ARB: a software environment for sequence data

Wolfgang Ludwig*, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier[1], Yadhukumar, Arno Buchner, Tina Lai, Susanne Steppi, Gangolf Jobb[1], Wolfram Förster[1], Igor Brettske[1], Stefan Gerber[1], Anton W. Ginhart[1], Oliver Gross, Silke Grumann[1], Stefan Hermann[1], Ralf Jost[1], Andreas König[1], Thomas Liss[1], Ralph Lüßmann[1], Michael May[1], Björn Nonhoff[1], Boris Reichel[1], Robert Strehlow[1], Alexandros P. Stamatakis[1,] Norbert Stuckmann[1], Alexander Vilbig[1], Michael Lenke[1], Thomas Ludwig[2], Arndt Bode[1]and Karl-Heinz Schleifer.

Lehrstuhl für Mikrobiologie, Technische Universität München, D-853530 Freising Germany, [1]Lehrstuhl für Rechnertechnik und Rechnerorganisation, Parallelrechnerarchitektur, Technische Universität München, D-85748 Garching, Germany, [2]Institut für Informatik, Ruprecht-Karls-Universität Heidelberg, D-69120 Heidelberg, Germany

* To whom correspondence should be addressed. Tel: +8161 71 5451; Fax: +8161 71 5475; Email: ludwig@mikro.biologie.tu-muenchen.de

Abstract

The ARB (arbor, latin: tree) project was initiated almost ten years ago. The ARB program package comprises a variety of directly interacting software tools for sequence database maintenance and analysis which are controlled by a common graphical user interface. Although it was initially designed for ribosomal RNA data it can also be used for any nucleic and amino acid sequence data. A central database contains processed (aligned) primary structure data. Any additional descriptive data can be stored in database fields assigned to the individual sequences or linked via local or world wide network. A phylogenetic tree visualised in the main window can be used for data access and visualisation. The package comprises of furthermore tools for data im- and export, sequence alignment, primary and secondary structure editing, profile and filter calculation, phylogenetic analyses, specific hybridisation probe design and evaluation and other components for data analysis. Currently, the package is more used by the scientific community all over the world.

Introduction

The ARB (arbor, latin: tree) project was established as an interdisciplinary bioinformatics initiative of the Lehrstuhl für Mikrobiologie and the Lehrstuhl für Rechnertechnik und Rechnerorganisation, Parallelrechnerarchitektur of the Technical University of Munich almost ten years ago. In that time, comparative sequence analysis of the small subunit rRNAs or the respective genes already had been established as the most commonly applied approach for phylogeny inference as well as microbial taxonomy and identification. Furthermore, improved and automated sequencing techniques promoted a rapid increase of the number of small subunit rRNA primary structure data available from databases such as GenBank (1) or EBI (European Bioinformatics Institute; 2). However, these databases provide only raw data and additional descriptive information which cannot interactively be extended by the user. Although the RDP (ribosomal database project; 3) and the Antwerpen (4, 5) projects offered data sets of aligned sequences, data handling and analysis remained difficult for scientists applying rRNA based methods. A variety of individual software tools for sequence editing, alignment and phylogenetic analysis was available from the different database projects (1-4) and other sources (6; http://www.gcg.com). However, a comprehensive package of interacting tools was missing. Furthermore, the number of different input and output formats which had to be used reflected the variety of individual software programs which uncomfortably had to be applied sequentially to achieve a comprehensive analysis of molecular data. Unfortunately, a promising initiative - the GDE project (genetic data environment; http://bimas.dcrt.nih.gov/gde_sw.html) – focussing on the development of a common graphical interface for data handling and analysis was not further continued. Against this background, microbiologists and computer scientists at the Technical University of Munich decided to develop their own software package to become capable to properly manage the upcoming data flood.

The two major tasks according to the ARB concept – formulated in the early days of the project and maintained up to now – are (i.) the maintenance of a structured integrative secondary database combining processed primary structures and any type of additional data assigned to the individual sequence entries and (ii.) a comprehensive selection of software tools directly interacting with one another as well as the central database which are controlled via a common graphical interface. Software and rRNA databases are accessible to public (http://www.arb-home.de) and in use worldwide for several years.

Materials and Methods

Sequence data

The raw data used to establish databases and perform data analysis were taken from own sequencing projects, provided by other research groups and periodically retrieved from public databases such as EBI (European Bioinformatics Institute; 2), Genbank (1), RDP (Ribsomal Database Project; 3) and the Antwerpen database on small (4) and large (5) subunit RNA's. Complete releases were downloaded from the latter two locations. The search and retrieval tools of the former two institutions were used to select and download the primary structure and additional information on rRNA or other genes. Furthermore, sequence data determined at the Lehrstuhl für Mikrobiologie of the Technical University of Munich or by other groups were imported and processed.

Operating systems and programming languages

When the project was started scientific software relying on powerful hardware most commonly was developed for UNIX systems and their derivatives. Initially, the respective operating systems from Sun (UNIX) and Digital (ULTRIX) running on Sparc or Alpha workstations, respectively, were used. Nowadays, SuSE LINUX (http://www.suse.com) running on PC's is preferred.

The greater part of the source code was written in C++ and C, some parts in Perl and other script languages. The graphical environment is based upon the Open Motif library.

Integrated external software tools

Functionalities from the GDE project (genetic data environment; http://bimas.dcrt.nih.gov/gde_sw.html) concerning sequence editing, were adopted and implemented in the ARB package. Some programs of the PHYLIP package for phylogeny inference (6) were incorporated as components directly interacting with the central database. Furthermore, fastDNAml (7) and protml of the Molphy package (8), components of the Puzzle package (9) and AxML a new accelerated fastDNAml derivate (10) were included for maximum likelihood based phylogenetic analyses of nucleic and amino acid sequence data.

