TCS: estimating gene genealogies

Version 1.09

2001 © Mark Clement, Jacob Derington (Computer Science) and David Posada (Zoology). Brigham Young University, Provo, UT 84602, US.

This software is distributed as it is. We are not responsible for the results of obtained with it.

TCS is a computer program that implements the estimation of gene genealogies from DNA sequences as described by (Templeton et al. 1992). This cladogram estimation method is also known as statistical parsimony. Some useful references are indicated below.

The *TCS* software package, including executables for Mac and PC, documentation, and Java source code, is distributed freely and is available at our web site, along with a host of other programs for population genetic and phylogenetic analyses:
http://bioag.byu.edu/zoology/crandall_lab/programs.htm

Version History (Current version 1.09)
Alpha 1.00: First version of the program.
Alpha 1.01: distances file included
Alpha 1.02: outgroups weights estimation included
Version 1.06: several cosmetic changes and some bugs fixed
Version 1.07-1.09: Fixed bug that was creating several unconnected haplotyped (when they should be connected) for some big data sets. The progress of the calculations are showed in the GUI.

Bugs report, questions, etc…
We will add new features to the next versions and try to correct potential bugs. If you are using this program, we should know that, so we can inform you about mistakes and new versions. Please feel free to make questions. Please contact David Posada at david.posada@byu.edu

Program Citation
Clement, M., D. Posada and K. A. Crandall 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology 9 (10): 1657-1660.

Input file (DNA / absolute distances)

The TCS software works with DNA nucleotide sequence. It opens DNA alignment files in either Nexus [Maddison, 1997 #2791] or PHYLIP (Felsenstein 1991) sequential format. Alternatively, absolute distance files in modified NEXUS or PHYLIP files can can also be used.

Aligned DNA sequences

Sequential NEXUS:

```
#NEXUS

Begin data;
      Dimensions ntax=4 nchar=6;
      Format datatype=nucleotide gap=- missing=? ;
      Matrix

Seq1      AAAAA-
Seq2      AAAAC-
Seq3      AAAAA?
Seq4      AAAAAA
      ;
end;
```

Sequential Phylip:

```
4 6
Seq1 AAAAA-
Seq2 AAAAC-
Seq3 AAAAA?
Seq4 AAAAAA
```

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 outrgroup probabilities, which correlate with haplotype age (Donnelly and Tavaré 1986; Castelloe and Templeton 1994).

Distance file

We included an option to read a matrix of absolute distances AMONG HAPLOTYPES. The matrix should be LOWER DIAGONAL in NEXUS (example_dis.nex) or PHYLIP (example_dis.phy) format.

IMPORTANT: you have to add the "nchar" to these files, so the 95% connection limit can be calculated. Look a the example files:

Modified Nexus format

#NEXUS

```
Begin taxa;
	Dimensions ntax=10;
	Taxlabels
		Seq1
		Seq2
		Seq3
		Seq4
		Seq5
		Seq6
		Seq7
		Seq8
		Seq9
		Seq10
		;
End;

Begin distances;
	Format triangle=lower labels nodiagonal;
	Matrix
Seq1
Seq2	2
Seq3	2 2
Seq4	3 3 3
Seq5	4 4 4 3
Seq6	4 4 4 3 2
Seq7	3 3 3 2 1 1
Seq8	4 4 4 3 2 2 1
Seq9	3 3 3 2 3 3 2 3
Seq10	2 2 2 1 2 2 1 2 1
	;
End;
```

Modified Phylip format

```
10 404
Seq1
Seq2        2
Seq3        2  2
Seq4        3  3  3
Seq5        4  4  4  3
Seq6        4  4  4  3  2
Seq7        3  3  3  2  1  1
Seq8        4  4  4  3  2  2  1
Seq9        3  3  3  2  3  3  2  3
Seq10       2  2  2  1  2  2  1  2  1
```

Treatment of Gaps (5th state / missing data)
By default, gaps are counted as events (i.e. treated as a fifth state). You can turn off this option in the program interface (Figure 1) and treat gaps as missing data.

Limits of parsimony (estimated/user defined)
The probability of parsimony (as defined in Templeton *et al.* [1992], equations 6, 7, and 8) is calculated for DNA pairwise differences until the probability exceeds 0.95. The number of mutational differences associated with the probability just before this 95% cutoff is then the maximum number of mutational connections between pairs of sequences justified by the "parsimony" criterion. Alternatively, this limit can be set up by the user (see Figure 1).

