BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 9 2005, pages 2101–2103
doi:10.1093/bioinformatics/bti259

Phylogenetics

Selecton: a server for detecting evolutionary forces at a single amino-acid site

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

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

Received on October 27, 2004; revised on December 27, 2004; accepted on December 30, 2004
Advance Access publication January 10, 2005

ABSTRACT

Summary: We present an algorithmic tool for the identification of biologically significant amino acids in proteins of known three dimensional structure. We estimate the degree of purifying selection and positive Darwinian selection at each site and project these estimates onto the molecular surface of the protein. Thus, patches of functional residues (undergoing either positive or purifying selection), which may be discontinuous in the linear sequence, are revealed. We test for the statistical significance of the site-specific scores in order to obtain reliable and valid estimates.

Availability: The Selecton web server is available at: http://selecton.bioinfo.tau.ac.il

Contact: selecton@bioinfo.tau.ac.il

Supplementary information: More information is available at http://selecton.bioinfo.tau.ac.il/overview.html. A set of examples is available at http://selecton.bioinfo.tau.ac.il/gallery.html

MOTIVATION

Detecting biologically significant amino acid sites is critical for understanding protein function and structure. Conserved sites may be indicative of structurally important sites (Melamed *et al*., 2004), active sites (Drory *et al*., 2004), ligand binding sites (Gertow *et al*., 2004) or be a part of protein–protein interaction surfaces (Bridges and Moorhead, 2004). Mutations in these sites may reduce the fitness of their carriers, hence will be selected against, and the carriers are likely to be removed from the population (Graur and Li, 2000). Such sites are called purified sites. Highly variable sites are usually regarded as being tolerant to functional constraints (Glaser *et al*., 2003). However, such sites may be undergoing positive Darwinian selection, conferring evolutionary advantage to the organism. Mutations in these sites have higher probabilities of becoming fixed in the population (Graur and Li, 2000).

Analyzing sequences at the codon level, as opposed to the amino acid level (Glaser *et al*., 2003; Gu and Vander Velden, 2002), enabled us to detect both positively selected and purified selected sites. By contrasting silent (synonymous) substitutions against amino acid altering (non-synonymous) substitutions, it is possible to detect the different selection forces operating on each amino acid site.

Here, we develop a web server (Selecton) that computes synonymous and non-synonymous substitutions from coding DNA sequences. The scores are then projected onto the three-dimensional (3D) structure of the protein. Thus, Selecton enables the detection of 'patches' that are evolutionary meaningful: positive or purified selected amino acid sites that are clustered together in the 3D space.

METHODOLOGY

The ratio of non-synonymous to synonymous substitutions, known as the Ka/Ks ratio, is used to estimate both purifying and positive Darwinian selection (Li, 1993; Li *et al*., 1985; Liberles *et al*., 2001; Miyata and Yasunaga, 1980; Nei and Gojobori, 1986). A Ka/Ks ratio significantly greater than 1 is indicative of positive selection, whereas values significantly smaller than 1 are indicative of purifying selection. Our method calculates the Ka/Ks ratio for each codon site in a codon-based multiple sequence alignment (MSA). In order to achieve maximum accuracy, the algorithm for site-specific Ka/Ks ratios estimation presented here explicitly takes into account the evolutionary relationships among the sequences and the underlying stochastic process. The stochastic process assumed here is a modification of the codon-based evolutionary model suggested by Goldman and Yang (1994), where the maximum likelihood estimates of the Ka/Ks ratios are computed for each site. The significance of the Ka/Ks scores is obtained by using the likelihood ratio test (LRT). This test compares two nested models: a null model which assumes no selection and an alternative model which does. A more detailed description of the methodology is provided at http://selecton.bioinfo.tau.ac.il/overview.html

Available tools that detect conserved sites in proteins such as ConSurf (Glaser *et al*., 2003) or tools aimed at identifying 'rate-shifts' such as DIVERGE (Gu and Vander Velden, 2002), perform their analyses using amino acid sequences only. While this approach is useful when the sequences are highly diverged (Goldman and Yang, 1994), the codon-based model is more sensitive when the sequences are more closely related since synonymous substitutions are explicitly taken into account. More importantly, using the codon-based model enables us to detect both purified and positive selected sites simultaneously. Finally, the LRT is easily incorporated into Ka/Ks calculations.

Selecton accepts as input a set of coding DNA sequences and a protein data bank (PDB) id (Sussman *et al*., 1998). The DNA sequences are translated to amino acids, and are aligned using Clustal W (Thompson *et al*., 1994). The alignment is reverse translated to obtain a codon-based multiple sequence alignment. Ka/Ks scores for each position are then computed based on the algorithm described above and are visualized on the 3D structure using the protein explorer engine (Martz, 2002).

EXAMPLE—HIV-1 PROTEASE

The human immunodeficiency virus type 1 (HIV-1) protease is an essential enzyme for viral replication and thus, is the target for design of drug inhibitors (Flexner, 1998; Peng *et al*., 1989). Specific patterns of drug resistance mutations are associated with each inhibitor

*To whom correspondence should be addressed.
†These authors contributed equally.

