Published online February 25, 2004
本页已使用福昕阅读器进行编辑。
福昕软件(C)2005-2009，版权所有，
仅供试用。
Nucleic Acids Research, 2004, Vol. 32, No. 4 ***1363–1371***
DOI: 10.1093/nar/gkh293

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, Stefan Gerber, 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 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

Received January 13, 2004; Revised and Accepted January 28, 2004

ABSTRACT

The ARB (from Latin *arbor*, tree) project was initiated almost 10 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 be used for any nucleic and amino acid sequence data as well. 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 worldwide networks. A phylogenetic tree visualized in the main window can be used for data access and visualization. The package comprises additional tools for data import and export, sequence alignment, primary and secondary structure editing, profile and filter calculation, phylogenetic analyses, specific hybridization probe design and evaluation and other components for data analysis. Currently, the package is used by numerous working groups worldwide.

INTRODUCTION

The ARB (from Latin *arbor*, 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 10 years ago. In that time, comparative sequence analysis of the small subunit rRNAs or the respective genes had already 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 in the number of small subunit rRNA primary structure data available from data sources 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 Ribosomal Database Project (RDP) (3) and the Antwerpen projects (4,5) offered datasets 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 were 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 Genetic Data Environment (GDE) project (http://bimas.dcrt.nih.gov/gde_sw.html), focusing on the development of a common graphical interface for data handling and analysis was not continued. Consequently, microbiologists and computer scientists at the Technical University of Munich decided to develop their own software package capable of properly managing the upcoming data flood.

The two major tasks according to the ARB concept, formulated in the early days of the project and maintained to the present, 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 publicly

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

本页已使用福昕阅读器进行编辑。
福昕软件(C)2005-2009，版权所有，
仅供试用。

accessible (http://www.arb-home.de) and have been 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 our own sequencing projects, provided by other research groups or periodically retrieved from public data sources such as the EBI (1), Genbank (2), the RDP (3) and the Antwerpen databases for small (4) and large (5) subunit RNAs. 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

The ARB software was developed for UNIX systems and their derivatives. Currently, the development is performed using SuSE LINUX (http://www.suse.com) running on PCs.

The greater part of the source code was written in C++ and C; some parts were written 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 (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. Additionally, fastDNAml (7) and protml of the Molphy package (8), components of the Puzzle package (9) and AxML, a new accelerated fastDNAml derivative (10), were included for maximum-likelihood-based phylogenetic analyses of nucleic and amino acid sequence data.

看到这里

RESULTS AND DISCUSSION

A selection of tools and functionalities of the ARB packages will be briefly described in the following sections. The network in Figure 1 schematically visualizes 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.

The central database

The sequences representing organisms, genes or gene products are stored in individual database fields as the central components and a unique identifier (short_name) is automatically generated and assigned to each of them. Databases created using ARB are hierarchically structured. Following the ARB concept of an integrative database, any type of additional data

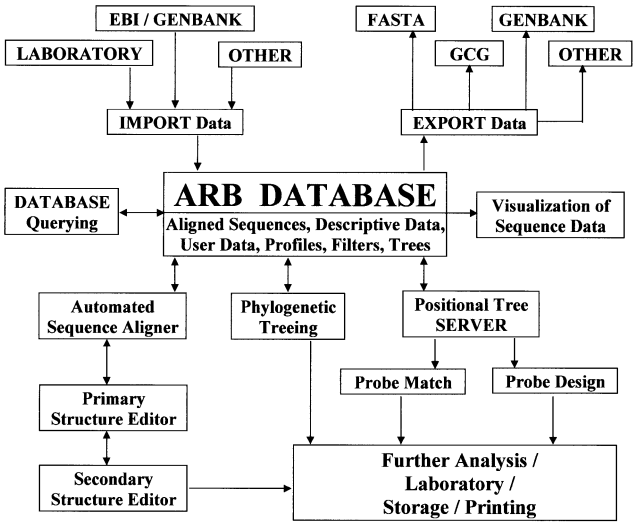

Figure 1. The interacting components and tools of the ARB software package and database.

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 database, 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 customized 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 visualization

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 a user-defined selection of database fields can be visualized on the screen in respective windows (Fig. 2). The layout of the visualization windows, i.e. selection, size and positioning of database field entries, can be customized by the user. Simple algorithms are included.

