BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 2 2003
Pages 301–302

r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock

Michael J. Sanderson

Section of Evolution and Ecology, One Shields Avenue, University of California, Davis, CA 95616, USA

Received on August 2, 2002; revised on August 29, 2002; accepted on September 3, 2002

ABSTRACT

Summary: Estimating divergence times and rates of substitution from sequence data is plagued by the problem of rate variation between lineages. R8s version 1.5 is a program which uses parametric, nonparametric and semiparametric methods to relax the assumption of constant rates of evolution to obtain better estimates of rates and times. Unlike most programs for rate inference or phylogenetics, r8s permits users to convert results to absolute rates and ages by constraining one or more node times to be fixed, minimum or maximum ages (using fossil or other evidence). Version 1.5 uses truncated Newton nonlinear optimization code with bound constraints, offering superior performance over previous versions.

Availability: The linux executable, C source code, sample data sets and user manual are available free at http://ginger.ucdavis.edu/r8s.

Contact: mjsanderson@ucdavis.edu

Development of this program was motivated by two observations. First, the abundance of molecular sequence data has renewed interest in estimating divergence times (Korber *et al.*, 2000; Wikstrom *et al.*, 2001), but there is wide appreciation that such data typically show strong departures from a molecular clock. This has prompted considerable work aimed at developing methods for estimating rates and times in the absence of a clock (Huelsenbeck *et al.*, 2000; Kishino *et al.*, 2001; Rambaut and Bromham, 1998; Sanderson, 1997, 2002; Takezaki *et al.*, 1995; Thorne *et al.*, 1998). However, software tools for this are still specialized and generally do not permit the addition of flexible calibration methods for converting relative times and rates to an absolute scale. Second, there are surprisingly few published estimates of absolute rates of molecular evolution obtained in a phylogenetic framework. Most estimates are based on pairwise distance methods with simple fixed calibration points (Nei *et al.*, 2001; Wray *et al.*, 1996), which can suffer from problems of statistical nonindependence (Pagel, 1999). The r8s program is designed to facilitate estimation of rates of molecular evolution, particularly estimation of the *variability* in such rates across a tree. This should improve prospects for identifying causal factors underlying rate differences among genes and organisms.

R8s implements several methods for estimating time and rate parameters. Methods range from standard maximum likelihood approaches in the context of global or local molecular clocks to more experimental semiparametric and nonparametric methods that relax the stringency of the clock assumption using smoothing methods (Hastie *et al.*, 2001). Its starting point is a user-provided phylogenetic tree with estimated branch lengths (numbers of substitutions along each branch) obtained by some other program, such as PAUP* (Swofford, 2002). The program's speed derives partly from the fact that branch length estimation is left to other programs. This permits r8s to concentrate on lineage variation and its impact on rate and time estimation. One or more calibration points can be added to permit scaling of rates and times to absolute units. These calibrations can take one of two forms: assignment of a fixed age to a node, or assignment of a minimum or maximum age constraint to a node, which is generally a better reflection of the information content of fossil evidence. Terminal nodes are permitted to occur at any point in time, allowing investigation of rate variation in phylogenies such as those obtained from 'serial' samples of viral lineages through time.

The program emphasizes nonparametric and semiparametric approaches (in contrast to fully parametric methods of e.g. Huelsenbeck *et al.*, 2000; Kishino *et al.*, 2001; Rambaut and Bromham, 1998). Analogous to smoothing techniques in regression analysis, these methods estimate unknown divergence times by smoothing the rapidity of rate change among lineages. The latter is done with a function that penalizes rates that change too quickly from branch to neighboring branch. Nonparametric rate smoothing (NPRS: Sanderson, 1997) uses nothing but this penalty function as an objective function. Since the penalty depends on the unknown divergence times,

optimization using this objective function permits estimation of these times. NPRS, however, can sometimes overfit the data leading to rapid fluctuations in rate in regions of a tree that have short branches. This can be overcome using penalized likelihood (Sanderson, 2002), a method that combines likelihood and the nonparametric penalty function used in NPRS. It permits specification of the relative contribution of the rate smoothing and the data-fitting parts of the estimation procedure. The user can select any level of smoothing from severe, which essentially leads to a molecular clock, to highly unconstrained, allowing very rapid changes in rate across a tree. A cross validation procedure is included to provide a data-driven method for finding the optimal level of smoothing (Hastie *et al.*, 2001).