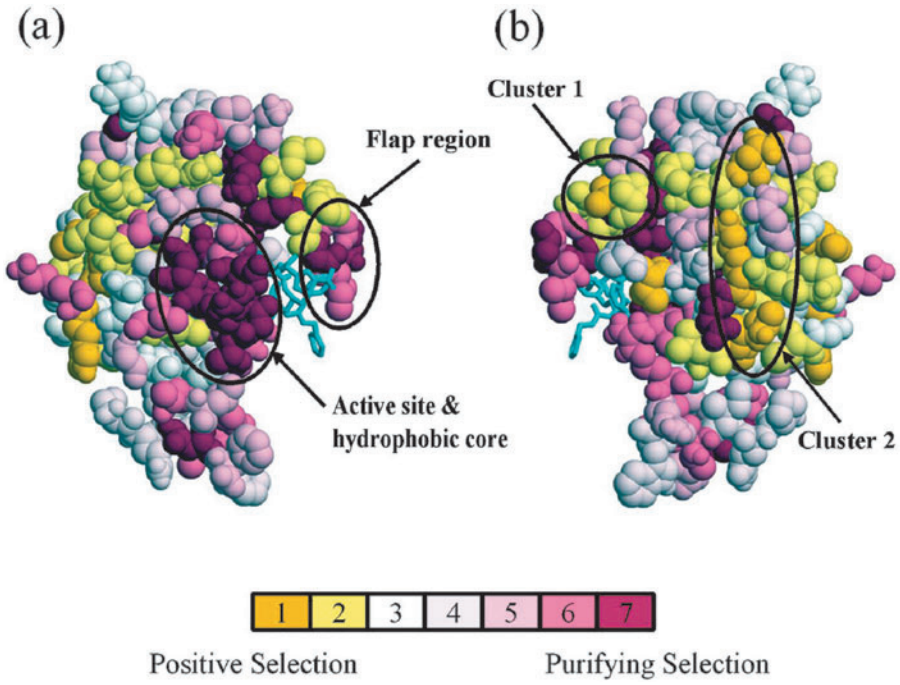

Fig. 1. Selecton results for HIV-1 protease chain A complexed with the inhibitor. The protein is represented as a spacefill model, where the Ka/Ks scores are color-coded onto its Van-der-Vaals surface. The inhibitor (Ritonavir) is shown in light blue as a backbone model. Significant purifying and positive selected sites (p-value < 0.05) are colored in bordeaux (color number 7) and dark yellow (color number 1) respectively. (**a**) View of the active site (residues 22–33), f lap region (residues 47–52) and hydrophobic core (residues 74–87); (**b**) View of two clusters of positively selected sites. Cluster number 1 contains residues Met46, Phe53, Ile54 and Pro79. Cluster number 2 contains residues Leu10, Val11, Thr12, Leu19, Lys20, Met36 and Asn37.

available today. Mutations are divided into two categories: primary resistance mutations, located at the active site, and secondary mutations, remote from the active site. Primary mutations generally cause decreased inhibitor binding, whereas secondary mutations may compensate and restore normal viral replication capacity (Ho *et al*., 1994; Lerma and Heneine, 2001; Molla *et al*., 1996).

Seventy HIV-1 protease gene sequences from patients treated with Ritonavir, a specific protease inhibitor, were extracted from the Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu/). The sequences, together with the PDB structure of the protease dimer, complexed with Ritonavir (Kempf *et al*., 1995), were given as input to the Selecton server and the results were projected on the 3D structure (Fig. 1).

Purifying selection is evident in the three known functional regions identified previously (Loeb *et al*., 1989). These domains include: (1) an active-site loop (residues 22–33) including the Asp–Thr–Gly catalytic triad, characteristic of aspartic proteases, (2) the flap region (residues 47–52) and (3) the hydrophobic core of the molecule (74–87). Selecton successfully identified the three domains as highly conserved (Fig. 1a). The Consurf server was compared with the Selecton server regarding the identification of these domains. As expected, Selecton was found to be more discriminating in the results obtained. Consurf indeed identified the three above-mentioned domains, but also identified an assortment of other residues as being highly conserved. In fact, Consurf identified a total of 53 out of 99 residues of the protein as being highly conserved. Selecton identified 29 conserved residues, the majority of which belong to the functional regions, pointing at the added precision of this novel tool.

Positive selected sites represent sites responsible for the HIV-1 evading the treatment. Positive selection is evident in 29 residues with Ka/Ks > 1. Nine of these residues have been previously reported as conferring drug resistance (Hirsch *et al*., 2000). Residue 82 is the only site known to belong to the primary mutation category and is a part of the hydrophobic core. Indeed, this site was detected by Selecton as undergoing significant positive selection. Apart from site 82, all other previously reported sites belong to the secondary mutation category (Hirsch *et al*., 2000; Lerma and Heneine, 2001).

