BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 7 2001
Pages 660–661

ADAPTSITE: detecting natural selection at single amino acid sites

Yoshiyuki Suzuki[1,*], *Takashi Gojobori*[2] *and Masatoshi Nei*[1]

[1]*Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA and* [2]*Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima-shi, Shizuoka-ken 411-8540, Japan*

Received on January 15, 2001; revised on March 9, 2001; accepted on March 15, 2001

ABSTRACT

Summary: ADAPTSITE is a program package for detecting natural selection at single amino acid sites, using a multiple alignment of protein-coding sequences for a given phylogenetic tree. The program infers ancestral codons at all interior nodes, and computes the total numbers of synonymous (c_S) and nonsynonymous (c_N) substitutions as well as the average numbers of synonymous (s_S) and nonsynonymous (s_N) sites for each codon site. The probabilities of occurrence of synonymous and nonsynonymous substitutions are approximated by $s_S/(s_S+s_N)$ and $s_N/(s_S+s_N)$, respectively. The null hypothesis of selective neutrality is tested for each codon site, assuming a binomial distribution for the probability of obtaining c_S and c_N.

Availability: ADAPTSITE is available free of charge at the World-Wide Web sites http://mep.bio.psu.edu/adaptivevol.html and http://www.cib.nig.ac.jp/dda/yossuzuk/welcome.html. The package includes the source code written in C, binary files for UNIX operating systems, manual, and example files.

Contact: yis1@psu.edu

In the study of adaptive evolution of protein molecules, it is important to detect amino acid sites at which natural selection is operating. Since different amino acid sites have different biochemical functions, the type and strength of natural selection may be different at different amino acid sites.

Natural selection may be detected by comparing the rate of nonsynonymous substitutions per nonsynonymous site per unit time (r_N) with that of synonymous substitutions per synonymous site per unit time (r_S). The observation $r_N > r_S$ suggests positive selection, whereas $r_N < r_S$ suggests negative selection (Hughes and Nei, 1988, 1989). For detecting natural selection at single amino acid sites, we have developed a method for comparing r_S and r_N at single codon sites, using a multiple alignment of protein-coding sequences for a given phylogenetic tree (Suzuki, 1999; Suzuki and Gojobori, 1999). Briefly, we apply the following algorithm to each codon site (Figure 1). We infer ancestral codons at all interior nodes of the phylogenetic tree. We then compute the total numbers of synonymous (c_S) and nonsynonymous (c_N) substitutions as well as the average numbers of synonymous (s_S) and nonsynonymous (s_N) sites for the entire phylogenetic tree. Note that c_S, c_N, s_S, and s_N correspond to s_c, n_c, s_t, and n_t, respectively, in Suzuki (1999) and Suzuki and Gojobori (1999). To detect natural selection, we assume a series of nucleotide substitutions as a series of Bernoulli trials, where two possible outcomes are synonymous and nonsynonymous substitutions and they occur with the probabilities of $s_S/(s_S + s_N)$ and $s_N/(s_S + s_N)$, respectively. The number of trials (n) is the sum of the observed numbers for c_S and c_N. Then, the probability ($f(c_N)$) of obtaining c_N ($0 \leqslant c_N \leqslant n$) (and $c_S = n - c_N$) follows a binomial distribution given by

$$f(c_N) = \frac{n!}{(n-c_N)!\cdot c_N!}\cdot\left(\frac{s_S}{s_S+s_N}\right)^{n-c_N}\cdot\left(\frac{s_N}{s_S+s_N}\right)^{c_N}.$$

The null hypothesis of selective neutrality ($c_N/s_N = c_S/s_S$) is tested by computing the sum (p) of all $f(c_N)$s which are equal to or smaller than that for the observed value of c_N (two-tailed test) (Sokal and Rohlf, 1995). If p is less than a given significance level (α) and the relationship $c_N/s_N > c_S/s_S$ is observed, positive selection is indicated, whereas the relationship $c_N/s_N < c_S/s_S$ with $p < \alpha$ suggests negative selection. We can also test the null hypothesis of $c_N/s_N \leqslant c_S/s_S$ by computing the p-value that c_N is equal to or greater than the observed value (one-tailed test). If p is less than α, positive selection is indicated. Similarly, $c_N/s_N \geqslant c_S/s_S$ is tested by computing the p-value that c_N is equal to or smaller than the observed value. The observation $p < \alpha$ suggests negative selection.

