TREE-PUZZLE 5.0: Maximum Likelihood Phylogenetic Analysis Using Quartets and Parallel Computing

Heiko A. Schmidt[1], Korbinian Strimmer[2], Martin Vingron[1], and Arndt von Haeseler[3]

[1] Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany, {h.schmidt|m.vingron}@dkfz.de
[2] University of Oxford, Department of Zoology, South Parks Road, Oxford OX1 3PS, UK, korbinian.strimmer@zoo.ox.ac.uk
[3] Max-Planck-Institut für Evolutionäre Anthropologie, Evolutionäre Genetik, Inselstr. 22, D-04103 Leipzig, Germany, haeseler@eva.mpg.de

Introduction

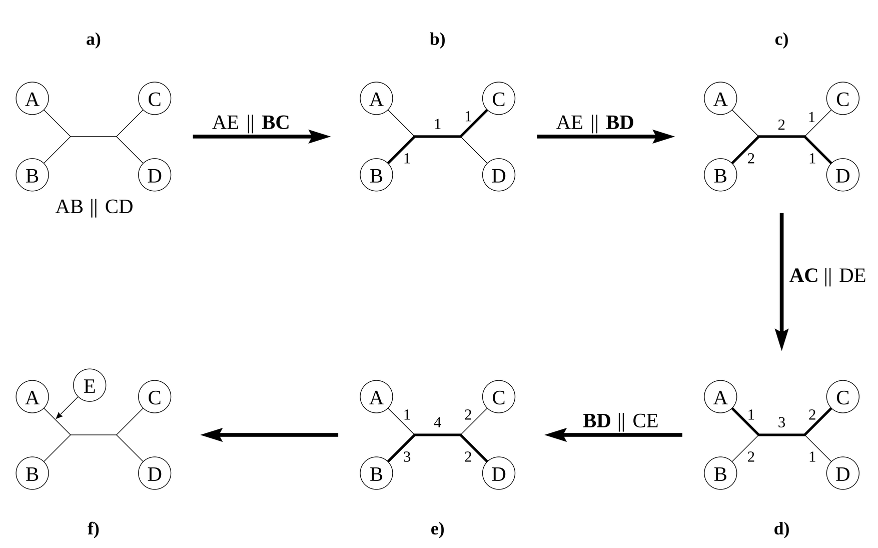
TREE-PUZZLE 5.0 is a program package to reconstruct phylogenetic trees from molecular sequence data using maximum likelihood. TREE-PUZZLE 5.0 is the current version of the PUZZLE program first published by Strimmer & von Haeseler (1996).

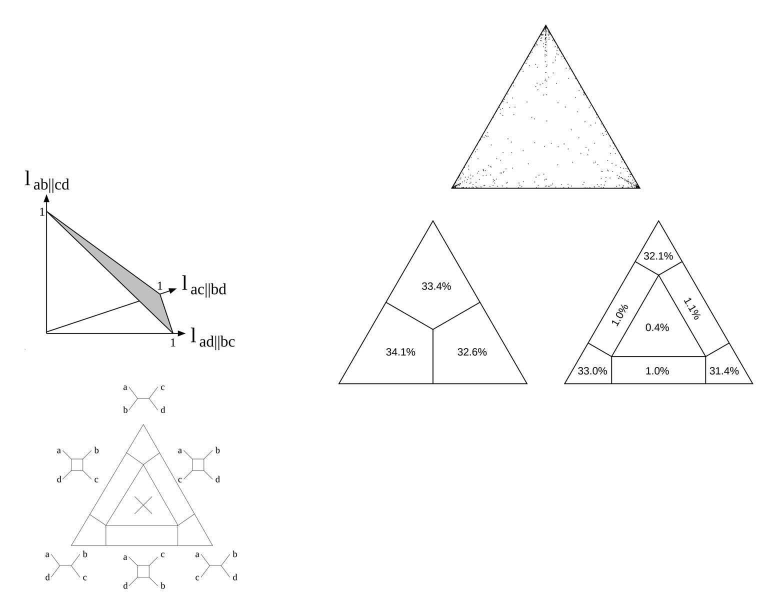
Types of Analysis TREE-PUZZLE allows different types of phylogenetic analysis:

- **Reconstruction of phylogenetic trees** using the quartet puzzling algorithm based on all possible quartets of the input sequences (Strimmer & von Haeseler, 1996).
- **Likelihood mapping analysis** can be used to examine the clustering of specified subgroups of the aligned sequences of to visualize the phylogenetic content or the alignment in general to find out whether the data are suitable for a phylogenetic reconstruction or not (Strimmer *et al.*, 1997).
- **Evaluation of the maximum-likelihood value** of a given tree topology under a given evolutionary model (Felsenstein, 1981).
- **Kishino-Hasegawa test** to compare different tree topologies to find out the most likely ones from a set of given topologies (Kishino & Hasegawa, 1989).

Evolutionary Models PUZZLE also offers a broad variety of evolutionary models:

- **Models for DNA sequences:** TN (Tamura & Nei, 1993), HKY (Hasegawa *et al.*, 1985), F84 (Felsenstein, 1984)
- **Models for protein sequences:** Dayhoff (Dayhoff *et al.*, 1978), JTT (Jones *et al.*, 1992), mtREV24 (Adachi & Hasegawa, 1996), BLOSUM 62 (Henikoff & Henikoff, 1992), VT (Müller & Vingron, 2000), WAG (Whelan & Goldman, 2001)
- **Model for doublets** (pairs of dependent nucleotides): SH (Schöniger & von Haeseler, 1994)
- **Model for binary state data:** F81 (Felsenstein, 1981)
- **Rate heterogeneity:** modeled by a discrete Gamma distribution and by allowing invariable sites.
- The branch lengths can also be calculated under the assumption of a **molecular clock**.





