BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 3 2002
Pages 500–501

DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family

Xun Gu and Kent Vander Velden*

Department of Zoology and Genetics, Program of Bioinformatics and Computational Biology, Iowa State University, IA 50011, USA

Received on August 16, 2001; revised on October 15, 2001; accepted on October 22, 2001

ABSTRACT

Summary: DetectIng Variability in Evolutionary Rates among GEnes (DIVERGE) is a software system to study functional divergence of a protein family by detecting site-specific change in evolutionary rate using a multiple alignment of amino acid sequences for a given phylogenetic tree. The program first conducts a statistical test for site-specific rate shifts along the tree, and predicting candidate amino acid residues responsible for functional divergence based on posterior analysis. These results can then be mapped on the 3D protein structure if available.

Availability: DIVERGE is available free of charge from http://xgu1.zool.iastate.edu/. Distribution packages for both Linux and Microsoft Windows operating systems are available, including manual and example files.

Contact: xgu@iastate.edu; kent@iastate.edu

In the study of functional divergence of a protein family it is important to detect amino acid sites that have varying evolutionary conservation among member genes (e.g. Livingstone and Barton, 1996; Gu, 1999). If the function or structure of the protein is changing, some residues may be subject to altered functional constraints during evolution (Landgraf *et al.*, 1999; Dermitzakis and Clark, 2001). This implies that the evolutionary rates at these sites will vary in different homologous genes of a gene family, i.e. type I functional divergence (Gu, 1999).

Site-specific altered functional constraint (or shifted evolutionary rates) can be detected by comparing the rate correlation between gene clusters, when the phylogeny is given (Gu, 1999). In particular, a two-state model is developed to capture the site-specific rate shift during protein family evolution. Consider a phylogeny with two monophyletic clusters generated by gene duplication or speciation. It is proposed that, with probabilities, an amino acid site has two states. In one state (S_0), the site has the same rate in both clusters. In the other state (S_1), the rates in two clusters are so different that over sites they are statistically independent. In each state, the evolutionary rate among sites varies according to the gamma distribution. The coefficient of functional divergence (θ) between two clusters is defined as the probability of a site being S_1, i.e. $\theta = P(S_1)$. Rejection of the null hypothesis $\theta = 0$ suggests that the evolutionary rates (or functional constraints) at some sites have shifted between two gene clusters significantly.

A likelihood ratio test is implemented in DetectIng Variability in Evolutionary Rates among GEnes (DIVERGE) for testing the null hypothesis $\theta = 0$. If the result from the statistical testing is significant, i.e. $\theta > 0$, it is desirable to identify amino acid residues in the protein that have experienced a shift in their functional constraints. These sites are likely to be relevant to the functional–structural basis of the differences between proteins. The posterior probability, denoted by $P(S_1|X)$, of a site being in the S_1 state (i.e. type I functional divergence-related), given the observed amino acid pattern X, is used as an indicator for this purpose.

The software system DIVERGE (Figure 1) follows this two-step procedure of statistical testing and then posterior predictions. The emphasis in DIVERGE has been centered on four main concerns: accuracy of results, ease of use, expandability, and accessibility. It requires the user to input a multiple alignment of amino acid sequences, in either FASTA or CLUSTAL format. Two options are available for the phylogenetic tree: (1) it can be input by the user using the PHYLIP format, standard to several software packages (e.g. PHYLIP, PAUP*, or CLUSTAL); (2) a neighbor-joining tree can be generated and re-rooted by DIVERGE. Gene clusters of interests are selected by simply clicking the internal nodes of the tree. If multiple clusters are selected, DIVERGE performs the statistical analysis (Gu, 1999) for all pairs of clusters, as well as the site-specific profile (posterior analysis) to predict critical amino acid residues for functional divergence. If the user sets a cut-off value, usually >50%, residues with

*To whom correspondence should be addressed.