ADAPTSITE is a program package implementing the above method and consists of programs ADAPTSITE-P,

*To whom correspondence should be addressed.

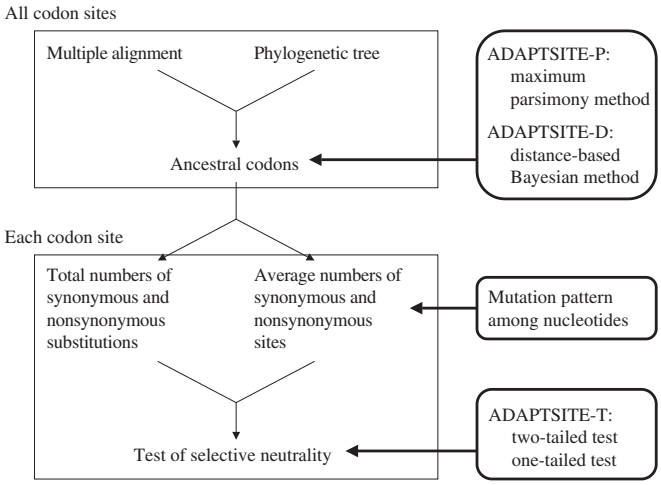

Fig. 1. Schematic diagram of the method for detecting natural selection at single amino acid sites.

ADAPTSITE-D, and ADAPTSITE-T. ADAPTSITE-P and ADAPTSITE-D compute c_S, c_N, s_S, and s_N for each codon site, but the former infers ancestral codons by using the maximum parsimony method (Hartigan, 1973) and the latter by using the distance-based Bayesian method (Zhang and Nei, 1997; Zhang *et al.*, 1998). In both programs, users can specify any mutation pattern among nucleotides for computing s_S and s_N. ADAPTSITE-T reads the output files generated by ADAPTSITE-P and ADAPTSITE-D and computes the p-value for each codon site. The formats of output files from all of these three programs are compatible with StarOffice Calc and Microsoft Excel, so that users can easily make figures and analyze data if necessary.

The performance of the above algorithm has been studied by using computer simulations and real data analyses (Suzuki, 1999; Suzuki and Gojobori, 1999). It has been shown that the false positive rate of detecting natural selection is generally low but the true positive rate increases as the strength of natural selection and the total branch length in the phylogenetic tree increase. For obtaining higher efficiency of detecting natural selection, it is recommended to use many closely related sequences so that the estimates of ancestral codons and c_S, c_N, s_S, and s_N are reliable.

Although the binary files are now available only for UNIX operating systems, we will soon make the binary files for Microsoft Windows operating systems. We will also upgrade these programs periodically to improve the efficiency of detecting natural selection at single amino acid sites.

ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers for providing valuable comments. This work was partially supported by a research grant from the National Institutes of Health to M.N. Y.S. is supported by the JSPS Research Fellowships for Young Scientists.

REFERENCES

Hartigan,J.A. (1973) Minimum mutation fits to a given tree. *Biometrics*, **29**, 53–65.

Hughes,A.L. and Nei,M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Hughes,A.L. and Nei,M. (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl Acad. Sci. USA*, **86**, 958–962.

Sokal,R.R. and Rohlf,F.J. (1995) *Biometry*, 3rd edn, Freeman, New York.

Suzuki,Y. (1999) *Molecular Evolution of Pathogenic Viruses*. PhD Dissertation, Department of Genetics, School of Life Science, The Graduate University for Advanced Studies, Hayama, Japan.

Suzuki,Y. and Gojobori,T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, **16**, 1315–1328.

Zhang,J. and Nei,M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.*, **44**, S139–S146.

Zhang,J., Rosenberg,H.F. and Nei,M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA*, **95**, 7308–7313.