BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 5 2003, pages 671–672
DOI: 10.1093/bioinformatics/btg043

TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees

Steve Woolley[1], Justin Johnson[2], Matthew J. Smith[2], Keith A. Crandall[2,3] and David A. McClellan[2,*]

[1]Department of Computer Science, Brigham Young University, Provo, UT 84602, USA, [2]Department of Integrative Biology, Brigham Young University, Provo, UT 84602, USA and [3]Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA

Received on May 7, 2002; revised on August 19, 2002; October 18, 2002; accepted on October 25, 2002

ABSTRACT

Summary: The software program TreeSAAP measures the selective influences on 31 structural and biochemical amino acid properties during cladogenesis, and performs goodness-of-fit and categorical statistical tests.

Availability: The TreeSAAP package (executables for Windows PC or Macintosh OSX, Java source code, documentation, and instruction manual) is available at http://genome.cs.byu.edu/treesaap.htm. UNIX version is available upon request.

Contact: David_McClellan@byu.edu

Approaches for detecting and estimating the influences of selection at the molecular level have included those that rely on the calculation of the global nonsynonymous to synonymous substitution rate ratio, K_a/K_s (e.g. Li, 1993; Pamilo and Bianchi, 1993), methods that compare levels of intra- and interspecific genetic variation (e.g. McDonald and Kreitman, 1991; Sawyer and Hartl, 1992), and maximum likelihood strategies, including both random- and fixed-sites models (e.g. Yang *et al.*, 2000; Yang and Swanson, 2002; Yang and Nielson, 2002). The K_a/K_s approaches assume that if positive selection has significantly influenced the molecular evolution of a protein-coding DNA sequence, $K_a/K_s > 1.0$. This assumption, however, has been found to be too conservative for practical use (e.g. Sharp, 1997) because conservative genes will almost always have more synonymous than nonsynonymous substitutions (Crandall *et al.*, 1999) even if several sites have been influenced by positive selection. Approaches that compare intra- and interspecific variation assume that a greater K_a/K_s ratio between species than within indicates that at least some of the changes between species have preferentially persisted due to selection. These approaches are also limited in scope because they are applicable to only very closely related species at the population level, and thus only can detect molecular adaptation that is recent in derivation (Sharp, 1997). Although more sensitive than the global K_a/K_s techniques, they are inapplicable for most data at the species, genus, and family levels, etc. Maximum likelihood strategies overcome many of these shortcomings because they apply the K_a/K_s approach to individual sites. However, once positive selection is detected, they provide little evidence that suggests the cause of selection, short of inferring causation from domain function only.

In reaction to many of these shortcomings, a few additional statistical models are emerging, including those that incorporate changes in quantitative amino acid properties (Xia and Li, 1998; McClellan and McCracken, 2001). These approaches use calculations of expected random distributions of possible amino acid changes based on fixed differences between residues given a particular physicochemical property. By comparing expected distributions with changes inferred from well corroborated phylogenetic trees, detection of significant deviations from neutral expectations is possible (Figure 1). In order to better implement these models, we have developed TreeSAAP (Selection on Amino Acid Properties using phylogenetic trees), a Java application that implements and automates analysis.

The user communicates with TreeSAAP using a menu-driven interface. DNA sequence input is accomplished with a standard NEXUS format. A TREES block is included to define phylogenetic relationships and infer amino acid substitution events. The user chooses the number of magnitude categories that will be considered and the appropriate genetic code. The user may select from among 31 amino acid properties (see Table 1) and the phylogenetic trees included in the TREES block. The user specifies a model of evolution for ancestral node reconstruction, which is accomplished using the PAML (Yang, 1996) algorithm baseml. TreeSAAP compares sequences in the context of the specified phylogenetic topology,

*To whom correspondence should be addressed.