Results and Discussion

A selection of tools and functionalities of the ARB packages will be briefly described in the following section. The network in Figure 1 schematically visualises these tools and their interactions with one another and the central database. Most tools developed for ARB directly interact with a copy of the database in the main storage, whereas the integrated second party tools are provided with data from ARB and their results are written back to the database. Thus, any changes or rearrangements are immediately known to the peripheral software components.

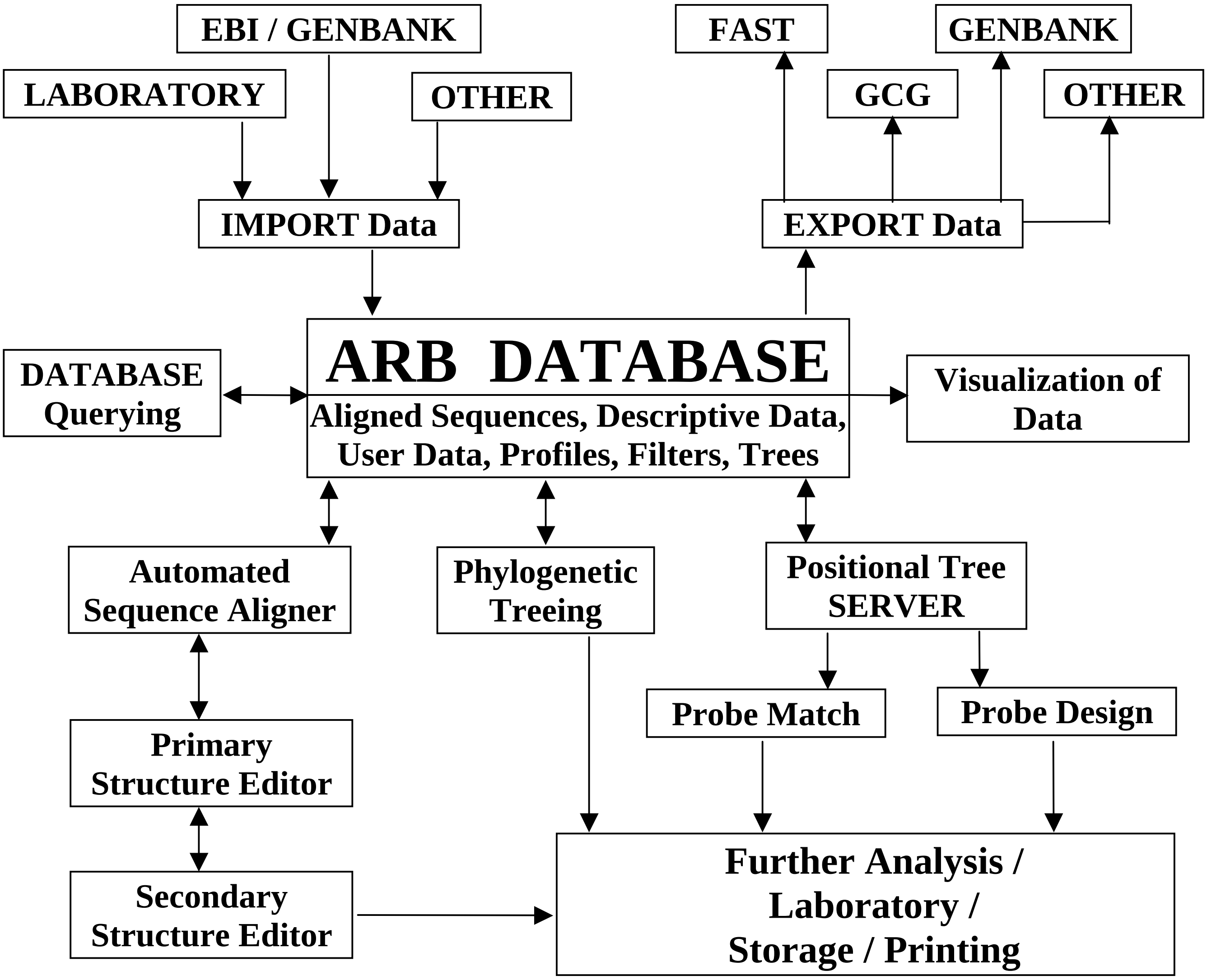


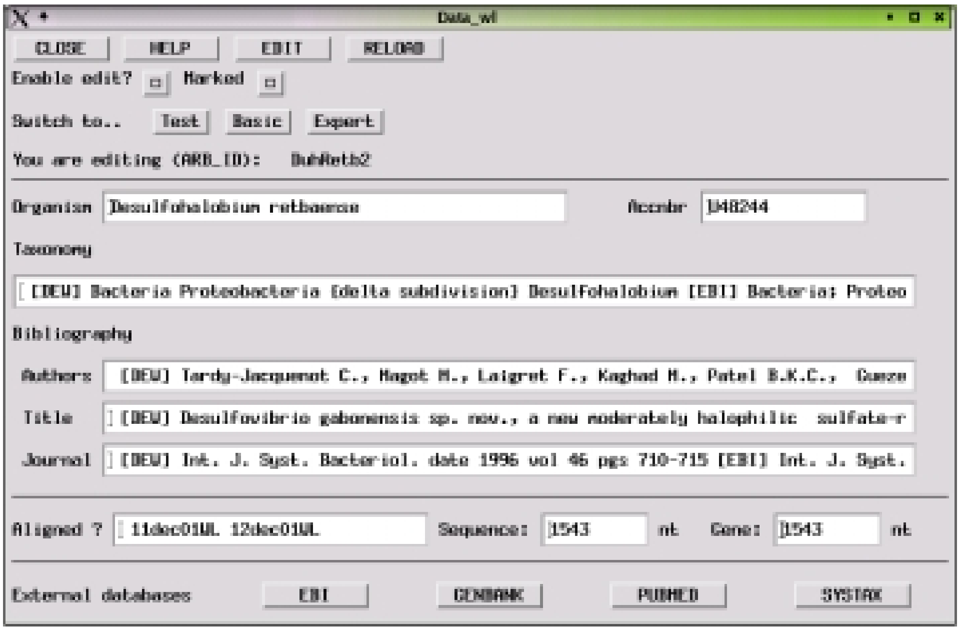
Figure 1. The interacting components and tools of the ARB software package and database.

