

ARB as a tool for marine biodiversity research



ARB: a comprehensive software package for phylogenetic analysis

By Johanna Wesnig
& Frank Oliver Glöckner

MOLECULAR PHYLOGENY PROVIDES a stable framework to address the diversity of organisms. The understanding of the basic principles of sequence alignment and phylogenetic tree reconstruction is therefore essential. To link phylogeny with community composition and structure in the environment, the application of group- or taxon-specific molecular probes has become a standard technique over the last years.

To facilitate sequence database maintenance, phylogenetic analysis and molecular probe design, the powerful, integrated software package ARB was developed at the Technical University of Munich. After 10 years of continuous improvement it has become a standard tool used by experts all around the world. In order to use the system efficiently, participation in a training workshop is highly recommended. The workshops give a thorough introduction into the theoretical background of phylogenetic tree reconstruction and probe design, followed by a

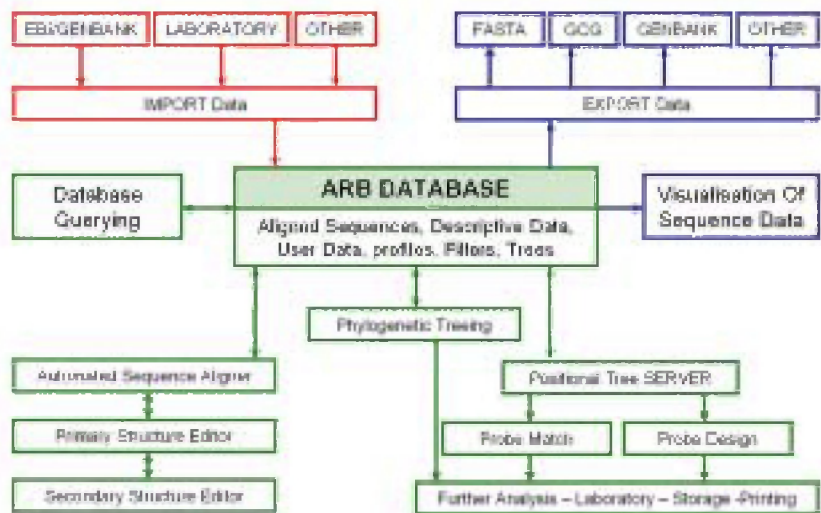


Fig 1. ARB: a software environment for sequence data. Ludwig, W., et al (2004). *Nucleic Acids Research*, 2004, Vol 32, No 4, 1,363-1,371.

'hands-on' experience on phylogenetic tree-making, using the freely available software package ARB.

Description of ARB

ARB (Latin: *arbor*; tree) is a comprehensive software package originally developed for ribosomal RNA (rRNA) data. Nowadays, it can be used for any nucleic- or amino-acid sequence data. The central idea is to arrange a database of sequences and associated descriptive information according to the phylogenetic relationships of the corresponding organisms. This phylogenetic tree is visualised and can be used for walking through the database via simple mouse clicks.

ARB imports raw data and primary structure data from existing sources like GenBank and EBI. For rRNA data, ARB already provides databases with processed primary structures (aligned sequences). Any additional data related to the individual sequences can be stored in structured database fields or linked via local or worldwide networks. Furthermore, to facilitate in-depth analysis of molecular data, a comprehensive selection of software tools is integrated into ARB. These are controlled via a common graphical user interface and they interact directly with one another as well as with the database (see **Figure 1**). Software and start-up databases are freely available at <http://www.arb-home.de>.

Technical details and features:

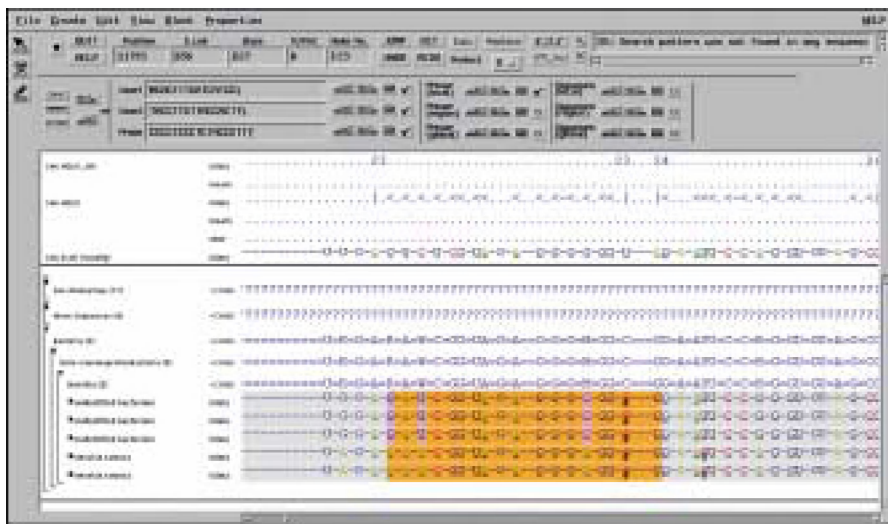
ARB Database

The designation and hierarchy of database fields can be customised by each user. Links to other databases are possible, with several levels of protection facilitating security management. All information stored in the database, along with the sequence data such as bibliography, user-made entries or information calculated on-line from the database entries, can be shown at the terminal nodes of the tree.

