**ABSTRACT**

**Summary:** ProfDist is a user-friendly software package using the profile-neighbor-joining method (PNJ) in inferring phylogenies based on profile distances on DNA or RNA sequences. It is a tool for reconstructing and visualizing large phylogenetic trees providing new and standard features with a special focus on time efficiency, robustness and accuracy.

**Availability:** A Windows version of ProfDist comes with a graphical user interface and is freely available at [http://profdist.bioapps.biozentrum.uni-wuerzburg.de](http://profdist.bioapps.biozentrum.uni-wuerzburg.de).

**Contact:** tobias.mueller@biozentrum.uni-wuerzburg.de

**INTRODUCTION AND FEATURES**

In phylogenetic analysis subclade topologies are often known or easily inferred and well supported by bootstrap analysis. However, we face the problem that basal branching patterns cannot be unambiguously estimated by the usual methods maximum parsimony (Camin and Sokal, 1965), neighbor-joining (Saitou and Nei, 1987), or maximum likelihood (Felsenstein, 1981), nor are they well supported.

The ProfDist-software implements the profile neighbor-joining method (PNJ) as described by Müller et al. (2004). This method enables one to calculate distances between preselected groups of homologous sequences and to reconstruct the corresponding phylogenetic tree. However, beyond this, if one is not willing or not able to preselect monophyletic groups the new ProfDist-software has generalized the PNJ algorithm to deal with an alignment consisting of single sequences and/or groups of sequences, using a robust iterative selection procedure.

ProfDist inherits the accuracy and robustness of profiles and the cubic time complexity of neighbor-joining. The user can represent subclades by a sequence profile and estimate evolutionary distances between profiles to obtain a matrix of distances between subclades.

Based on calculated bootstrap values subclades can be defined automatically, semi-automatically or manually, before a variant of the neighbor-joining algorithm (BIONJ, Gascuel (1997)) is iteratively applied to reconstruct a phylogenetic tree. The robustness of the tree estimation based on profiles is validated by running a variant of the bootstrap procedure (Felsenstein, 1985; Efron et al., 1996). Resulting trees can be visualized by standard tree viewers like ATV (Zmasek and Eddy, 2001), TREEVIEW (Page, 1996), NJplot (Perriere and Gouy, 1996) or hyperformy (De Praetere et al., 2004).

The main feature of ProfDist is the new and efficient implementation of the PNJ algorithm, focusing on the optional iteration and the robust selection of monophyletic sequence families. However, ProfDist combines this new algorithm with standard methods that are used in phylogenetics. All these can be started step by step or independently by the buttons on top of the ProfDist main window (Fig. 1).

To improve the usability, several tree and sequence file formats, e.g., NEWICK, FASTA, EMBL and PHYLIP, are supported.

In the preferences section window, see (Fig. 2), the essential parameters of PNJ are shown and can be set or adjusted: (1) the number of bootstrap replicates, (2) the distance estimation method, (3) a user defined substitution model, (4) the PNJ agglomeration procedure and (5) the path to the preferred tree viewer. Implemented distance estimation methods are JC (Jukes and Cantor, 1969), K2P (Kimura, 1980), GTR (Lanave et al., 1984) and the LogDet transformation (Barry and Hartigan, 1987).

For the iterative PNJ algorithm there are only two driving parameters: (1) the minimal bootstrap value supporting a group of sequences representing a trustworthy monophyletic group which is transformed to a sequence profile, and (2) a percent identity threshold representing an a priori profile to prevent low bootstrap values due to high sequence identities.

As we use memory as well as time efficient algorithms and data structures, PNJ is suitable for reconstructing large phylogenetic trees. In particular, if one knows in advance certain aligned monophyletic sequence groups given in
In the input window of ProfDist the user can load sequence data in FASTA or EMBL format. In the output window the resulting trees of PNJ are shown in NEWICK format.

EMBL-format (e.g., from the European rRNA database (Wuyts et al., 2004)), the time and memory efficiency of PNJ is further improved.

REFERENCES


ACKNOWLEDGMENTS

Special thanks to Annette Wolf, Sven Rahmann and Hannes Luz for valuable discussions and proofreading the paper.