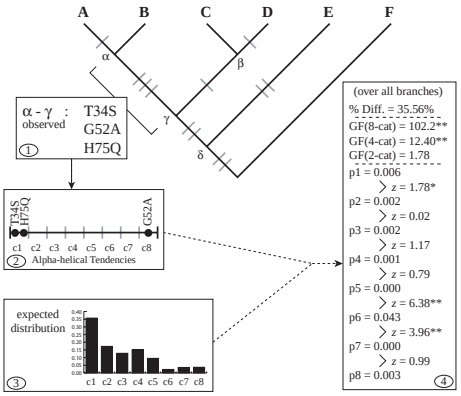

Fig. 1. Algorithmic flow-chart describing the TreeSAAP analyses: (1) nucleotide characters are optimized onto a well-corroborated phylogenetic tree to infer amino acid replacement events; (2) observed amino acid replacements are analyzed in the context of the several physicochemical properties to determine the magnitude classes (c1–c8) of change; (3) the codon compositions of the extant DNA sequences are analyzed in terms of relative frequencies of magnitude classes of evolutionary pathways for each amino acid property; and (4) percent difference between expected and observed mean changes (Xia and Li 1998) and goodness-of-fit (assuming a χ^2-distribution) of observed to expected distributions (McClellan and McCracken, 2001) are calculated for three scales, and the hypothesis of equality between contiguous proportions (p1–p8) of observed substitutions to expected evolutionary pathways is tested (using a normal distribution) for each property (probabilities denoted with * for $p < 0.05$, and ** for $p < 0.01$).

Table 1. Physicochemical amino acid properties available in TreeSAAP for selection analysis (references included in online documentation)

$\alpha-$helical tendencies (P_α)	Molecular weight (M_w)
Average # surrounding residues (N_s)	Normal. consensus hydrophob. (H_{nc})
β-structure tendencies (P_β)	Partial specific volume (V^0)
Bulkiness (B_l)	Polar requirement (P_r)
Buriedness (B_r)	Polarity (p)
Chromatographic index (R_F)	Power to be – C-term., α-helix (α_c)
Coil tendencies (P_c)	Power to be – middle, α-helix (α_m)
Composition (c)	Power to be – N-term., α-helix (α_n)
Compressibility (K^0)	Refractive index (μ)
Equil. Const. – ioniza., COOH (pK')	Sh.- & med.-range n.b. energy (E_{sm})
Helical contact energy (C_a)	Solvent accessible reduct. ratio (R_a)
Hydropathy (h)	Surrounding hydrophobicity (H_p)
Isoelectric point (pH_i)	Thermodyn. transfer hydrophob. (H_t)
Long-range n.b. energy (E_l)	Total n.b. energy (E_t)
Mean r.m.s. fluctuat. displace. (F)	Turn tendencies (P)
Molecular volume (M_v)	

codon by codon, to infer amino acid replacement events.

The inferred pattern of amino acid replacement is then analyzed by using the Xia and Li (1998) and McClellan and McCracken (2001) methods. Both models estimate distributions of potential changes in physicochemical amino acid properties by assuming that every possible amino acid replacement is equally likely under neutral conditions. Expected and observed mean changes in amino acid properties are summarized by the mean percent difference (Xia and Li, 1998). The relative shapes of expected and observed distributions are analyzed two different ways; by goodness-of-fit and by statistically comparing proportions of observed amino acid replacements to expected evolutionary pathways for each contiguous magnitude category (McClellan and McCracken, 2001). Under neutral conditions, the proportions of all contiguous categories are expected to be approximately equal. Generally speaking, low mean percent differences, moderately high goodness-of-fit scores, and significant differences between the most radical magnitude categories correlate with positive selection.

ACKNOWLEDGEMENTS

This project was supported by a fellowship from the Japan Society for the Promotion of Science (D.A.M.), the Office of Research and Creative Activities (S.W.), the National Science Foundation (DEB0120718), and the Pharmaceutical Researchers and Manufacturers of America (PhRMA) Foundation (D.A.M., K.A.C.).

REFERENCES

Crandall,K.A., Kelsey,C.R., Imamichi,H. and Salzman,N.P. (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.*, **16**, 372–382.

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

McClellan,D.A. and McCracken,K.G. (2001) Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. *Mol. Biol. Evol.*, **18**, 917–925.

McDonald,J. and Kreitman,M. (1991) Adaptive protein evolution at the *adh* locus in *Drosophila*. *Nature*, **351**, 652–654.

Pamilo,P. and Bianchi,N.O. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.*, **10**, 271–281.

Sawyer,S.A. and Hartl,D.L. (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.

Sharp,P.M. (1997) In search of molecular Darwinism. *Nature*, **385**, 111–112.

Xia,X. and Li,W.-H. (1998) What amino acid properties affect protein evolution? *J. Mol. Evol.*, **47**, 557–564.

Yang,Z. (1996) Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, **42**, 294–307.

Yang,Z. and Nielson,R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908–917.

Yang,Z., Nielson,R., Goldman,N. and Pedersen,A.-M.K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yang,Z. and Swanson,W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.*, **19**, 49–57.