The central database

The sequences representing organisms, genes or gene products are stored in individual database fields as the central components and an unique identifier (short_name) is automatically generated and assigned to each of them. Databases created by using ARB are hierarchically structured. Following the ARB concept of an integrative data base, any type of additional data can be assigned to the individual sequence data entry and stored within default or user defined database fields. These data can either be kept as intrinsic components of the database or linked to it via local networks or the internet. In the latter case the path to the respective file or the URL of an external data base - optionally including commands and search strings - have to be stored within the respective ARB database fields. The designations and hierarchy of the database fields can be customised by the user. The default structuring is according to the phylogeny of the organisms derived from the respective sequence data. However, it can also be changed according to other criteria defined by database field entries. This hierarchy is used by special algorithms for highly effective data compression. Different protection levels (0-6) can be assigned to the individual database fields. Database as well as security management is facilitated by this tool.

Data access and visualisation

A powerful search tool allows simple (strings and combination of strings) and complex (default or user defined algorithms) searches in one or more of the database fields. The information in all or an user defined selection of database fields can be visualised on the screen in respective windows (Figure 2). The layout of the visualisation windows i.e. selection, size and positioning of database field entries can be customised by the user. Simple algorithms are included.



Enable edit? □ Marked □

You are editing (ARB_ID): DuhRetb2

Organism: Desulfohalobium retbaense

Accnbr: U48244

Taxonomy: [DEW] Bacteria Proteobacteria (delta subdivision) Desulfohalobium [EBI] Bacteria; Proteo

Bibliography

Authors: [DEW] Tardy-Jacquenot C., Magot M., Laigret F., Kaghad M., Patel B.K.C., Gueze

Title: [DEW] Desulfovibrio gabonensis sp. nov., a new moderately halophilic sulfate-r

Journal: [DEW] Int. J. Syst. Bacteriol. date 1996 vol 46 pgs 710-715 [EBI] Int. J. Syst.

Aligned ? 11dec01WL 12dec01WL

Sequence: 1543 nt

Gene: 1543 nt

Figure 2. Example of a data visualisation window. Bibliographic data stored in respective database fields are shown. The selection of database fields, extraction of data as well as the layout of the visualisation window can be customised by the users.

An alternative way of data access and visualisation is provided by the ARB main window. Phylogenetic trees generated by intrinsic ARB tree reconstruction tools or imported from external sources are stored in the database and can be visualised in different formats within the ARB main window (Figure 3). Any (combination of) database field entries can be visualised at the terminal nodes of the tree currently shown. Selection and order of data entries, the results of data analysis or extraction to be visualised are defined by the NDS (node display settings) tool. Irrespective of the visualisation mode used, the ARB SRT (search and replacement tool) and ACI (command interpreter) can be used for extraction of (combinations of) (sub)strings as well as for analysis of database field entries, respectively.

