PROGRAM NOTE

TCS: a computer program to estimate gene genealogies

M. CLEMENT,* D. POSADA† and
K. A. CRANDALL†

**Department of Computer Science, Brigham Young University, Provo, UT 84602, USA, †Department of Zoology, 574 Widtsoe Building, Brigham Young University, Provo, UT 84602, USA*

Keywords: cladogram estimation, gene genealogy, phylogeography, population genetics, software

Received 11 February 2000; revision accepted 26 April 2000

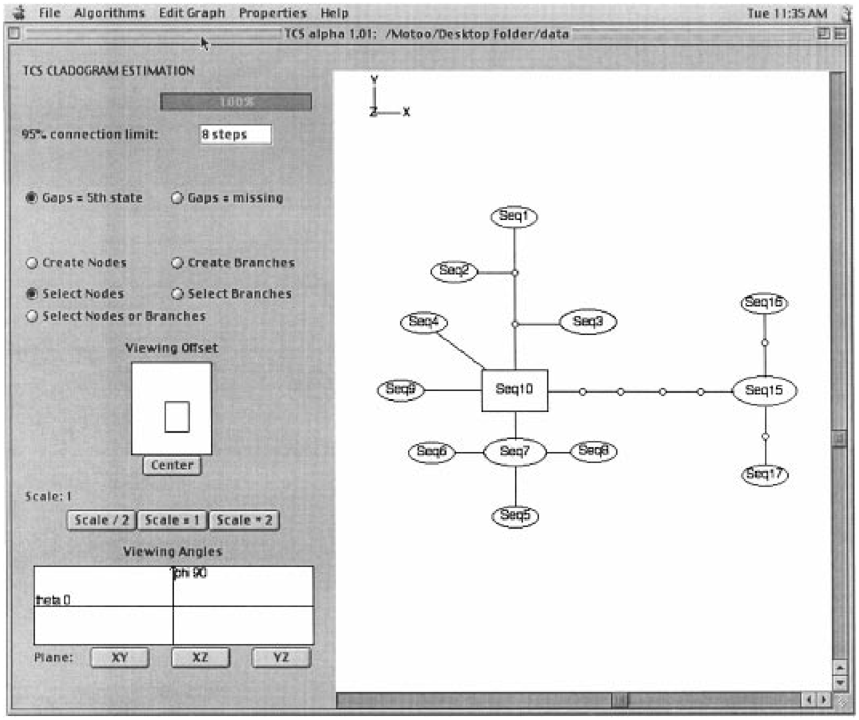
Correspondence: Keith A. Crandall. Fax: (801) 378 7423; E-mail: kac@email.byu.edu

Phylogenies are extremely useful tools, not only for establishing genealogical relationships among a group of organisms or their parts (e.g. genes), but also for a variety of research once the phylogenies are estimated. In a recent review, Pagel (1999) eloquently outline a number of uses for phylogenetic information from discovery of drug resistance to reconstructing the common ancestor to all of life. Phylogenies have been used to predict future trends in infectious disease (Bush *et al*. 1999) and have even been offered as evidence in a court of law (Vogel 1997). Yet phylogenies are only as useful as they are accurate.

Estimating genealogical relationships among genes at the population level presents a number of difficulties to traditional methods of phylogeny reconstruction. These traditional methods such as parsimony, neighbour-joining, and maximum-likelihood make assumptions that are invalid at the population level. For example, these methods assume ancestral haplotypes are no longer in the population, yet coalescent theory predicts that ancestral haplotypes will be the most frequent sequences sampled in a population level study (Watterson & Guess 1977; Donnelly & Tavaré 1986; Crandall & Templeton 1993). Traditional methods require reasonably large numbers of variable characters to accurately reconstruct relationships (Huelsenbeck & Hillis 1993) and population level studies typically lack such variation. Also, recombination is a real possibility among sequences at the population level and traditional methods assume recombination does not occur. The failure to incorporate the possibility of recombination in phylogeny reconstruction can lead to grave errors in the resulting estimated phylogeny. The combination of these effects can lead parsimony methods to infer a cumbersome amount of most parsimonious trees at the population level with no resolution among the set (e.g. over one billion trees for a set of human mitochondrial DNA (mtDNA), Excoffier & Smouse 1994). These effects can also lead neighbour-joining and traditional maximum-likelihood methods to be over confident in the resulting relationships (Bandelt *et al*. 1995). Therefore, an alternative approach is needed to provide accurate estimates of gene genealogies at the population level that take into account these population level phenomena not addressed by traditional methods.