Furthermore, all integrated and second-party tools are linked to the ARB database. Thus, any changes or rearrangements are immediately known to all software components without user intervention.

The following tools are integrated:

- Tools for importing and exporting aligned and unaligned nucleic acid and protein sequences in various database formats;
- Tools for defining sequence profiles and for finding and evaluating organism or group specific identifiers (probes or primers);
- Database management with full text search in any database field, as well as complex search options;
- A script language for writing small programmes to facilitate recurring tasks.



Sequence editor and alignment

A powerful sequence editor which can be used for nucleic acid and amino acid sequences allows sequence editing, string search, visualisation of base pairing, positional variability or likelihood. All colours and symbols are user-defined, and separate "align" and "edit" modes prevent erroneous changes. A special feature for rRNA sequences is the simultaneous secondary structure editor which helps the scientist to evaluate probe targets.

The following features are implemented:

- Automated alignment of a sequence according to its next relative (only for nucleic acids);
- Synchronisation of nucleic acid and amino acid sequence alignments;
- Implemented secondary structure editor (only for rRNA);
- Visualisation of patterns and probes.

Phylogenetic reconstruction

For phylogenetic tree-making, several alternative methods are available as part of the package, ranging from simple character



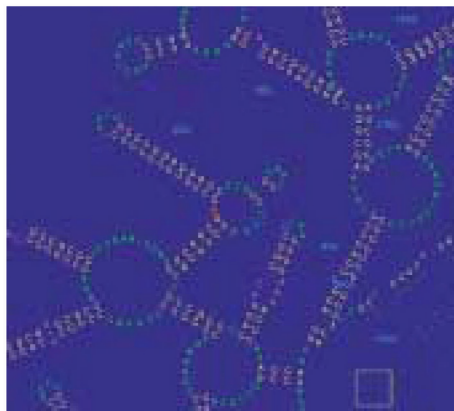
counting to a special maximum parsimony approach that allows the reconstruction and evaluation of big trees. One prominent feature is the possibility to add sequences to an existing tree without changing the overall topology. Optimisation of trees can be applied to the complete or user-selected sub-trees, and intermediate stages can be stored. Various programmes for phylogenetic analyses are directly cooperating within ARB, e.g. some programmes of the PHYLIP package for phylogeny inference are incorporated.

Useful features include:

- Distance matrix, maximum parsimony and maximum likelihood methods;
- Confidence tests (bootstrapping);
- Partial- and full-length sequences can be added to existing trees without altering the general topology;
- Extended possibilities for tree visualisation and optimisation.

Probe design and evaluation

In just three steps, the selection of the target organisms, probe design and probe match is done and can be visualised. For fluorescence in situ hybridisation (FISH), the ARB multi-probe



software is especially useful. A peculiarity is the position tree (PT) server, which has to be established by every user locally. Analyses with the PT server do not rely upon aligned sequences and can be used for fast and reliable probe design and probe check.

The advantages of ARB in this respect are:

- Easy graphical selection of target sequences;
- Fast and reliable probe design and probe check based on suffix trees;
- Graphical visualisation of results in the tree, alignment and secondary structure.

Everybody with an interest in sequence alignment, phylogenetic analysis and/or probe design and hybridisation should be able to profit from using ARB. ARB sequence databases are currently available for small subunit rRNA. Ongoing developments are focusing on:

- Sequence databases for further phylogenetic markers;
- A tool for visualisation of accessibility maps;
- Further extensions so that ARB can handle complete genomes. •

Dr Johanna Wesnigk &
Prof Dr Frank Oliver Glöckner,
Max Planck Institute for Marine
Microbiology, Celsiusstraße,
128359 Bremen, Germany.

Further reading

Amann, R., W. Ludwig, K.-H. Schleifer, (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59, 43-69.

Ludwig W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Foerster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Koenig, T. Liss, R. Luemann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode and K.-H. Schleifer, (2004). ARB: a software environment for sequence data. *Nucleic Acids Research*, 2004, Vol 32, No 4, 363-437.

Ludwig, W. & H.-P. Klenk, (2001). Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Garrity, G. (ed.) *Bergey's Manual of Systematic Bacteriology* (2nd edn). Springer, New York, 49-65.