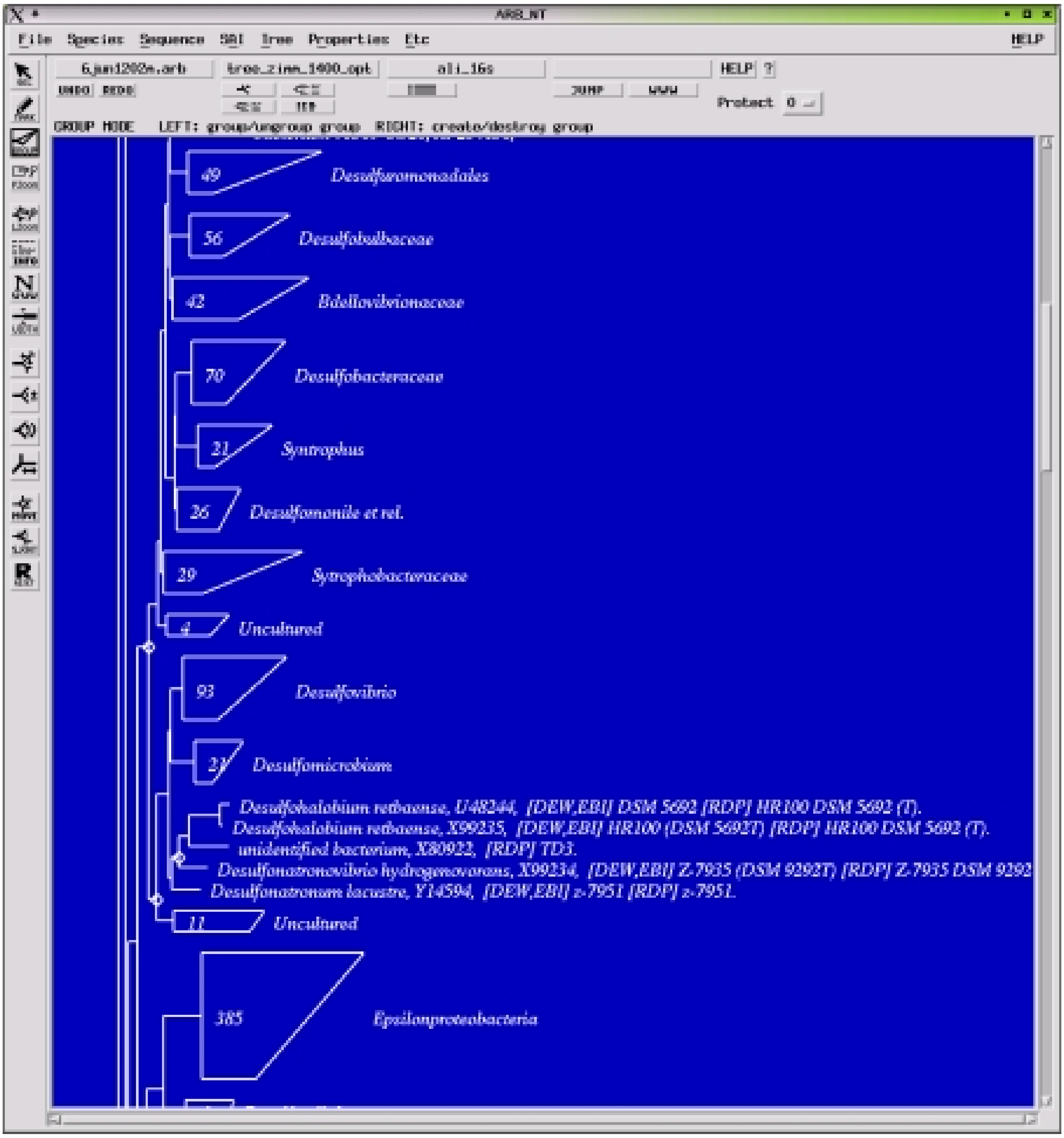

ARB_NT
File
Species
Sequence
SAI
Tree
Properties
Etc
HELP
6jun1202m.arb
tree_zimm_1400_opt
ali_16s
HELP
?
UNDO
REDO
JUMP
WWW
Protect
0
GROUP MODE   LEFT: group/ungroup group   RIGHT: create/destroy group
49
Desulfuromonadales
56
Desulfobulbaceae
42
Bdellovibrionaceae
70
Desulfobacteraceae
21
Syntrophus
26
Desulfomonile et rel.
29
Sytrophobacteraceae
4
Uncultured
93
Desulfovibrio
21
Desulfomicrobium
Desulfohalobium retbaense, U48244, [DEW,EBI] DSM 5692 [RDP] HR100 DSM 5692 (T).
Desulfohalobium retbaense, X99235, [DEW,EBI] HR100 (DSM 5692T) [RDP] HR100 DSM 5692 (T).
unidentified bacterium, X80922, [RDP] TD3.
Desulfonatronovibrio hydrogenovorans, X99234, [DEW,EBI] Z-7935 (DSM 9292T) [RDP] Z-7935 DSM 9292
Desulfonatronum lacustre, Y14594, [DEW,EBI] z-7951 [RDP] z-7951.
11
Uncultured
385
Epsilonproteobacteria

Figure 3. The ARB main window showing part of an ARB-parsimony generated dendrogram. The rectangles represent ‘online compressed’ monophyletic groups which can be ‘unfolded’ by mouse click. Database field entries such as taxonomic name, public database accession number and strain designation as reported in EBI (2), RDP (3) and the European rRNA databases (DEW; 4,5) are visualised at the terminal nodes of the ‘unfolded’ Desulfohalobiaceae.

Sequence editors

The sequence data can be visualised and modified with a powerful editor (Figure 4). The original data as well as virtually transformed (e.g. purine-pyrimidine or simplified amino acid presentation) data are displayed in user defined colour codes. Keyboard customisation is possible for data entry and modification. Two different editing modes can be selected: the ‘Align’ and ‘Edit’ modes, respectively. The ‘Align’ mode allows only to insert/remove alignment gaps and to move sequence characters. In addition to these functions, character changes can be performed in the ‘Edit’ mode. The rights to overcome protection of the individual sequence

entries can be given for the two modes independently. This helps to prevent unwanted character changes when manually modifying the sequence data or alignment.
Sets of search strings can be defined and optionally stored. Their occurrence can be visualised within the displayed sequences by user defined background colours. Virtual compression – removal of alignment gaps common to all or a certain fraction of the displayed sequences - is possible. This helps to a more convenient data handling in case of large insertions occurring in only part of the selected sequences. Groups of sequences can be interactively defined or are automatically shown if defined in the pylogenetic trees. Consensus sequences are determined for each defined group of sequences according to default or user defined criteria and optionally visualised along with or instead of the individual sequences. This consensus is editable and changes made concern any sequence in the group. A special feature of the editor is the simultanous secondary structure check if rRNA (gene) data are visualised. Symbols indicating the presence or absence as well as the character of base pairing are shown below the individual characters and immediately refreshed during sequence editing. A (3-domain) consensus secondary structure mask which was established according to commonly accepted secondary structure models (11) functions as a guide for this tool. Thereby, the users are greatly supported with regard to the evaluation of sequences, alignment and probe targets.

