Short Communication

FreeTree – Freeware Program for Construction of Phylogenetic Trees on the Basis of Distance Data and Bootstrap/Jackknife Analysis of the Tree Robustness. Application in the RAPD Analysis of Genus *Frenkelia*.

(molecular phylogenetics / fingerprinting / UPGMA / neighbor-joining / software / population structure / coccidia)

A. PAVLÍČEK[1], Š. HRDÁ[2], J. FLEGR[2]

[1]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[2]Department of Parasitology, Faculty of Science, Charles University, Prague, Czech Republic

Abstract. The Win95 program for computation of distance matrixes and construction of phylogenetic or phenetic trees on the basis of RAPD, RFLP and allozyme data was presented. In contrast with other presently available software, the program FreeTree can also assess the robustness of the tree topology by bootstrap, jackknife or OTU-jackknife analysis. Moreover, the program can be used also for an analysis of data obtained in several independent experiments performed with nonidentical subsets of taxa. The function of the program was demonstrated by an analysis of RAPD data from 22 strains of *Frenkelia*. The program is available as an autoextractive archive containing the installation files of FreeTree and TreeView, manual in MS Word format and a sample of the input file at http://www.natur.cuni.cz/~flegr/programs/freetree.

The advent of molecular taxonomy techniques offered a solution for many problems which were for a long time out of reach of classical taxonomy methods and approaches. At the present time the methods of construction of phylogenetic trees on the basis of molecular data are widely used not only in systematic and comparative biology, but also in ecology, ethology, sociobiology and epidemiology. The methods of molecular taxonomy can be divided into two groups: single-locus methods and multi-loci methods. The result of single-locus methods such as DNA sequencing, microsatellite analysis and single-strand conformation polymorphism (SSCP) analysis is a so called gene tree, the topology of which reflects the evolution of a particular locus (gene). Mostly, however, the aim of our analysis is to get a species tree, the phylogenetic tree of the taxons under the study or the genealogical tree of the individuals in the studied population. Under favourable conditions (long intervals between speciation events, absence of any horizontal gene transfer between the species), the topology of the gene tree can reflect the topology of the desired species tree. However, such single locus-based species trees often contain errors (Takahata and Nei, 1985; Neigel and Avise, 1986). More reliable results can be obtained using the multi-loci methods like DNA hybridization, randomly amplified polymorphic DNAs (RAPD), or restriction fragment lenghth polymorphism (RFLP). These methods provide a species tree based on phylogenetically relevant information contained in many loci or (in the ideal case) in the whole genome. These methods are often even cheaper and easier to perform than the usual single-locus methods. There is, however, one serious technical obstacle for routine application of multi-loci methods. For single-locus methods, a broad collection of application software exist for all stages of the data analysis including the programs for automatic input of data, construction of dendrograms and statistical analysis of the reliability of the results. For the multi-loci methods, such programs are scarce and for some steps of the analysis are even not available. For example, no program exists either in the public or in the commercial domain for the construction of phylogenetic/genealogic trees on the basis of RFLP and RAPD data, which can also perform the bootstrap or jackknife analysis of the robustness of the tree topology. While all trees constructed on the basis of DNA sequencing and described in the scientific literature presently contain the information about the robustness of the tree topology (mostly the bootstrapping values for internal

Received November 5, 1998. Accepted January 21, 1999.

The work was supported by the grants GAUK 107/1998, GACR 206/95/0638 and VS96142.

Corresponding author: Jaroslav Flegr, Department of Parasitology, Faculty of Science, Charles University, Viničná 7, 128 44 Prague 2, Czech Republic. Fax (4202) 299 713. Tel. (4202) 2195 3289. E-mail: flegr@beba.cesnet.cz.

Abbreviations: AP-PCR – arbitrarily primed PCR, OTU – operational taxonomic units, PCR – polymerase chain reaction, RAPD – randomly amplified polymorphic DNAs, RFLP – restriction fragment length polymorphism, SSCP – single-strand conformation polymorphism, UPGMA – unweighted pair group method with arithmetic averages.

branches of the tree), this fundamental information is never provided for the trees constructed on the basis of multi-loci methods. At the same time the information about the robustness of the tree is often critical from the point of view of biological interpretation of the data.

The purpose of our program FreeTree was to fill in this gap in the molecular taxonomy software. It was originally intended for the analysis of results of DNA fingerprinting methods (RFLP, RAPD, arbitrarily primed polymerase chain reaction (AP-PCR)) or other methods which provide binary character data (presence/absence of the characters). For such data the program computes the distance matrix, constructs the phylogenentic or phenetic tree using unweighted pair group method with arithmetic averages (UPGMA) or neighbor-joining, and computes bootstraping or jackknifing values for internal branches of the tree. The program can also be used for the construction of trees on the basis of any distance data or on the basis of frequency data (e.g. results of isoenzyme analysis). In such cases, however, the program cannot test the reliability of the trees.

