LIKEWIND

Users Manual

A maximum likelihood-based method for detecting conflicting phylogenetic signal in nucleotide sequence alignments

John M. Archibald and Andrew J. Roger

Canadian Institute for Advanced Research, Program in Evolutionary Biology,
Department of Biochemistry and Molecular Biology, Dalhousie University,
Halifax, Nova Scotia, B3H 4H7, Canada,

email: aroger@is.dal.ca.
tel: (902) 494 2620, FAX: (902) 494 1355,

Introduction

Recombination is well known as a complicating factor in the interpretation of molecular phylogenies. Currently, a wide variety of methods are available for detecting recombination in nucleotide sequence alignments. Most of these methods work best at identifying recombination events between closely related genes. We have developed a maximum likelihood (ML) method that has proven useful for detecting recombination between more divergent sequences. Our method is based on the likelihood ratio test and uses a sliding window approach to identify regions in an alignment that are inconsistent with an *a priori* phylogenetic hypothesis (typically the ML tree inferred from the complete alignment). We have written a set of three Perl programs (the LIKEWIND package) that allow this method to be implemented.

What do you need to have to run LIKEWIND?

The LIKEWIND package consists of 2 Perl utilities that build 'nexus blocks' for execution in PAUP* (Swofford 1998, see

http://www.sinauer.com/Titles/Text/swofford.html) with its ML likelihood options – these Perl utilities are: *likewind.pl* and *simblock.pl*. A 3rd post-processing utility (*getlikes.p*l) will generate your likelihood difference profiles from the real or simulated data as a column of numbers that can easily be plotted with any graphing tool. To successfully complete LIKEWIND analyses you will need to have installed the following programs or utilities on your computers:

PAUP* (preferably running in a Unix environment, although this is not required)

Perl version 5.0 or later (running on a Unix/Linux environment) (http://www.perl.com)

Seq-Gen version 1.2.4 or later (http://evolve.zoo.ox.ac.uk/) running in the Unix/Linux environment

For ease of use, it is best to have all of these programs in your 'path' (i.e. they must be executable from your user directories). If this terminology confuses you, then consult your system administrator to verify whether this is the case.

You must also download: *likewind.pl*, *simblock.pl* and *getlikes.pl* and either run them in your 'working' directory or have them installed in the normal place for programs on your computer, then make sure they are specified in your 'path'.

likewind.pl

Likewind.pl generates an executable 'nexus' file that allows the user to perform an ML sliding window analysis with PAUP* (Swofford 1998) to identify regions of the alignment that are inconsistent with a phylogenetic hypothesis calculated *a priori*. By executing the program (by typing '*likewind.pl*'at the prompt), you will be prompted to provide information on the command line regarding options or filenames of files that are in your current working directory. The user is first asked to specify the total number of characters in the alignment, and whether a pre-existing user-defined tree is to be imposed on each window spanning the alignment or whether the *a priori* phylogenetic hypothesis is to be inferred from the data. Next, the user has the option to enforce user-defined branch-lengths on the likelihood estimates for each window, and, if analyzing protein-coding genes, to calculate the *a priori* phylogenetic hypothesis using only the first and second positions as well as all sites in the alignment. Finally, the user specifies the size of the sliding window as well as the increment with which the window is to be moved along the alignment. We have found that a window size of 100 nucleotides and an increment of 10 nucleotides work well, and are thus the default settings. Once the 'nexus' output file is generated you can execute your nexus formated alignment in PAUP* (the dataset you want to test), then execute

the 'nexus' file that *likewind.pl* created. Execution of the block will infer ML trees using a general time reversible plus Γ plus invariable sites model (GTR+Γ+P_{INV}), with the GTR rate matrix values, the Γ shape parameter α and the proportion of invariable sites estimated from the data. Sliding window analyses can also be performed with a ML-distance approximation (see below).