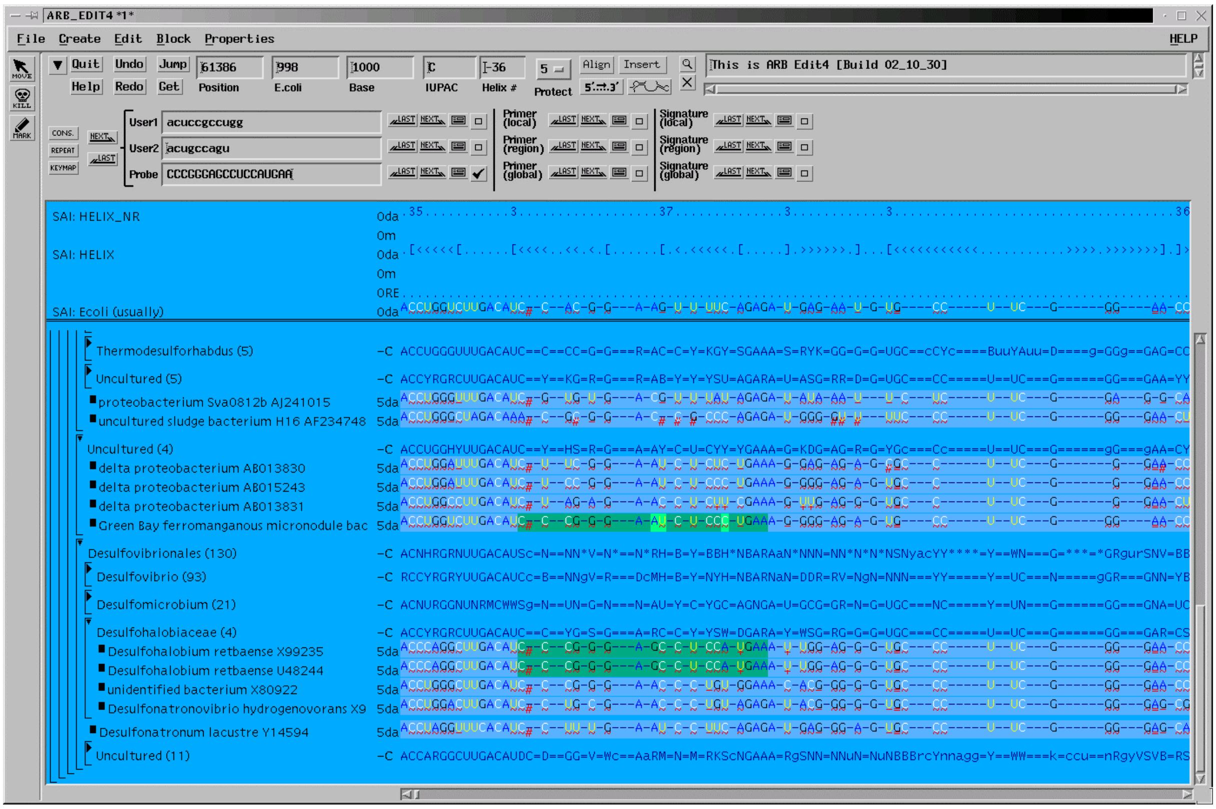
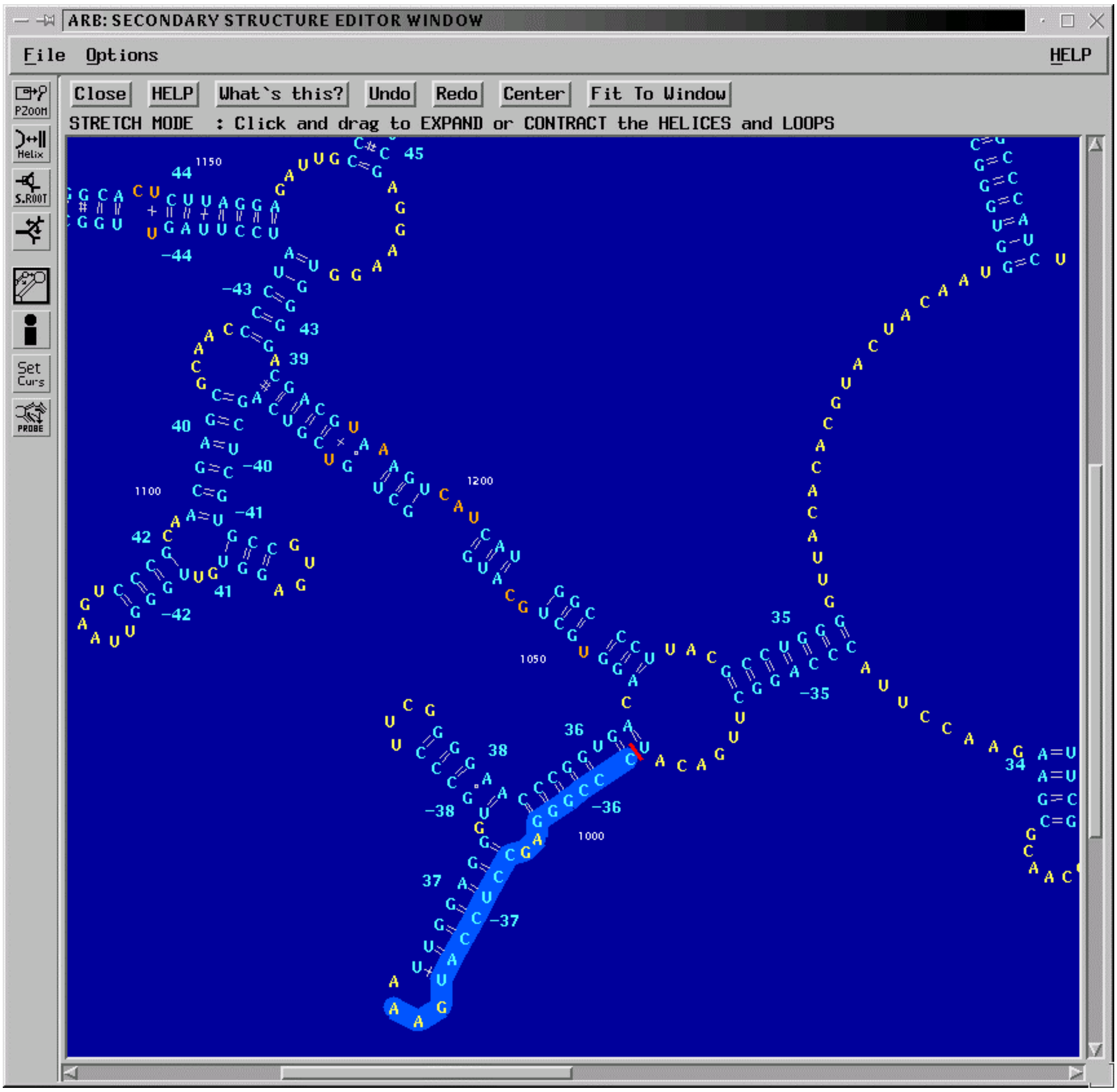


Figure 4. The ARB primary structure editor. As an example for highlighting a search string a probe target site is shown by background colour. Perfect and mismatched pairing is colour coded as well.

