealing protein function. *Proc. Natl Acad. Sci. USA*, **102**, 3179–3180.
4. Doron-Faigenboim,A., Stern,A., Mayrose,I., Bacharach,E. and Pupko,T. (2005) Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*, **21**, 2101–2103.
5. Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
6. Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A.M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

7. Swanson,W.J., Nielsen,R. and Yang,Q. (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.*, **20**, 18–20.
8. Doron-Faigenboim,A. and Pupko,T. (2006) A combined empirical and mechanistic codon model. *Mol. Biol. Evol.*, **24**, 388–397.
9. The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
10. Bustamante,C.D., Fledel-Alon,A., Williamson,S., Nielsen,R., Hubisz,M.T., Glanowski,S., Tanenbaum,D.M., White,T.J., Sninsky,J.J. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.
11. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
12. Edwards,R.J. and Shields,D.C. (2005) BADASP: predicting functional specificity in protein families using ancestral sequences. *Bioinformatics*, **21**, 4190–4191.
13. Afonnikov,D.A. and Kolchanov,N.A. (2004) CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res.*, **32**, W64–W68.
14. Gu,X. and Vander Velden,K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics*, **18**, 500–501.
15. Korber,B. (2000) *HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences,* Kluwer Academic Publishers, Dordrecht, The Netherlands.
16. Pond,S.L. and Frost,S.D. (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, **21**, 2531–2533.
17. Liang,H., Zhou,W. and Landweber,L.F. (2006) SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. *Nucleic Acids Res.*, **34**, W382–W384.
18. Phylemon: a suite of web-tools for molecular evolution, phylogenetics and phylogenomics. http://phylemon.bioinfo.cipf.es/cgi-bin/home.cgi. (Last accessed date April 10, 2007)
19. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
20. Ota,T. and Nei,M. (1994) Variance and covariances of the numbers of synonymous and non-synonymous substitutions per site. *Mol. Biol. Evol.*, **11**, 613–619.
21. Sawyer,S.L., Wu,L.I., Emerman,M. and Malik,H.S. (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl Acad. Sci. USA*, **102**, 2832–2837.
22. Yap,M.W., Nisole,S. and Stoye,J.P. (2005) A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction. *Curr. Biol.*, **15**, 73–78.
23. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
24. Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **119**, 716–723.
25. Anisimova,M., Bielawski,J.P. and Yang,Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.*, **18**, 1585–1592.
26. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
27. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. Ser. B.*, **39**, 1–38.
28. Mayrose,I., Friedman,N. and Pupko,T. (2005) A Gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, **21**(Suppl. 2), ii151–ii158.
29. Friedman,N., Ninio,M., Pe'er,I. and Pupko,T. (2002) A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.*, **9**, 331–353.
30. Mayrose,I., Graur,D., Ben-Tal,N. and Pupko,T. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
31. Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.
32. Huelsenbeck,J.P. and Dyer,K.A. (2004) Bayesian estimation of positively selected sites. *J. Mol. Evol.*, **58**, 661–672.
33. Yang,Z., Wong,W.S. and Nielsen,R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**, 1107–1118.
34. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
35. 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.
36. Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**(Suppl. 1), S71–S77.
37. Mayrose,I., Mitchell,A. and Pupko,T. (2005) Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J. Mol. Evol.*, **60**, 345–353.
38. Pie,M.R. (2006) The influence of phylogenetic uncertainty on the detection of positive Darwinian selection. *Mol. Biol. Evol.*, **23**, 2274–2278.
39. Kosakovsky Pond,S.L. and Frost,S.D. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, **22**, 1208–1222.
40. Reymond,A., Meroni,G., Fantozzi,A., Merla,G., Cairo,S., Luzi,L., Riganelli,D., Zanaria,E., Messali,S. *et al.* (2001) The tripartite motif family identifies cell compartments. *EMBO J.*, **20**, 2140–2151.
41. Hatziioannou,T., Perez-Caballero,D., Yang,A., Cowan,S. and Bieniasz,P.D. (2004) Retrovirus resistance factors Ref1 and Lv1 are species-specific variants of TRIM5alpha. *Proc. Natl Acad. Sci. USA*, **101**, 10774–10779.
42. Hatziioannou,T., Cowan,S., Goff,S.P., Bieniasz,P.D. and Towers,G.J. (2003) Restriction of multiple divergent retroviruses by Lv1 and Ref1. *EMBO J.*, **22**, 385–394.
43. Xu,L., Yang,L., Moitra,P.K., Hashimoto,K., Rallabhandi,P., Kaul,S., Meroni,G., Jensen,J.P., Weissman,A.M. *et al.* (2003) BTBD1 and BTBD2 colocalize to cytoplasmic bodies with the RBCC/tripartite motif protein, TRIM5delta. *Exp. Cell Res.*, **288**, 84–93.
44. Stremlau,M., Owens,C.M., Perron,M.J., Kiessling,M., Autissier,P. and Sodroski,J. (2004) The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature*, **427**, 848–853.