Note: TCS is not for RFLPs
TCS calculations are for only for DNA sequence data. If your data is RFLPs you might think you could input absolute distances, but that would not work. The problem is that for each pair of RFLP haplotypes, the parsimony connection limit could be different, depending on the number of shared sites. This is because for RFLPs the total number of characters minus the number of characters with a different state does not necessarily equal the number of shared characters (which is true for DNA sequences). The difference with DNA sequences is that ++ is a shared site, while -- is not a shared site. But you could build an RFLP network by hand. The parsimony probability for RFLP data can be calculated using the program ParsProb (http://bioag.byu.edu/zoology/crandall_lab/programs.htm) for each pair of RFLP haplotypes.

Logfile
Each time that the TCS analysis is performed, a graph file (GML format) is saved (*logfile*). This file contains information on the run: probabilities of parsimony for mutational steps, the pairwise absolute distance matrix, a test listing of connections made and missing intermediates generated, outgroup weights for each haplotype, a graph description, and the date and time ellapsed for the analysis.

Graphfile
Each time that the TCS analysis is performed, a graph file (GML format) is saved. The name of this file will be *datafilename.graph* This graph can be open posteriorly in TCS.

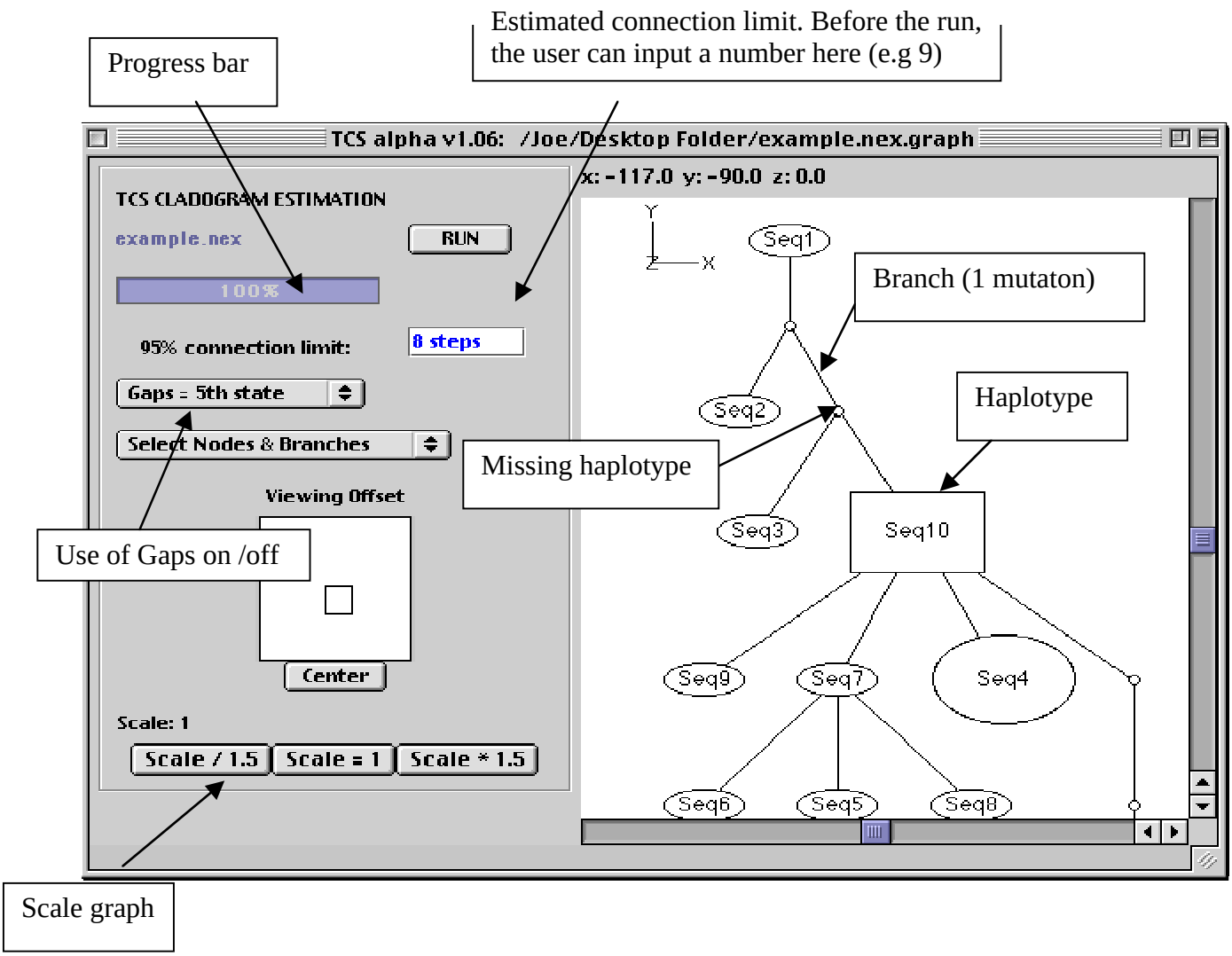
VGJ
For graphic purposes, the freeware VGJ 1.0.3, distributed under the terms of the GNU General Public License, Version 2), is packaged within the TCS program.
(http://www.eng.auburn.edu/department/cse/research/graph_drawing/graph_drawing.html;

