BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 12 2004, pages 1977–1979
doi:10.1093/bioinformatics/bth180

LumberJack: a heuristic tool for sequence alignment exploration and phylogenetic inference

Carolyn J. Lawrence[1,†,‡], Christian M. Zmasek[3,†], R. Kelly Dawe[1,2] and Russell L. Malmberg[1,*]

[1]Department of Plant Biology and [2]Department of Genetics, The University of Georgia, Athens, GA 30602, USA and [3]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA

Received on August 7, 2003; revised on March 3, 2004; accepted on March 4, 2004
Advance Access publication March 25, 2004

ABSTRACT

Summary: LumberJack is a phylogenetic tool intended to serve two purposes: to facilitate sampling treespace to find likely tree topologies quickly, and to map phylogenetic signal onto regions of an alignment in a revealing way. LumberJack creates non-random jackknifed alignments by progressively sliding a window of omission along the alignment. A neighbor-joining tree is built from the full alignment and from each jackknifed alignment, and then the likelihood for each topology (given the original full alignment) is calculated. To determine whether any of the topologies generated is significantly more likely than the others, Kishino-Hasegawa, Shimodaira-Hasegawa and ELW tests are implemented.

Availability and Supplementary information: http://www.plantbio.uga.edu/~russell/software.html

Contact: russell@plantbio.uga.edu

BACKGROUND

In a previous analysis, Lawrence *et al*. (2002) reported the phylogeny of the kinesin superfamily using maximum likelihood (ML) methods. Subsequently, we noticed that there was a region of the alignment that contributed a different phylogenetic signal than other regions of the alignment; this suggested the need for the utility LumberJack reported here. We built neighbor-joining (NJ) trees based upon abbreviated (jackknifed) alignments of kinesin motor domains to test whether regions of the kinesin alignment contributed differentially to overall tree topology. Each jackknifed alignment was constructed by sequentially leaving out windows of 50 consecutive positions from Lawrence *et al*.'s (2002) masked progressive alignment. The NJ trees built from these jackknifed alignments were nearly identical at the sequence family level (nodes several steps above the leaves). However, one notable exception occurred within a single clade of the C-terminal I/Kar3 family comprising only the *Caenorhabditis elegans* kinesins CelZ81048, CelU80450 and CelZ66521. This clade only grouped with the other C-terminal kinesins if either of two adjacent 50 amino acid windows was omitted from the masked progressive alignment. Further sequence omissions near these windows revealed that the misleading portion of these three *C.elegans* kinesin sequences lies between positions 372 and 557 in the full progressive alignment (EMBLALIGN_0356). In addition, the tree topology generated by omitting that region closely approximated the most likely tree (in a statistical sense) obtained using more sophisticated ML methods (Lawrence *et al*., 2002). Our results suggested that small regions of phylogenetically misleading sequence may be present in other sequence alignments, and that constructing phylogenies with partial alignments might be an effective means of sampling to identify such regions.

In this report, we describe a semi-automatic utility designed to make this jackknifing method more generally available for use by other researchers to address their alignment-treebuilding problems.

APPLICATION DEVELOPMENT

Our aim was to design an application that could: (1) identify regions of an alignment that can mislead NJ treebuilding (as was the case for our kinesin dataset) and (2) define reasonably likely topologies for large datasets consisting of nucleotide or amino acid sequences in a short amount of time relative to likelihood-based methods such as ML star decomposition (Saitou, 1990; Adachi and Hasegawa, 1992) or ML quartet puzzling (Strimmer and von Haeseler, 1996). Since this approach generates multiple trees, we include a

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[‡]Present address: Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA.

calculation of tree likelihood scores (Felsenstein, 1981) as a basis for the user to compare the output trees with each other.

LumberJack creates non-random jackknifed alignments by omitting regions of sequence from the alignment. Given an alignment input file, the size of the window to be omitted, the amount windows are to overlap, and a substitution model, LumberJack generates all the jackknifed alignments and then calculates pairwise ML distances (based on the substitution model selected by the user). Next, NJ trees are constructed for each of the jackknifed alignments and for the original, unabridged alignment. Subsequently, the likelihood is calculated for each topology generated (based upon the original, unabridged alignment and the chosen substitution model), and the likelihoods of all topologies generated are compared using expected likelihood weight confidence values, two versions of the Kishino-Hasegawa test and the Shimodaira-Hasegawa test (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1989; Shimodaira and Hasegawa, 1999; reviewed in Strimmer and Rambaut, 2001).