The ARB secondary structure editor (Figure 5) fits any sequence into the common consensus model. The particular sequence to be visualised is selected by cursor positioning in the primary structure editor. The layout of the structure i.e. colour coding of base paired, non-paired and loop positions as well as probe target sites can be customised according to the users preferences. Any of the search strings activated in the primary structure editor can be indicated in the secondary structure model. This helps the experts to evaluate probe targets. The evaluation of target position with respect to higher order rRNA structure is of importance especially when probes should be used for in situ cell hybridisations (12). The structure can be exported to xfig - a simple open source graphics program (http://ww.xfig.org) - for further modification and/or transformation into various formats.



STRETCH MODE : Click and drag to EXPAND or CONTRACT the HELICES and LOOPS

Figure 5. Secondary structure editor. The sequence selected in the primary structure editor (Figure 4) is automatically fitted into a consensus secondary structure model.

Profiles, masks and filters

Conservation or base composition profiles, higher order structure masks as well as filters including or excluding particular alignment positions are important tools for sequence data analyses, especially for phylogeny inference (13). The ARB package provides tools for determining such profiles based upon the full database or user defined sub sets. The underlying methods range from simple character counting to maximum parsimony based column statistics. These profiles, masks and filters are stored in the central database as so called SAIs (sequence associated informations) and can be visualised and modified by the primary structure editor. The filter selection tool does not only allow to choose sets of particular filters but also to perform a fine tuning with respect to the inclusion or exclusion of alignment positions in case of multiple character filters.

Phylogenetic treeing

As mentioned in the Materials and Methods section, software implementations of several alternative treeing methods are incorporated in the package. They operate as intrinsic tools with all the respective ARB components and database elements such as alignment and filters. The central treeing tool of the package – ARB-parsimony – is a special development for the handling of several thousand sequences (more than 30.000 in the current small subunit rRNA ARB database). New sequences are successively added to an existing tree according to the parsimony criterion. An intrinsic software component superimposes branch length to the parsimony generated tree topology. These branch lengths reflect the significance of the individual 'tetra-furcations' by expressing the difference of the most and the two less parsimonious solutions when performing NNI (nearest neighbour interchange of adjacent branches or sub trees). These relative distances are standardised according to a distance matrix deduced from primary structure comparison. Thus branch lengths in ARB-parsimony generated trees in first instance visualize the significance of topologies, in second instance reflect a degree of estimated sequence divergence. Given the limited information content of any potential phylogenetic marker, treeing should generally be based upon full sequences only (13). Intermixing of full and partial sequence data obviously destroys the quality of a tree irrespective of the method applied. A prominent feature of ARB-parsimony is the possibility to add sequences to an existing tree without allowing any changes in the initial tree. This enables the user to reconstruct and optimise an initial tree based upon the best (full sequences) and most comprehensive (wide variation of phylogenetic levels) sequence data and also to include partial sequences without destroying the optimised tree. The second peculiarity of the treeing software concerns the tree optimisation performing cycles of NNI (nearest neighbour interchange) and KL (Kernigham and Lin, 14) tree modifications. This optimisation can not only be applied to the complete tree but confined to user selected sub trees. Thus tree optimisation is possible applying the appropriate filters for the respective phylogenetic levels and groups. In this context, it is of interest that performing stepwise optimisations the intermediates are stored until the user defines the version to be permanently stored in the database. Furthermore, different trees generated applying various parameters can be permanently kept in the database and optionally used for data visualisation in the ARB main window.

The positional tree server

The ARB PT-server (positional tree) once established allows rapid finding of sequence identity or peculiarity. Thus, it is the central tool for fast searching of closest relatives for automated sequence alignment or to define diagnostic sequence stretches for primer and probe design. Establishing a prefix tree server of any oligonucleotide sequence up to 20-mers occurring in the underlying database and assignment of the individual oligonucleotides to the sequences or organisms containing them is the basis for these procedures. PT-server based analyses do not rely upon aligned sequences. The advantages of these logarithmic algorithms over linear ones such as Blast (15) or Fasta (16) are the effectiveness and rapidity.