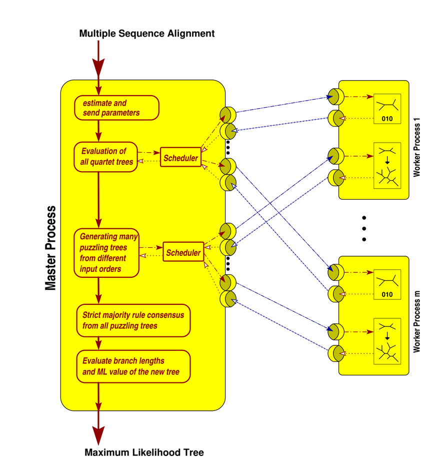


Figure 3: Structure of the parallel implementation of TREE-PUZZLE. The yellow boxes are separate processes in the parallel run.

1 The Quartet Puzzling Algorithm

The quartet puzzling algorithm implemented in TREE-PUZZLE is a quartet based three-step procedure (Strimmer & von Haeseler, 1996).

maximum likelihood step: all $\binom{N}{4}$ quartet maximum likelihood trees are reconstructed in order to find the most likely relationship of every set of four sequences.

puzzling step: these quartet trees are composed into an overall tree adding sequences one by one. The branch a new sequence has to be added is determined by penalizing the paths between sequences distant from the new sequence in the quartet trees. Since this step highly depends on the order of sequences, many trees with different input orders are constructed.

consensus step: a *strict majority rule consensus tree* is built from these puzzle trees.

2 Likelihood Mapping

Quartet trees can also be used to examine the clustering of groups of sequences given by the user. Given four groups of sequences A, B, C, D we draw four sequences a, b, c, d $(a \in A, b \in B, c \in C, d \in D)$ and compute likelihoods values for the informative tree topologies $L_{ab\|cd} = L_1$, $L_{ab\|cd} = L_2$, $L_{ab\|cd} = L_3$. Those are then normalized using Bayes' formular

$$l_i = \frac{L_i}{L_1 + L_2 + L_3}$$

Using (l_1, l_2, l_3) as coordinates all resulting dots fall in a plane (cf. Fig 2a). This plane can be used to examine splits from the dataset. The more dots are near a certain angle of the triangular the more suppurt has this refering topology. This analysis can be done for 4, 3, and 2 groups or with random quartets to examine the phylogenetic information content in the alignment.

3 Parallelization of the Algorithm

To reduce the wallclock time needed by analyses of large datasets, we parallelized the tree reconstruction part of the PUZZLE program to make usable with parallel computers and workstation clusters (COWs).

In our parallelization we used master/worker concepts. This means one master process coordinates the tasks and sends them to the worker processes, which then do the computation work (Fig. 1).

Also we used a variant of the *guided self-scheduling* algorithm (GSS, Hagerup 1997). We adapted the GSS algorithm for the puzzle algorithm, to reduce communication by grouping together tasks, but on the other hand trying to keep all the workers equally busy. For this purpose the scheduling algorithm starts with larger blocks of tasks and ends with a minimum amount per block.

For the parallelization we used the Message Passing Interface (MPI, Snir *et al.* 1998; Gropp *et al.* 1998). Since MPI has become the *de facto* standard in parallel computing using message passing, implementations of the MPI libraries are available on almost all parallel platforms from massively parallel computers to clusters of workstations (COW).

4 Availability

The TREE-PUZZLE 5.0 package is available via Internet from `http://www.tree-puzzle.de`.

The program is written in ANSI C. PUZZLE can be run on UNIX as well as Windows and Mac systems.

To run the parallel version of PUZZLE, an MPI library has to be installed on the system. Free MPI implementations are available on the Web (cf. `http://www.mpi.nd.edu/MPI/`).

References

Adachi, J. & Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**, 459–468.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In: *Atlas of Protein Sequence Structure*, (Dayhoff, M. O., ed) volume 5. National Biomedical Research Foundation Washington DC.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

Felsenstein, J. (1984). Distance methods for inferring phylogenies: A justification. *Evolution,* **38**, 16–24.

Gropp, W., Huss-Lederman, S., Lumsdaine, A., Lusk, E., Nitzberg, B., Saphir, W., & Snir, M. (1998). *MPI: The Complete Reference - The MPI Extensions*, volume 2. Cambridge, Massachusetts: The MIT Press, 2 edition.

Hagerup, T. (1997). Allocating independent tasks to parallel processors: An experimental study. *J. Parallel Distrib. Comput.* **47**, 185–197.

Hasegawa, M., Kishino, H., & Yano, T.-A. (1985). Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA,* **89**, 10915–10919.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282.

Kishino, H. & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179.

Müller, T. & Vingron, M. (2000). Modeling amino acid replacement. *J. Comput. Biol.* **7**, 761–776.

Schöniger, M. & von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* **3**, 240–247.

Snir, M., Otto, S. W., Huss-Lederman, S., Walker, D. W., & Dongarra, J. (1998). *MPI: The Complete Reference - The MPI Core*, volume 1. Cambridge, Massachusetts: The MIT Press, 2 edition.

Strimmer, K., Goldman, N., & von Haeseler, A. (1997). Bayesian probabilities and quartet puzzling. *Mol. Biol. Evol.* **14**, 210–213.

Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: A quartet maximum–likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.

Acknowledgments

This work is supported by *Deutsche Forschungsgemeinschaft* under grants no. VI 160/3-1 and Ha 1628/4-1.