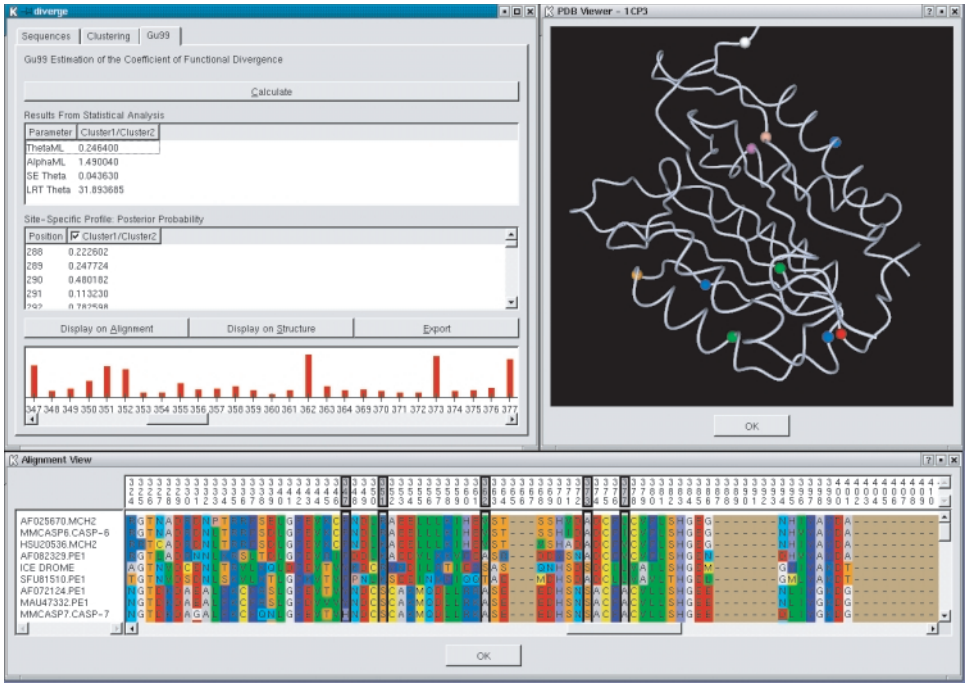

Fig. 1. The main interface of DIVERGE including the statistical results, the site-specific profile, and predicted residues highlighted on the alignment and protein structure of caspase-1. Data from Wang and Gu (2001).

values above the cut-off will be highlighted in the multiple alignment. Then, the user may investigate the underlying mechanism for functional divergence by associating these sites with other biological evidence such as domain, motif, etc. In particular, DIVERGE provides a graphic tool to show the 3D protein structure, provided by a file in PDB format. Thus, predicted amino acid residues by the posterior probability analysis can be highlighted on the protein structure (if available). By allowing interactions with the protein structure, new discoveries about the interrelationships of residues, such as the spatial clustering of those with shifted rates, may be identified.

The performance of the above algorithm has been examined by several case studies (Gu, 1999; Gaucher *et al.*, 2001; Wang and Gu, 2001). To obtain higher efficiency of detecting functional divergence-related residues, the use of a sequence dataset that satisfies the following conditions is recommended: (1) each cluster must have at least four amino acid sequences; (2) except for a large number of sequences, one should be cautious about the result when all pairwise sequence identities are $>90\%$, because of the lack of statistical power; and (3) multiple alignment should be reliable.

The current version provides one of many approaches to predict candidate important residues for functional divergence. We will upgrade our software periodically, (1) to improve the efficiency of detecting single amino acid sites that are involved in functional divergence in protein families, and (2) to improve the 3D structure viewer by including more user-friendly options.

ACKNOWLEDGEMENTS

This work is supported by the NIH grant RO1 GM62118 to X.G. We thank Dr Jianzhi Zhang in programming assistance, and Yuan Lin for website development.

REFERENCES

Dermitzakis,E.T. and Clark,A.G. (2001) Non-neutral diversification after duplication in mammalian developmental genes. *Mol. Biol. Evol.*, **18**, 557–562.

Gaucher,E.A., Miyamoto,M.M. and Benner,S.A. (2001) Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl Acad. Sci. USA*, **98**, 548–552.

Gu,X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.*, **16**, 1664–1674.

Landgraf,R., Fischer,D. and Eisenberg,D. (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.*, **12**, 943–951.

Livingstone,C.D. and Barton,G.J. (1996) Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol.*, **266**, 497–512.

Wang,Y. and Gu,X. (2001) Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**, 1311–1320.