Sequence alignment

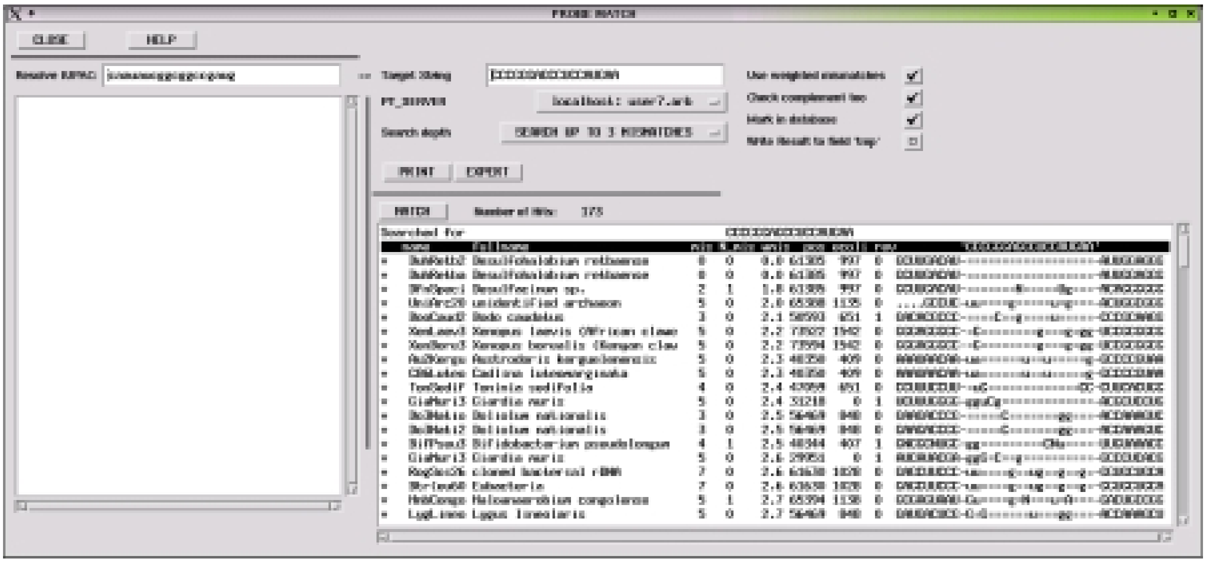
As mentioned in the materials and methods section, for de novo generating a nucleic or amino acid sequence alignment ClustalW (17) as implemented in the ARB package can be used. However, in most cases new sequence entries have to be integrated in an already existing database of aligned sequences. For this purpose the ARB-fast-aligner was developed and included. This aligner uses a (set of) selected aligned reference sequences as template(s) for rapid integration of a (set of) unaligned sequence(s). Individual entries i.e. sequences or consensus defined by the user or automatically determined by PT-server based search for most similar reference sequences are used as template.

In case of protein coding nucleic acid sequences the alignment usually is optimised on the amino acid level. The underlying nucleic acid alignment then can be adapted to the amino acid alignment by a back-translation based tool taking into consideration all known codon usages.

Probe design and evaluation

Taxon or gene specific probe design nowadays certainly plays a central role in many molecular biological researche and analysis projects be it for example the identification and detection of organisms in complex environmental samples or expression studies within the scope of genome projects, respectively. Algorithms of the ARB programs ‘Probe Design’ and ‘Probe Match’ are searching the PT-server to identify short (10 - 100 monomers) diagnostic sequence stretches which are evaluated against the background of all full and partial sequences in the respective database the PT-server has been built from. In principle, no alignment of the sequence data is needed for specific probe design. However, in case of taxon specific probes alignment and phylogenetic analyses are necessary to allow defining groups of phylogenetically (taxonomically) related organisms as the targets of specific probes. The design of taxon specific oligonucleotide probes with ARB is performed in three steps. Firstly, the user selects the organism or a group of organisms for which he wants to design a diagnostic probe. Secondly, the software ‘Probe Design’ searches the PT-server for potential target sites. The results are shown in a ranked list of proposed targets, probes and additional information. The ranking is according to several compositional and thermodynamical criteria (18 - 20). Thirdly, the proposed oligonucleotide probes are evaluated against the whole database by using the program "Probe Match". Local alignments are determined between the probe target sequence(s) and the most similar reference sequences (optionally from 0 to 5 mismatches) in the respective database (Figure 6). Furthermore, these sequence strings can automatically be visualised in the primary and secondary structure editors. Especially the latter information is of high importance when designing probes for in situ cell hybridisation.

A special advancement is the ARB multi probe software component. It determines sets of up to five probes optimally identifying the target group. These probe set can be used for multiple fluorescence in situ hybridisation experiments.



PT_SERVER localhost: user7.arb
Search depth SEARCH UP TO 3 MISMATCHES
Use weighted mismatches ✓
Check complement too ✓
Mark in database ✓
Write Result to field 'tmp' □
Number of Hits: 173

```
*  DshRetb2 Desulfohalobium retbaense       0  0  0.0 61305  997  0
*  DshRetba Desulfohalobium retbaense       0  0  0.0 61305  997  0
*  DfnSpec1 Desulfocinum sp.                2  1  1.0 61305  997  0
*  UnidArc28 unidentified archaeon          5  0  2.0 65308 1135  0
*  BodCaud2 Bodo caudatus                   3  0  2.1 58593  651  1
*  XenLaev3 Xenopus laevis (African clawe   5  0  2.2 73927 1942  0
*  XenBore3 Xenopus borealis (Kenyan claw   5  0  2.2 73994 1942  0
*  AuKergu Austrodoris kerguelenensis       5  0  2.3 48358  409  0
*  CdlLuteo Cadlina luteomarginata          5  0  2.3 48358  409  0
*  TonSedif Tonicia sedifolia               4  0  2.4 47059  651  0
*  GiaMuri3 Giardia muris                   5  0  2.4 33218    0  1
*  DolNatio Doliolum nationalis             3  0  2.5 56469  848  0
*  DolNati2 Doliolum nationalis             3  0  2.5 56469  848  0
*  BifPseu3 Bifidobacterium pseudolongum    4  1  2.5 48344  407  1
*  GiaMuri3 Giardia muris                   5  0  2.6 29951    0  1
*  [illegible] cloned bacterial rRNA        7  0  2.6 63638 1028  0
*  [illegible] [illegible]                  7  0  2.6 63658 1028  0
*  HabCongo Haloanaerobium congolense       5  1  2.7 65594 1138  0
*  LygLinea Lygus lineolaris                5  0  2.7 56469  848  0
```

Figure 6. Results of probe design and evaluation. Part of the primary structure alignment containing the probe target site is shown for the target organism *Desulfohalobium retbaense* and the non target organisms containing the most similar sequence stretches.

Data im- and export
The sequence as wells as additional data can be im- and exported in commonly used flat file formats. The parsing from and to tagged flat files can be customised by advanced users. There is also contained a tool for automated completion of database submission forms for those users determining sequences by their own.