In coding LumberJack, we used PAL (a Java library for molecular evolution and phylogenetics) extensively for alignment analyses and treebuilding (Drummer and Strimmer, 2001) and incorporated TREE-PUZZLE 5.1 for likelihood score calculations and comparisons (Strimmer *et al*., 2000). Using TREE-PUZZLE (a C application) makes this application faster but less portable than were it to have been coded entirely in Java. We found that using Java to calculate likelihoods was unreasonably time consuming for amino acid sequence alignments. We also incorporated ATV (Zmasek and Eddy, 2001) into LumberJack to facilitate visualization of trees generated. The LumberJack application can be run directly from the command line or in an interactive mode.

DISCUSSION

LumberJack can identify regions with potentially misleading phylogenetic information or regions with high levels of phylogenetic signal by mapping them onto the underlying alignment. The most likely trees built are based upon alignments wherein phylogenetically misleading regions were omitted, and least likely trees are based upon alignments missing regions of potentially high phylogenetic signal. For many researchers, LumberJack may be a starting place for evaluating information along an alignment. Subsequent to carrying out the LumberJack analysis, one might further investigate the information content of such regions by comparing the regions' substitution rates to other portions of the alignment or by evaluating entropy values along the alignment.

Because the biological processes that result in misleading phylogenetic signal are great in number (e.g. hybridization, recombination, gene conversion, paralog sampling, etc.), many methods have been developed to detect the past occurrence of such events (see http://evolution.genetics.washington.edu/phylip/software.html for Felsenstein's curated list of programs that implement some such methods). For instance, various statistical methods and software packages (e.g. PLATO, TOPAL, LARD, etc.) have been developed to evaluate whether any region of an alignment supports the hypothesis that interspecific recombination has occurred among the DNA sequences represented in a given alignment [reviewed in Husmeier and Wright (2001)]. LumberJack differs from such methods in that it does not test any particular hypothesis *per se*, and could be used to identify misleading regions regardless of how these regions came into existence. After having used LumberJack to confirm the presence of such regions within an alignment, applications (such as those listed above) could be used to determine which evolutionary process might underlie the observation of differential phylogenetic signal along the alignment.

Researchers interested in generating a reasonable phylogenetic tree quickly from a large dataset also could use LumberJack. Since a distribution of trees is generated and compared on the basis of likelihoods, LumberJack can be used as a tool for sampling treespace to find reasonable trees quickly. Other methods used to sample treespace quickly include bootstrapping (Felsenstein, 1985) and jackknifing a given number of sites from the alignment at random, but these methods cannot be used to reveal interesting regions of the alignment simultaneously. It should be noted that recombination (and other events) can cause the tree-like structure used to depict evolutionary history to breakdown, yielding a reticulate, web-like network that more accurately conveys the sequences' relatedness (reviewed in Legendre and Makarenkov, 2002). In such instances, any method that yields results representing the history as a bifurcating tree would be inappropriate for determining the phylogenetic relatedness of a set of sequences, including the one presented here.

LumberJack is under continual development. Current efforts to improve LumberJack include creation of a web server to run LumberJack online, development of a user-friendly GUI and increased documentation. It has been suggested that LumberJack be modified to query protein domain databases, enabling users to test the hypothesis that regions containing divergent phylogenetic signal could correspond to conserved domains. Modifications like this one will increase the utility of the LumberJack package, and such features will be added for future version releases.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their critical reviews of this work. This research was supported by an Energy Biosciences grant (DEFG02-97ER286) from the Department of Energy to R.L.M. Additional support was provided to C.J.L. by a National Science Foundation interdisciplinary research training grant (BIR20329).

REFERENCES

Adachi,J. and Hasegawa,M. (1992) MOLPHY: programs for molecular phylogenetics. I.—PROTML: maximum likelihood inference of protein phylogeny. In *Computer Science Monograph*. The Institute of Statistical Mathematics, Tokyo.

Drummer,A. and Strimmer,K. (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, **17**, 662–663.

Felsenstein,J. (1981) Evolutionary trees form DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Felsenstein,J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Hasegawa,M. and Kishino,H. (1989) Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution*, **43**, 672–677.

Husmeier,D. and Wright,F. (2001) Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics*, **1**, 1–8.

Kishino,H. and Hasegawa,M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.*, **29**, 170–179.

Lawrence,C.J., Malmberg,R.L., Muszynski,M.G. and Dawe,R.K. (2002) Maximum likelihood methods reveal conservation of function among closely related kinesin families. *J. Mol. Evol.*, **54**, 42–53.

Legendre,P. and Makarenkov,V. (2002) Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.*, **51**, 199–216.

Saitou,N. (1990) Maximum likelihood methods. *Methods Enzymol.*, **183**, 584–598.

Shimodaira,H. and Hasegawa,M. (1999) Multiple comparison of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.*, **16**, 1114–1116.

Strimmer,K. and Rambaut,A. (2001) Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. Lond.*, **269**, 137–142.

Strimmer,K., Schmidt,H.A., Vingron,M. and von Haeseler,A. (2000) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

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

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