BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 4 2002
Pages 646–647

PATH: a task for the inference of phylogenies

Coral del Val, Peter Ernst, Rüdiger Bräuning, Karl-Heinz Glatting and Sandor Suhai

Department of Molecular Biophysics, DKFZ, INF280, D-69120 Heidelberg, Germany

Received on August 6, 2001; revised on October 29, 2001; accepted on November 26, 2001

ABSTRACT

Summary: Phylogenetic Analysis Task in Husar (PATH) is a task for the inference of phylogenies. It executes three phylogenetic methods and automatically chooses the evolutionary model for each set of data. The output of the tasks shows the consensus trees together with full results obtained from all executed methods.

Availability: PATH is available at the German EMBnet node after registration via www at http://genome.dkfz-heidelberg.de

Contact: genome@dkfz.de

INTRODUCTION

The inference of phylogenies represents a good example of commonly performed analyses that can be implemented into a task. A large number of tools are available which use different phylogenetic methods and different models of evolution distinguished by their parameterizations regarding the average rates of replacements in DNA and protein sequences. The best possible phylogenetic estimates will arise from using the best suited models of evolution along with the use of good quality input data. This is labour intensive and requires previous knowledge of the underlying concepts from the user.

Phylogenetic Analysis Task in Husar (PATH) is implemented under the W3H-Task-System which provides a very flexible way to configure program and data flow among different biocomputational methods (Ernst *et al.*, in preparation). The W3H-Task-System uses Heidelberg Unix Sequence Analysis Resources (HUSAR) (http://genome.dkfz-heidelberg.de) a GCG-based sequence analysis software package. HUSAR is running at the DKFZ (German EMBnet Node).

FEATURES

PATH presents a user-friendly interface with the possibility to start with aligned or unaligned DNA or protein sequences. If the data is not aligned PATH calculates a multiple alignment using *ClustalW* (Thompson *et al.*, 1994). The distance distribution of the sequences is checked using *Distances* from the GCG package (Devereux *et al.*, 1984). Following guidelines from Jin and Nei (1990) the average distance per residue is used as an estimate of the average number of nucleotide substitutions per site (d). The value of d determines which evolutionary model will be used to estimate the pairwise distance tree. Then the quality of the input data is assessed using the splits decomposition method (Bandelt and Dress, 1992) with the program *Splits-Tree* (Hudson, 1998). The splittability index is a measure of the accuracy of the representation of the data set while the isolation index indicates how far apart two data-subsets are. Both indices are checked to give information in the final result about how 'tree-like' the input data is.

At the same time the multiple alignment is used as input for *PUZZLE* (Strimmer and von Haeseler, 1996). This method reconstructs an ML tree for all possible quartets that can be formed from the alignment. The percentage of 'bad quartets,' those with similar best and second best maximum-likelihood values, is used as a measure of the background noise in the data (Strimmer and von Haeseler, 1996).

Then PATH executes the pairwise distance and parsimony method using the PHYLIP package (Felsenstein, 1989). The precision of the phylogenetic inference is indicated by a non-parametric bootstraping using *SeqBoot* (Felsenstein, 1989). The number of multiple datasets and the type of data resampling can be selected by the user. The output file generated by *SeqBoot* is used as input for the estimation of the parsimony and distance analysis. The task configuration will select the appropriate programs for both inference methods, depending on the sequence type (nucleotide or protein) as indicated in Figure 1. *Consense* is then used to construct a majority rule consensus tree for parsimony and distance. Afterwards, the consistency of subgroups is assessed comparing the ML tree, distances and the parsimony consensus tree.

The final output of PATH provides information about the quality of the input data depending on the combination of values obtained for the average distance, splittability and isolation indices from the splits method and the percentage of bad quartets. A summary shows the results from the three phylogenetic methods. Additional information generated in the process, like the multiple alignment, the

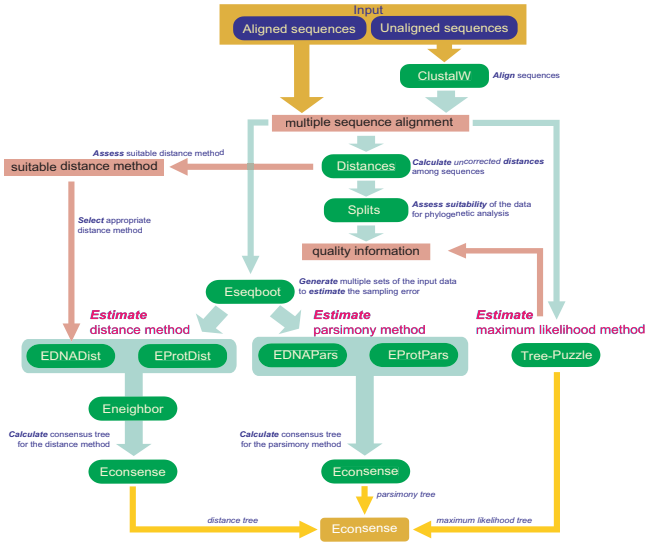

Fig. 1. Program flow in PATH.

splits graph and the distance matrix is also presented. All full text and graphical outputs can be checked using hyperlinks. Generated trees are visualized through a Java tool called *ATV* (Zmasek and Eddy, 2001).

IMPLEMENTATION

PATH has been implemented under the *W3H-Task-System*. This framework allows the integration of applications and methods to create tailor-made analysis task flows, which can be used in high throughput analysis without the usual necessity of customized programming. In such a task system it is necessary to describe the program flow and dependency of applications, the data flow and the merging of the individual outputs (or parts thereof) into a common output report.

The meta-data approach of the *W3H-Task-System* allows the immediate integration of PATH into the W2H web interface (Senger *et al.*, 1998), which is the graphical www interface to HUSAR (http://www.w2h.dkfz-heidelberg.de).

At present PATH has some restrictions in terms of sizes of the submitted jobs, the maximum number of sequences allowed is 200. The running time of the task depends on the different programs and varies between 1 and 4 h on our Sun Multiprocessor machine. In the future, alternative methods for the inference of phylogenies and Likelihood Ratio Tests (LRTs) will be included.

ACKNOWLEDGEMENTS

C.del Val was supported by the German Cancer Research Center (DKFZ) as visiting scientist. We would like to thank Mark van der Linden for the careful reading of the manuscript.

REFERENCES

Bandelt,H.-J. and Dress,A.W.M. (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.*, **1**, 242–252.

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Dress,A.W.M., Huson,D.H. and Moulton,V. (1996) Analyzing and visualizing sequence and distance data using splits-tree. *Discrete Appl. Math.*, **71**, 95–109.

Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics*, **5**, 164–166.

Goldman,N. (1998) Phylogenetic information and experimental design in molecular systematics. *Proc. R. Soc. Lond.* B, **265**, 1779–1786.

Hudson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.

Jin,L. and Nei,M. (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, **7**, 82–102.

Senger,M., Flores,T., Glatting,K., Ernst,P., Hotz-Wagenblatt,A. and Suhai,S. (1998) W2H: www interface to the GCG sequence analysis package. *Bioinformatics*, **14**, 452–457.

Strimmer,K. and von Haeseler,A. (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, **13**, 964–969.

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.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.