Running TCS
1. Open the DNA data file in the FILE menu
2. Click on Run
3. The program reads the file and collapses sequences to haplotypes
4. An absolute distance matrix is then calculated for all pairwise comparisons of haplotypes.
5. The parsimony connection limit is calculated. Alternatively, this limit can be set up by the user (see Figure 1).
6. These justified connections are then made resulting in a 95% set of plausible networks (1 or more)
7. A graph is generated and automatically opened. In this graph, haplotytpes are drawn in a size proportional to their frequency.

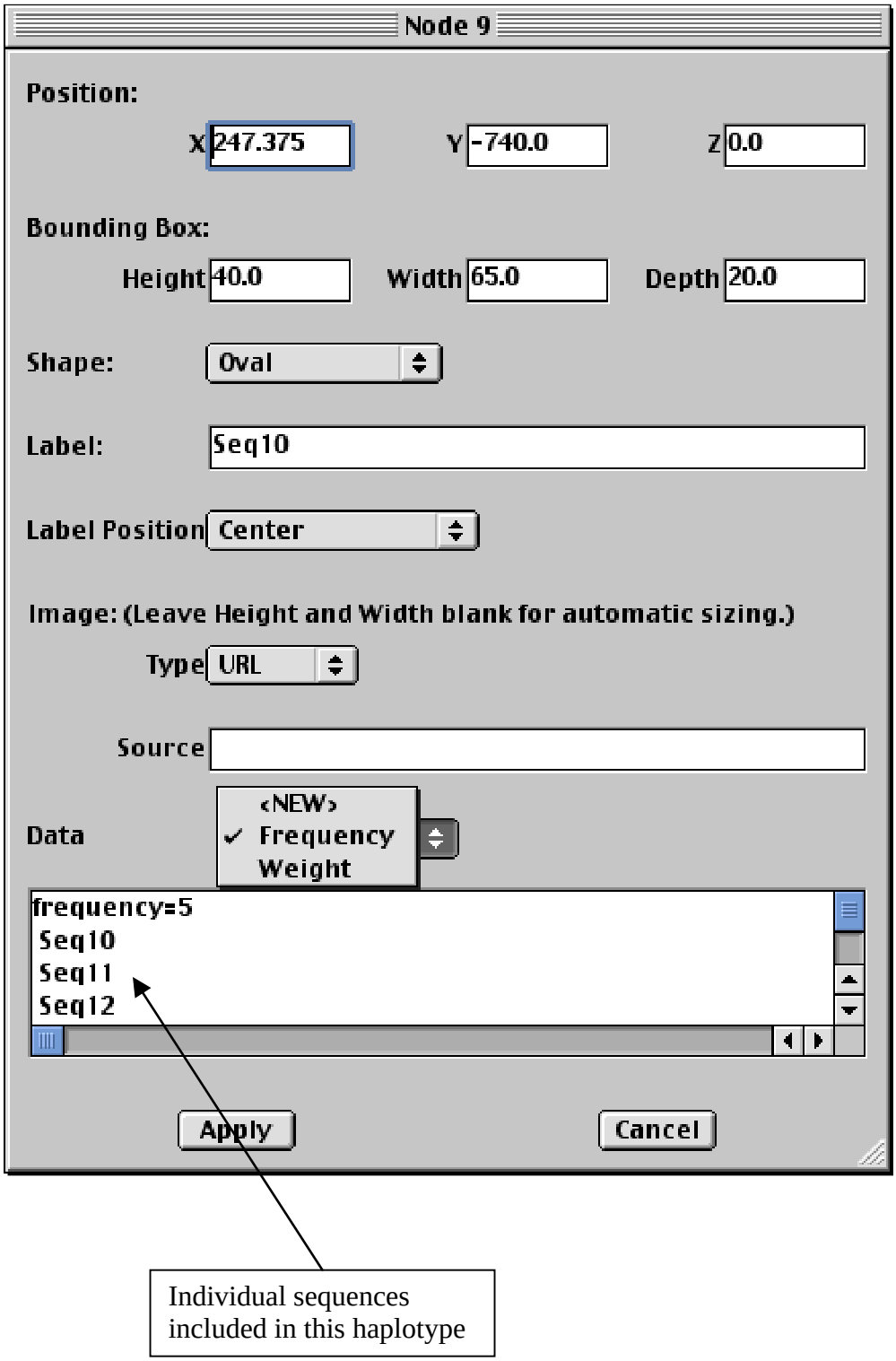
Editing the graph
You can select (by clicking), create and delete nodes (haplotypes) o branches on the graph. Automatic algorithms to order the graph are available in the menus. You can move the nodes and branches around and save the file as postcript. By double-clicking on a haplotype node, you will be able of displaying its frequency and its outgroupt weight. The haplotoype in a square has the biggest outgroup weight.