In addition to known mutations, Selecton identified sites not previously associated with resistance to protease inhibitors as having undergone positive selection. These sites appear as two clusters in Figure 1b. All of these additional sites appear on the surface of the protein. It is likely that these sites also belong to the secondary mutation category. Further experimentation is required to validate the contribution of these sites to viral replication in the presence of Ritonavir.

ACKNOWLEDGEMENTS

We would like to thank Eric Martz for allowing us to use his Protein Explorer program for 3D visualization used in the Selecton server. We thank Nimrod Rubinstein and Ofir Cohen for their helpful comments. T.P. is supported by a grant in Complexity Science from the Yeshaia Horvitz Association and by a grant from the Israel Science Foundation, number 1208/04. Eran Bacharach is supported by a grant from the Israel Science Foundation, number: 731/01-1.

REFERENCES

Bridges,D. and Moorhead,G.B. (2004) 14-3-3 proteins: a number of functions for a numbered protein. *Sci. STKE*, **2004**, re10.

Drory,O., Frolow,F. and Nelson,N. (2004) Crystal structure of yeast V-ATPase subunit C reveals its stator function. *EMBO Rep.*, **5**, 1148–1152.

Flexner,C. (1998) HIV-protease inhibitors. *N. Engl. J. Med.*, **338**, 1281–1292.

Gertow,K., Bellanda,M., Eriksson,P., Boquist,S., Hamsten,A., Sunnerhagen,M. and Fisher,R.M. (2004) Genetic and structural evaluation of fatty acid transport protein-4 in relation to markers of the insulin resistance syndrome. *J. Clin. Endocrinol. Metab.*, **89**, 392–399.

Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.

Graur,D. and Li,W.H. (2000) *Fundamentals of molecular evolution*, 2nd edn. Sinauer Press, Sunderland, MA.

Gu,X. and Vander Velden,K. (2002) DIVERGE: phylogeny-based analysis for Functional–structural divergence of a protein family. *Bioinformatics*, **18**, 500–501.

Hirsch,M.S., Brun-Vezinet,F., D'Aquila,R.T., Hammer,S.M., Johnson,V.A., Kuritzkes,D.R., Loveday,C., Mellors,J.W., Clotet,B., Conway,B. *et al*. (2000) Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society—USA Panel. *JAMA*, **283**, 2417–2426.

Ho,D.D., Toyoshima,T., Mo,H., Kempf,D.J., Norbeck,D., Chen,C.M., Wideburg,N.E., Burt,S.K., Erickson,J.W. and Singh,M.K. (1994) Characterization of human immunodeficiency virus type 1 variants with increased resistance to a C2-symmetric protease inhibitor. *J. Virol.*, **68**, 2016–2020.

Kempf,D.J., Marsh,K.C., Denissen,J.F., McDonald,E., Vasavanonda,S., Flentge,C.A., Green,B.E., Fino,L., Park,C.H., Kong,X.P. *et al*. (1995) ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. *Proc. Natl Acad. Sci. USA*, **92**, 2484–2488.

Lerma,J.G. and Heneine,W. (2001) Resistance of human immunodeficiency virus type 1 to reverse transcriptase and protease inhibitors: genotypic and phenotypic testing. *J. Clin. Virol.*, **21**, 197–212.

Li,W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.

Li,W.H., Wu,C.I. and Luo,C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.

Liberles,D.A., Schreiber,D.R., Govindarajan,S., Chamberlin,S.G. and Benner,S.A. (2001) The adaptive evolution database (TAED). *Genome Biol.*, **2**, Research0028.

Loeb,D.D., Swanstrom,R., Everitt,L., Manchester,M., Stamper,S.E. and Hutchison,C.A., 3rd. (1989) Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.

Martz,E. (2002) Protein explorer: easy yet powerful macromolecular visualization. *Trends Biochem. Sci.*, **27**, 107–109.

Melamed,D., Mark-Danieli,M., Kenan-Eichler,M., Kraus,O., Castiel,A., Laham,N., Pupko,T., Glaser,F., Ben-Tal,N. and Bacharach,E. (2004) The conserved carboxy terminus of the capsid domain of human immunodeficiency virus type 1 gag protein is important for virion assembly and release. *J. Virol.*, **78**, 9675–9688.

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.

Molla,A., Korneyeva,M., Gao,Q., Vasavanonda,S., Schipper,P.J., Mo,H.M., Markowitz,M., Chernyavskiy,T., Niu,P., Lyons,N. *et al*. (1996) Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat. Med.*, **2**, 760–766.

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

Peng,C., Ho,B.K., Chang,T.W. and Chang,N.T. (1989) Role of human immunodeficiency virus type 1-specific protease in core protein maturation and viral infectivity. *J. Virol.*, **63**, 2550–2556.

Sussman,J.L., Lin,D., Jiang,J., Manning,N.O., Prilusky,J., Ritter,O. and Abola,E.E. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. D. Biol. Crystallogr.*, **54**, 1078–1084.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.