Multiple groups have looked to network representations for population level genealogical information (Bandelt & Dress 1992; Templeton *et al*. 1992; Excoffier & Smouse 1994; Fitch 1997). Networks allow one to naturally incorporate the often-times nonbifurcating genealogical information associated with population level divergences. The method of Templeton *et al.* (1992) (TCS) has been used extensively with restriction site and nucleotide sequence data to infer population level genealogies when divergences are low (Georgiadis *et al*. 1994; Routman *et al*. 1994; Gerber & Templeton 1996; Hedin 1997; Schaal *et al*. 1998; Vilá *et al*. 1999, Gómez-Zurita *et al*. 2000). TCS has been used with traditional methods to estimate relationships among organisms that span a wide range of divergence (Crandall & Fitzpatrick 1996; Benabib *et al*. 1997). The approach has also been used extensively with a nested analysis procedure to partition population structure from population history (Templeton *et al*. 1995; Templeton 1998) and explore the phylogeographic history of a diversity of organisms (e.g. Johnson & Jordon 2000; Turner *et al*. 2000). In this note, we announce the availability of a new software package, TCS, to estimate genealogical relationships among sequences using the method of Templeton *et al*. (1992).

The TCS software opens nucleotide sequence files in either NEXUS (Maddison *et al*. 1997) or PHYLIP (Felsenstein 1991) sequential format. Sequences should not be collapsed into haplotypes as frequency data can be incorporated into the output. The program collapses sequences into haplotypes and calculates the frequencies of the haplotypes in the sample. These frequencies are used to estimate haplotype outgroup probabilities, which correlate with haplotype age (Donnelly & Tavaré 1986; Castelloe & Templeton 1994). An absolute distance matrix is then calculated for all pairwise comparisons of haplotypes. The probability of parsimony [as defined in Templeton *et al*. (1992), equations 6, 7, and 8] is calculated for pairwise differences until the probability exceeds 0.95. The number of mutational differences associated with the probability just before this 95% cut-off is then the maximum number of mutational connections between pairs of sequences justified by the 'parsimony' criterion. These justified connections are then made resulting in a 95% set of plausible solutions. The program outputs the sequences, the pairwise absolute distance matrix, probabilities of parsimony for mutational steps just beyond the 95% cut-off, a test listing of connections made and missing intermediates generated, and a graph output file containing the resulting network (Fig. 1). This graph output file can be opened in the freeware VGJ 1.0.3 (http://www.eng.auburn.edu/department/cse/research/graph_drawing/graph_drawing.html; distributed under the terms of the GNU General Public License, Version 2), which is packaged with the TCS algorithm. The program can handle a reasonable number of sequences. For example, an HTLV data set with 69 haplotypes of length 725 bp took over one hour to run in a Macintosh G3. Memory requirements are



TCS CLADOGRAM ESTIMATION

100%

95% connection limit: 8 steps

● Gaps = 5th state ○ Gaps = missing

○ Create Nodes ○ Create Branches

● Select Nodes ○ Select Branches

○ Select Nodes or Branches

Scale: 1

Fig. 1 TCS Java interface. The maximum number of steps connecting parsimoniously two haplotypes is indicated. Gaps can be treated as a 5th state or as missing data. The graph can be edited and arranged using different algorithms. By double-clicking over a haplotype, some information is displayed, such as sequences included in the haplotype and outgroup weights. The haplotype with the highest outgroup probability is displayed as a square, while other haplotypes are displayed as ovals. The size of the square or oval corresponds to the haplotype frequency.

low, and the program will run with less than 1 MB RAM. The TCS software package, including executables for Mac and PC, documentation, and Java source code, is distributed freely and is available at our website, along with a host of other programs for population genetic and phylogenetic analyses: http://bioag.byu.edu/zoology/crandall_lab/programs.htm.