The function of the program can be shown on the example of analysis of RAPD data (information on the presence or absence of all 405 PCR fragments which were visible on 12 electrophoreograms) from 22 strains of *Frenkelia*, the coccidian parasite of raptors (definitive hosts) and small rodents (intermediate hosts). The data were prepared as an Excel matrix with the names of 22 strains of *Frenkelia* (9 strains of *F. microti* and 13 strains *F. glareoli*) and 2 strains of *Toxoplasma gondii* (outgroups) in the first row of the matrix. In the next 405 rows, the presence or absence of a particular DNA fragment obtained by the PCR amplification with 12 different random primers (Operon, CA) were for every strain coded by 1 or 0. The names or MW of particular fragments can be written in the first column of the spreadsheet. The matrix 25 × 405 cells (A1:Y405) was selected and copied to the clipboard (ALT C). The maximal number of operational taxonomic units (OTU) is 200 (columns), the maximal number of fragments (rows) is not limited by software and depends only on the size of operation memory of the computer. Then the "New analysis" the program FreeTree was started and the contents of the clipboard was pasted into the spreadsheet (by clicking the right button of the mouse and selecting the option Paste from the flying menu). Then the Nei-Li distance (Nei and Li, 1979) from the list with 7 different similarities/distances and the method of tree construction, e.g. neighbor-joining (Saitou and Nei, 1987) were selected. After 25 seconds the program showed the distance matrix. By clicking on the tab of the sheet Reference tree, we checked the rough topology of the tree computed on the basis of this distance matrix. Then, we opened the sheet Resampling methods and by clicking the button we selected the Bootstrapping method for a resampling analysis. Into the field Repetition count we wrote 100 (to generate one hundred of resampled data sets) and pressed the button Run. After 50 minutes of computation (Pentium 133) the program showed the list of different trees computed from the resampled data. On the sheet Reference tree we checked the bootstrapping values and copied the bracketed form of the tree with the bootstraping values into the clipboard. This form of the tree was pasted into the program for drawing dendrograms; in our case we used the freeware program TreeView (Page, 1996). In the program TreeView we rooted the tree by the outgroup species *Toxoplasma gondii* and edited the tree for printing. The result of our analysis of 22 strains of *Frenkelia* is shown in Fig. 1. The results suggest that *F. glareoli* and *F. microti* are two distinct species (bootstrapping values for both branches were 100%) despite the fact that cysts of some strains included into the analysis have a slightly intermediate phenotype. The *F. glareoli* strains from the same locality seem to have a tendency to cluster together. However, very low bootstrapping values within the branch show that any conclusion about the viscosity of *Frenkelia glareoli* populations based on our data would be highly premature.

The program FreeTree can also be used for an analysis of data obtained in several independent experiments performed with nonidentical subsets of taxa. When some

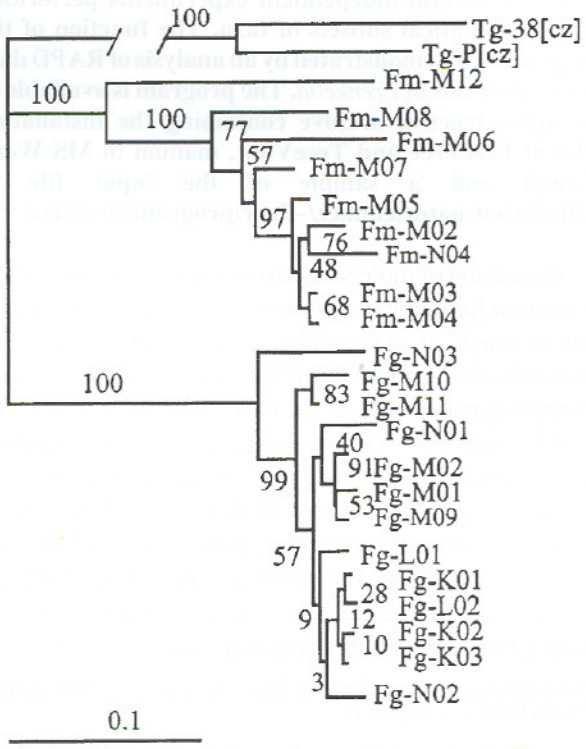

Fig. 1. Phylogenetic tree for twenty-two strains of *Frenkelia* constructed on the basis of RAPD data. The tree was rooted using two strains of *Toxoplasma gondii* Tg-P(CZ) and Tg-01529/38. Fm – *Frenkelia microti*, Fg – *Frenkelia glareoli*, Tg – *Toxoplasma gondii*. The localities of the strain isolation were labeled: M – Mikulov (South Moravia), N – Náchod (East Bohemia), L – Lednice (South Moravia), and K - Klentnice (South Moravia).

subset of taxa has been analyzed with only a limited number of random primers or when for some subset of taxa also the results from RFLP analysis are available, we can prepare a composite matrix containing the data from all experiments. In such matrix the presence or absence of the character should be coded by 1 or 0, the absence of an information about the character should be coded by an empty cell. In the matrix the groups of rows containing the characters obtained by different methods (for example by an amplification with different random primers or restriction by different enzymes) should be separated by an empty row. In the first cell of this row the name of the method (e.g. the name of restriction enzyme) can be written. When we intend to construct the tree on the basis of nucleotide distances computed from fractions of shared restriction fragments by the Nei & Li iterations method (Nei and Li, 1979), we must add a second empty row with the length of recognition site of a particular restriction enzyme written in its first cell.

FreeTree is the Windows 95/98/NT program. The bootstrapping analysis of large matrixes can take a relatively long time on slow computers. The program is available as an autoextractive archive containing the installation files of FreeTree and TreeView, manual in MS Word format and a sample of the input file at http://www.natur.cuni.cz/~flegr/programs/freetree.

Acknowledgement

We thank Jan Votýpka for providing the biological material and Tomáš Pavlíček for his help with programming.

References

Nei, M., Li, W. H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269-5273.

Neigel, J. E., Avise, A. C. (1986) Phylogenetic relationship of mitochondrial DNA under various models of speciation. In: *Evolutionary Processes and Theory*, eds. Karlin, S., Nevo, E., pp. 515-534, Academic Press, New York.

Page, R. D. M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Comp. Applicat. Biosc.* **12**, 357-358.

Saitou, N., Nei, M. (1987) The Neighbor-joining method: a new method for reconstruction of phylogenetics trees. *Mol. Biol. Evol.* **4**, 406-425.

Takahata, N., Nei, M. (1985) Gene genealogy and variance of interpopulational nucleotide difference. *Genetics* **110**, 325-344.