Availability and documentation
Although so far not officially published, previous versions of the software package and databases had been available since several years and the software is in use worldwide. Self installing program versions, the source code as well as some documentation are available at http://www.arb-home.de. HTTP Browsers have to be used, ftp connection is not accepted. Furthermore, there is an e-mail forum of the world wide ARB users community. Subscription is needed for those who are interested to join (subscribe@arb-home.de). ARB-sequence databases are currently available for the small subunit rRNA and those for other conserved genes will be provided soon. Checking for new releases and updates should be done at http:www.arb-home.de.

System and hardware requirements
The ARB group provides tested versions for SuSE LINUX and Sun Solaris systems. According to user provided information, the LINUX version is also running on Redhat and Mandrake LINUX systems. For running ARB on Mac OSX see http://www.microbiol.unimelb.edu.au/micro/staff/mds/ARB_OSX/ARB_to_MacOSX.html.
With respect to hardware requirements main frame memory is more important than processor performance. The users among the wet lab partners of the ARB group are performing their analyses on dual pentium III PCs with 1 Gb memory and 1 Gb swap space. 21' monitors at 1600 x 1200 are recommended. However, ARB is also routinely used on laptops or older PC's and workstations with less memory and monitors with lower resolution.

Future developments
The ongoing developments are focussing on two major tasks: 1. a web tool providing all potential probe target sites which can be derived from the current data base version and should phylogenetically (taxonomically) make sense. The user can not only search for hierarchical and multiple probes submitting organisms names, strain designations or accession numbers as search strings, but also may send his own probe sequences for *in silico* evaluation.
2. the package is adopted for handling and analysing databases of completed and annotated genomes. All ARB functionalities can be applied and genome maps can be used for visualisation and data access. In accordance with the ARB concept of integrative databases experimental parameters and data can be stored and assigned to the individual genomes or genes.
Many users ask for a windows compatible ARB version. Although comprehensive software redesign would be desirable, the current capacity and funding of the ARB group does not allow doing this in reasonable time with a source code developed by many individual scientists and programmers.

Literature

1. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L. (2002) GenBank. *Nucleic Acids Res*., 30, 17-20.
2. Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli,

M.A., Tzouvara, K. and Voughan, R. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res*., 30, 21-26.

3. Maidak, B.L., Cole, J.R., Lilburn, T.G., Parker, C.T., jr., Saxman, P.R., Farris, R.J., Garrity, G.M., Olsen, G.J., Schmidt, T.M., Tiedje, J.M. (2001) The RDP-II (ribosomal database project) *Nucleic Acids Res*., 29, 173-174.
4. Wuyts, J., Van de Peer, Y., Winkelmans, T., and De Wachter, R. (2002) The european database on small subunit ribosomal RNA. *Nucleic Acids Res*., 30, 183-185.
5. Wuyts, J., De Rijk, P., Van de Peer, Y., Winkelmans, T., and De Wachter, R. (2002) The european large subunit ribosomal RNA database. *Nucleic Acids Res*., 30, 175-177.
6. Felsenstein, J., (1989) PHYLIP – phylogeny inference package (version 3.2). *Cladistics,* 5, 164-166.
7. Olsen, G.J., Matsuda, H., Hagstrom, R. and Overbeek, R. (1994) FastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci*., 10, 41-48.
8. Adachi, J. and hasegawa, M. (1996) Molphy version 2.3, programs for molecular phylogenetics based on maximum likelihood. *Technical report*. The Institute of Statistical Mathematics.
9. Strimmer, K. and von Haeseler, A. (1996) Quartett Puzzling: a quartett maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol*., 13, 964-969.
10. Stamatakis, A.P., Ludwig, T., Meier, H., and Wolf, M.J.. Accelerating Parallel Maximum Likelihood-based Phylogenetic Tree Calculations using Subtree Equality Vectors. In: Proceedings of Supercomputing Conference (SC2002), Baltimore, Maryland, in press.
11. Cannone, J.J, Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Müller, K.M., Pande, N., Shang, Z., Yu, N., and Gutell, R.R. (2002). The comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BioMed Central Bioinformatics*. 3, 2.
12. Amann, R., Ludwig, W. and Schleifer, K.H. (1995) Phylogentic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev*. 59, 143-169
13. Ludwig, W. and Klenk, H.P. (2001) Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In Garrity, G. (ed.) *Bergey's Manual of Systematic Bacteriology,* second edition, New York: Springer, pp. 49-65.
14. Kernigham, B.W., and Lin, S. (1970) An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal,* 49, 291-307.
15. Altschul, S.F., Madden, T.L., Schaffer, A.A., Znang, J., Znang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-Blast: a new generation of protein-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., 25, 3389-3402.
16. Pearson, W.R. and Lipman, D.C. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad,. Sci. U.S.A.* 85, 2444-2448.
17. Thompson, J.D., Higgins, D.G. and Gibson, D.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment. *Computer Applications in the Biosciences,* 8, 189-191.
18. Amann, R., Ludwig, W. and Schleifer, K.H. (1995) Phylogentic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev*. 59, 143-16.
19. Amann, R. and Ludwig, W. (2000) Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS Microbiol. Rev*. 24, 555–565.
20. Ludwig, W., Amann, R., Martinez-Romero, E., Schönhuber, W., Bauer, S., Neef, A. and Schleifer, K.H. (1998) rRNA based identification systems for rhizobia and other bacteria. *Plant and Soil,* 204, 1-9.

Acknowledgements

ARB software development and database maintenance was partly supported by the European Union within the HRAMI project, by the German Research Foundation microbial ecology and the German Ministry of Research and Education BIOLOG project. The authors highly acknowledge F.O. Glöckner and R. Amann for redistributing the ARB database and software, answering user queries as well as for hosting and organising ARB workshops.