Printing
The graph is printed by being saved as a postcrip file and sent manually to the printer or as a PICT file.

Running times
The program can handle a reasonable number of sequences. For example, an HTLV data set with 69 haplotypes of length 725 bps took over one hour to run in a Macintosh G3. Memory requirements are low, and the program will run with less than 1 MB RAM.


Progress bar
Estimated connection limit. Before the run, the user can input a number here (e.g 9)
TCS alpha v1.06: /Joe/Desktop Folder/example.nex.graph
TCS CLADOGRAM ESTIMATION
example.nex
RUN
100%
95% connection limit:
8 steps
Gaps = 5th state
Select Nodes & Branches
Viewing Offset
Center
Scale: 1
Scale / 1.5
Scale = 1
Scale * 1.5
x: -117.0 y: -90.0 z: 0.0
Seq1
Seq2
Seq3
Seq10
Seq9
Seq7
Seq4
Seq6
Seq5
Seq8
Branch (1 mutaton)
Haplotype
Missing haplotype
Use of Gaps on /off
Scale graph

If you double click in the node “Seq10” you will display:



Useful references

Castelloe, J. and A. R. Templeton 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. **3:** 102-113.

Clement, M., D. Posada and K. A. Crandall 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology (in press)**:**

Crandall, K. A. 1994. Intraspecific cladogram estimation: Accuracy at higher levels of divergence. Syst. Biol. **43:** 222-235.

Crandall, K. A. 1995. Intraspecific phylogenetics: Support for dental transmission of human immunodeficiency virus. J. Virol. **69:** 2351-2356.

Crandall, K. A. 1996a. Multiple interespecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. Mol. Biol. Evol. **13:** 115-131.

Crandall, K. A. 1996b. Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. Mol. Biol. Evol. **13:** 115-131.

Crandall, K. A. and A. R. Templeton 1996. Applications of intraspecific phylogenetics. Pp. 81-99. *in* P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee, eds. New Uses for New Phylogenies. Oxford University Press, Oxford, England.

Crandall, K. A., A. R. Templeton and C. F. Sing 1994. Intraspecific phylogenetics: problems and solutions. Pp. 273-297. *in* R. W. Scotland, D. J. Siebert and D. M. Williams, eds. Models in Phylogeny Reconstruction. Clarendon Press, Oxford, England.

Donnelly, P. and S. Tavaré 1986. The ages of alleles and a coalescent. Adv. Appl. Prob. **18:** 1-19.

Felsenstein, J. 1991. PHYLIP: Phylogenetic Inference Package. 3.4. University of Washington, Seattle, WA.

Templeton, A. R. 1995. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. Genetics **140:** 403-409.

Templeton, A. R. 1998. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. Molecular Ecology **7:** 381-397.

Templeton, A. R., E. Boerwinkle and C. F. Sing 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics **117:** 343-351.

Templeton, A. R., K. A. Crandall and C. F. Sing 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, A. R., E. Routman and C. A. Phillips 1995. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger salamander, *Ambystoma tigrinum*. Genetics **140:** 767-782.

Templeton, A. R. and C. F. Sing 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics **134:** 659-669.

February 22, 2001

David Posada

dp47@email.byu.edu