getlikes.pl

To perform the sliding window analysis, the *likewind.pl* outfile is run as a 'commands block' in PAUP*, after the nexus-formatted alignment file has been executed. If no user-defined tree has been specified, the *a priori* phylogenetic hypothesis—the ML tree inferred from the complete alignment—is determined and the tree, with branch-lengths included, is written to a file. For each window spanning the alignment, the log likelihood of the best tree from a heuristic ML search is then obtained, as is the likelihood of the data in the window given the *a priori* phylogenetic hypothesis (imposed with or without branch-lengths). This is done for each window spanning the alignment, and the likelihoods are written and appended to files with root names *besttre.scores* and *pos123.scores*, respectively. For each window, the difference between the two lnLs (ΔlnL) represents the degree of conflict between the data in the window and the *a priori* phylogenetic hypothesis. The Perl script *getlikes.pl* calculates the ΔlnL's and writes them to a file called *allbest.deltalnL* and, if different codon sets were specified, a file called *pos12best.deltalnL*. These files contain a complete set of ΔlnL values, one for each of the sliding windows spanning the alignment. These can be plotted using a spreadsheet package (e.g., Microsoft Excel) to visualize trends.

simblock.pl

The ΔlnL profile obtained above may identify regions of your alignment that possess phylogenetic signal that is inconsistent with the *a priori* phylogenetic hypothesis. To determine whether these regions differ significantly from a null hypothesis of no recombination, a third Perl script, *simblock.pl*, generates the datasets necessary to perform a parametric bootstrapping analysis. *Simblock.pl* uses Rambaut and Grassly's Seq-Gen version 1.2.4 (1997) to simulate nucleotide sequence evolution over the user-defined tree used to obtain the ΔlnL profile. The user specifies the rate matrix values of the GTR model as well as the estimated base frequencies, the shape parameter α and the proportion of invariable sites (P_{INV}). Each of the simulated datasets is then written to a single file, and a sliding window analysis is performed on each. The program *getlikes.pl* (above) is used to obtain the ΔlnL values for each simulated dataset, and the largest ΔlnL value from each is taken to form a distribution of extreme ΔlnL values under the null hypothesis of no recombination with which empirical values can be compared. For 'n' simulated datasets, ordering the extreme ΔlnLs

from lowest to highest, the i^{th} ΔlnL value defines the $(i-0.5)/n^{th}$ quantile of the null distribution. If , for instance, this quantile = 0.99, then ΔlnL values from the observed data greater than this value are considered significant at $p<0.01$.

An ML-distance approximation

The use of ML to infer phylogenetic trees is computationally prohibitive for large datasets. We have found that performing the sliding window analysis using a ML-distance approximation (i.e., inferring likelihoods from ML-distance trees) produces ΔlnL profiles similar to those inferred using full ML. As expected, this approach is much faster, making it suitable for the analysis of alignments containing a large number of sequences. As a cautionary note, we have found that ΔlnL profiles inferred with the ML-distance option occasionally contain negative ΔlnL values, which correspond to regions of the alignment where the ML and ML-distance methods produce different tree topologies. Inferences about such regions should thus be made with care. As well, the latest version of PAUP* (4.0b8) contains a bug that prevents the ML-distance option from being used with some (but not all) datasets. This problem is currently under investigation.

Problems and Bugs

If you encounter problems with these programs that are not addressed above please contact Andrew Roger at: aroger@is.dal.ca

Citing LIKEWIND

This procedure has been described in detail and tested in:

Archibald, J.M. & Roger, A.J. A maximum likelihood-based method for detecting conflicting phylogenetic signal in nucleotide sequence alignments. ***Submitted***

Until this work is published, users should contact the authors to obtain permission to cite the method or to obtain the most up-to-date citation information.

References:

Grassly, N. C. & Holmes, E. C. (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**, 239-247.

Rambaut, A. & Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**, 235-238.

Swofford, D. L. (1998). PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Massachusetts.