Acknowledgements

This work was supported by the Alfred P. Sloan Foundation, a Shannon Award from the National Institutes of Health, and NIH R01-HD34350.

References

Bandelt H-J, Dress AWM (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, **1**, 242–252.

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics*, **141**, 743–753.

Benabib M, Kjer KM, Sites JW Jr (1997) Mitochondrial DNA sequence-based phylogeny and the evolution of viviparity in the *Sceloporus scalaris* group (Reptilia, Squamata). *Evolution*, **51**, 1262–1275.

Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. *Science*, **286**, 1921–1925.

Castelloe J, Templeton AR (1994) Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution*, **3**, 102–113.

Crandall KA, Fitzpatrick JF Jr (1996) Crayfish molecular systematics: Using a combination of procedures to estimate phylogeny. *Systematic Biology*, **45**, 1–26.

Crandall KA, Templeton AR (1993) Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics*, **134**, 959–969.

Donnelly P, Tavaré S (1986) The ages of alleles and a coalescent. *Advances in Applied Probability*, **18**, 1–19.

Excoffier L, Smouse PE (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. *Genetics*, **136**, 343–359.

Felsenstein J (1991) *PHYLIP: Phylogenetic Inference Package*. University of Washington, Seattle, WA.

Fitch WM (1997) Networks and viral evolution. *Journal of Molecular Evolution*, **44**, S65–S75.

Georgiadis N, Bischof L, Templeton A *et al.* (1994) Structure and history of African elephant populations: I. Eastern and Southern Africa. *Journal of Heredity*, **85**, 100–104.

Gerber AS, Templeton AR (1996) Population sizes and within-deme movement of *Trimerotropis saxatilis* (Acrididae), a grasshopper with a fragmented distribution. *Oecologia*, **105**, 343–350.

Gómez-Zurita J, Petitpierre E, Juan C (2000) Nested cladistic analysis, phylogeography and speciation in the *Timarcha goettingensis* complex (Coleoptera, Chrysomelidae). *Molecular Ecology*, **9**, 557–570.

Hedin MC (1997) Speciational history in a diverse clade of habitat-specialized spiders (Araneae: Nesticidae: *Nesticus*): Inferences from geographic-based sampling. *Evolution*, **51**, 1929–1945.

Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, **42**, 247–264.

Johnson JB, Jordon S (2000) Phylogenetic divergence in leatherside chub (*Gila copei*) inferred from mitochondrial cytochrome *b* sequences. *Molecular Ecology*, **9**, 1029–1035.

Maddison DR, Swofford DL, Maddison WP (1997) NEXUS: an extensible file format for systematic information. *Systematic Biology*, **46**, 590–621.

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

Routman E, Wu R, Templeton AR (1994) Parsimony, molecular evolution, and biogeography: The case of the North American Giant Salamander. *Evolution*, **48**, 1799–1809.

Schaal BA, Hayworth DA, Olsen KM, Rouscher JT, Smith WA (1998) Phylogeographic studies in plants: problems and prospects. *Molecular Ecology*, **7**, 465–474.

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–633.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.

Turner TF, Trexler JC, Harris JL, Haynes JL (2000) Nested cladisitic analysis indicates population fragmentation shapes genetic diversity in a freshwater mussel. *Genetics*, **154**, 777–785.

Vilá C, Amorim IR, Leonard JA *et al.* (1999) Mitochondrial DNA phylogeography and population history of the Gray Wolf *Canis lupus*. *Molecular Ecology*, **8**, 2089–2103.

Vogel G (1997) Phylogenetic analysis: getting its day in court. *Science*, **275**, 1559–1560.

Watterson GA, Guess HA (1977) Is the most frequent allele the oldest? *Theoretical Population Biology*, **11**, 141–160.