Time and rate estimation is subject to upper and/or lower bounds on divergence times. Older versions of r8s implemented bound constraints using general nonlinear constraint methods such as barrier/penalty functions (Gill *et al.*, 1981), with optimization relying on Powell's method without gradients. This was reasonably effective but slow. More recently, a quasi-Newton algorithm was added, which used gradients, but did not take constraints into account. Clearly this was faster and more accurate, but inadequate with respect to time calibration methods. Version 1.5 now uses a proper bound-constrained truncated Newton method (with gradients) developed by Nash (2000) and available in the NetLib library of tools for optimization (http://www.netlib.org).

Input files must adhere to the 'Nexus' format (Maddison *et al.*, 1997), an extensible data format widely used for phylogenetic software. Commands can either be issued interactively at a command line prompt, or imbedded in a special 'r8s block' in the Nexus file. Output consists of simple text plots of trees with or without time calibrations, and Nexus style output of tree descriptions with time-calibrated branch lengths. These can be read by more graphics-savvy programs such as PAUP (Swofford, 2002). Output includes age estimates and a description of the variability in absolute rates of molecular evolution on a tree, presented as both a table and as a tree in which branch lengths are drawn proportional to absolute rates. If a collection of identical tree topologies with different branch lengths is input, summary statistics across the trees for estimated branch lengths, rates or divergence times at a node, can be obtained. This is useful for examining the sensitivity of these estimates to a variety of factors or to use bootstrapping to estimate their confidence intervals. Miscellaneous convenience features include the ability to read ultrametric trees obtained from other programs and calibrate them with a single node (absolute) time. The user can also prune lineages, collapse zero-length branches to polytomies, and reroot trees. For performance evaluation studies, routines are included to simulate trees according to branching process models and to simulate substitutions on branches according to Poisson process substitution models.

REFERENCES

Gill,P.E., Murray,W. and Wright,M.H. (1981) *Practical optimization*. Academic Press, London, New York.

Hastie,T., Tibshirani,R. and Friedman,J.H. (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York.

Huelsenbeck,J.P., Larget,B. and Swofford,D. (2000) A compound Poisson process for relaxing the molecular clock. *Genetics*, **154**, 1879–1892.

Kishino,H., Thorne,J.L. and Bruno,W.J. (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.*, **18**, 352–361.

Korber,B., Muldoon,M., Theiler,J., Gao,F., Gupta,R., Lapedes,A., Hahn,B.H., Wolinsky,S. and Bhattacharya,T. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science*, **288**, 1789–1796.

Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1997) Nexus: an extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621.

Nash,S.G. (2000) A survey of truncated-Newton methods. *J. Comput. Appl. Math.*, **124**, 45–59.

Nei,M., Xu,P. and Glazko,G. (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA*, **98**, 2497–2502.

Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature (London)*, **401**, 877–884.

Rambaut,A. and Bromham,L. (1998) Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.*, **15**, 442–448.

Sanderson,M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.*, **14**, 1218–1231.

Sanderson,M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol. Biol. Evol.*, **19**, 101–109.

Swofford,D.L. (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, 4.0 edn, Sinauer, Sunderland, MA.

Takezaki,N., Rzhetsky,A. and Nei,M. (1995) Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.*, **12**, 823–833.

Thorne,J.L., Kishino,H. and Painter,I.S. (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.

Wikstrom,N., Savolainen,V. and Chase,M.W. (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. Biol. Sci. B*, **268**, 2211–2220.

Wray,G.A., Levinton,J.S. and Shapiro,L.H. (1996) Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science*, **274**, 568–573.