LVB: parsimony and simulated annealing in the search for phylogenetic trees

Supplementary Information for the paper:

Barker, D., 2004. LVB: parsimony and simulated annealing in the search for phylogenetic trees. *Bioinformatics*, **20**, 274–275.

Why use a heuristic?

The average size of phylogenetic analyses is increasing. Research is being performed on relationships among large numbers of species and large gene families. However, many of the popular methods of evaluating trees, including parsimony (Fitch 1971) and maximum likelihood (Felsenstein 1981), are so computationally intensive that an exact solution is not possible for more than about 20 sequences.

For larger numbers of sequences, heuristic methods may give a result similar to the exact solution, but in much shorter time. Heuristics used in phylogenetic inferrence include a greedy algorithm (stepwise addition, Swofford 1993), hill climbing (Croes 1958), simulated annealing (Kirkpatrick et al. 1983, Lundy 1985, Salter and Pearl 2001) and genetic algorithms (Holland 1975, Lewis 1998).

Why use simulated annealing?

A popular heuristic for seeking parsimonious trees is stepwise addition followed by hill climbing. This combination is implemented by, for example, DNAPARS in the PHYLIP package (Felsenstein 1989) and PAUP (Swofford 1993, 2002). However, as the 'search space' becomes more complicated, many random re-starts are required, since each search may become 'stuck' in a local optimum.

Simulated annealing is able to 'jump out' of local optima in a way that stepwise addition and hill climbing cannot, by occasionally accepting a worse tree during the course of the search. One or a small number of simulated annealing searches may be more economical than stepwise addition and hill climbing with many re-starts.

Implementation

LVB 2.0 is written in ANSI C. Internal documentation is provided by structured comments in POD format, which may be extracted to HTML files by `pod2html` (e.g., by using the supplied `Makefile`). Data matrix input code is re-used from PHYLIP <http://evolution.genetics.washington.edu/phylip.html>. LVB's source distribution includes documentation of some PHYLIP internals. LVB has an automated test suite, which uses Perl to launch LVB and collate results. The test suite is expected to grow

over time. A GUI for LVB 1.0, written in TCL/Tk by Mark Ludewig and Andreas Basios, is being updated for LVB 2.0.

Simulated annealing algorithm

The cooling cycle is controlled by several parameters. To simplify use, LVB 2.0 asks the user to chose between a 'fast' and a 'slow' analysis. Each of these settings uses a specific set of values for the simulated annealing parameters that have proved efficient (Table S1).

Should the user wish to use different values for the simulated annealing parameters, this can be achieved by modifying the source code. The annealing parameters are held in a data structure of derived type `Params`. The type is defined in `lvb.h`. The structure is filled with final values in function `params_change()` in `getparam.c`. Parameters could be set to the user-defined values there, providing the values are acceptable (Table S1). For LVB to output the values in use, set the `verbose` field of the structure to `LVB_TRUE`.

Temperature remains at a given level until a decrease is required by `maxpropose` or `maxaccept`. Following Kirkpatrick et al. (1983), temperature levels are related exponentially, with the *n*th temperature level defined as $T_n = (T_1/T_0)^n T_0$. This is not user-configurable but could be altered by modifying the source code for `anneal()` in file `solve.c`, for example to give linear decrease. Care must be taken to avoid numerical problems with temperatures near zero: the existing code provides guidance.

Where `runs` > 1, LVB outputs the most parsimonious trees found across all searches. With more than one search, all searches begin with a random tree except for the last search, which begins with one of the most parsimonious trees already found.

Parameter	Field in `Params`	Type	Acceptable range	'Fast' setting	'Slow' setting
Initial temperature	`t0`	`double`	$0 <$ `t0` ≤ 1	0.0001	0.0001
Second temperature	`t1`	`double`	$0 <$ `t1` $<$ `t0`	9.9e-05	9.9e-05
Maximum changes proposed at a temperature	`maxpropose`	`long`	`maxpropose` ≥ 1	1000	2000
Maximum new trees at least as good as the then-current tree accepted at a temperature	`maxaccept`	`long`	`maxaccept` ≥ 1	5	5
Maximum consecutive temparatures allowed at each of which less than `maxaccept` new better trees are accepted, per search	`maxfail`	`long`	`maxfail` ≥ 1	3	40
Number of independent searches	`runs`	`long`	`runs` ≥ 1	1	2

Table S1 Simulated annealing parameters in LVB 2.0.

References

Croes, G.A. 1958. A method for solving travelling-salesman problems. *Operations Research*, **6**, 791–812.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

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

Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, **20**, 406–416.

Holland, J.H. 1975. *Adaptation in natural and artificial systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, Ann Arbor.

Kirkpatrick, S., Gelatt, C.D. and Vecci, M.P. 1983. Optimization by simulated annealing. *Science*, **220**, 671–680.

Lewis, P.O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution*, **15**, 277–283.

Lundy, M. 1985. Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrica*, **72**, 191–198.

Salter, L.A and Pearl, D.K. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology*, **50**, 7–17.

Swofford, D.L. 1993. *PAUP: Phylogenetic Analysis using Parsimony*, Version 3.1. Illinois Natural History Survey, Champaign, Illinois.

Swofford, D.L. 2002. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*, 4.0 Beta. Sinauer Associates, Sunderland, Massachusetts.