An alternative method of data access and visualization 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 visualized in different formats within the ARB main window (Fig. 3). Any (combination of) database field entries can be visualized at the terminal nodes of the tree currently shown. Selection and order of data entries, the results of data analysis or extractions to be visualized are defined by the node display settings (NDS) tool. Irrespective of the visualization mode used, the ARB search and

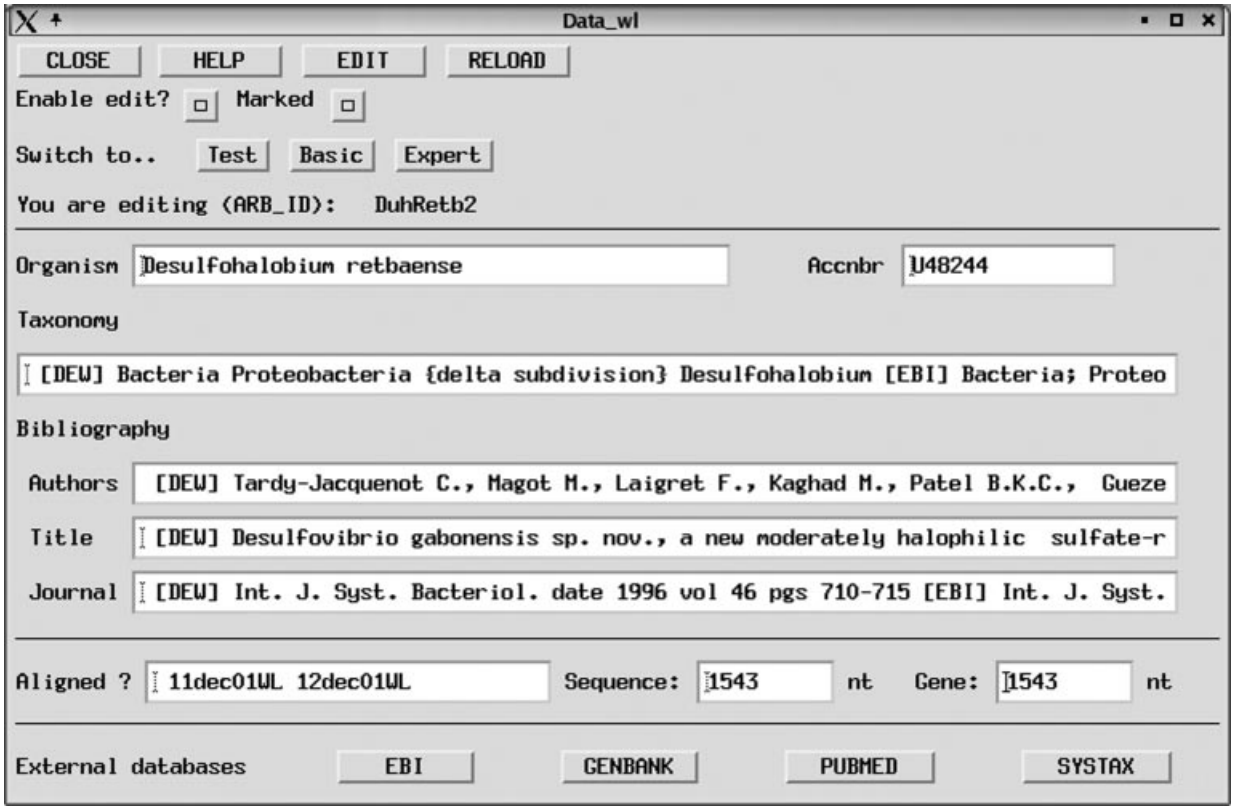

Figure 2. Example of a data visualization window. Bibliographic data stored in respective database fields are shown. The selection of database fields, extraction of data and the layout of the visualization window can be customized by the user.

replacement tool (SRT) and ARB command interpreter (ACI) can be used for extraction of combinations of (sub)strings as well as for analysis of database field entries, respectively.

Sequence editors

The sequence data can be visualized and modified with a powerful editor (Fig. 4). The original data as well as virtually transformed data (e.g. purine–pyrimidine or simplified amino acid presentation) are displayed in user-defined color codes. Keyboard customization is possible for data entry and modification. Two different editing modes can be selected. The 'Align' mode allows only insertion/removal of alignment gaps and movement of 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 visualized within the displayed sequences by user-defined background colors. Virtual compression (removal of alignment gaps common to all or a certain fraction of the displayed sequences) is possible. This makes data handling more convenient in the 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 phylogenetic trees. Consensus sequences are determined for each defined group of sequences according to default or user-defined criteria and optionally visualized along with or instead of the individual sequences. This consensus can be edited and changes made concern any sequence in the group. A special feature of the editor is the simultaneous secondary structure check if rRNA (gene) data are visualized. 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 (three-domain) consensus secondary structure mask established according to commonly accepted secondary structure models (11) functions as a guide for this tool. Thereby the users are strongly supported with regard to the evaluation of sequences, alignment and probe targets.

The ARB secondary structure editor (Fig. 5) fits any sequence into the common consensus model. The particular sequence to be visualized is selected by cursor positioning in the primary structure editor. The layout of the structure, i.e. color coding of base-paired, non-paired and loop positions as well as probe target sites, can be customized according to the user's 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 are used for *in situ* cell hybridizations (12–14). 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.

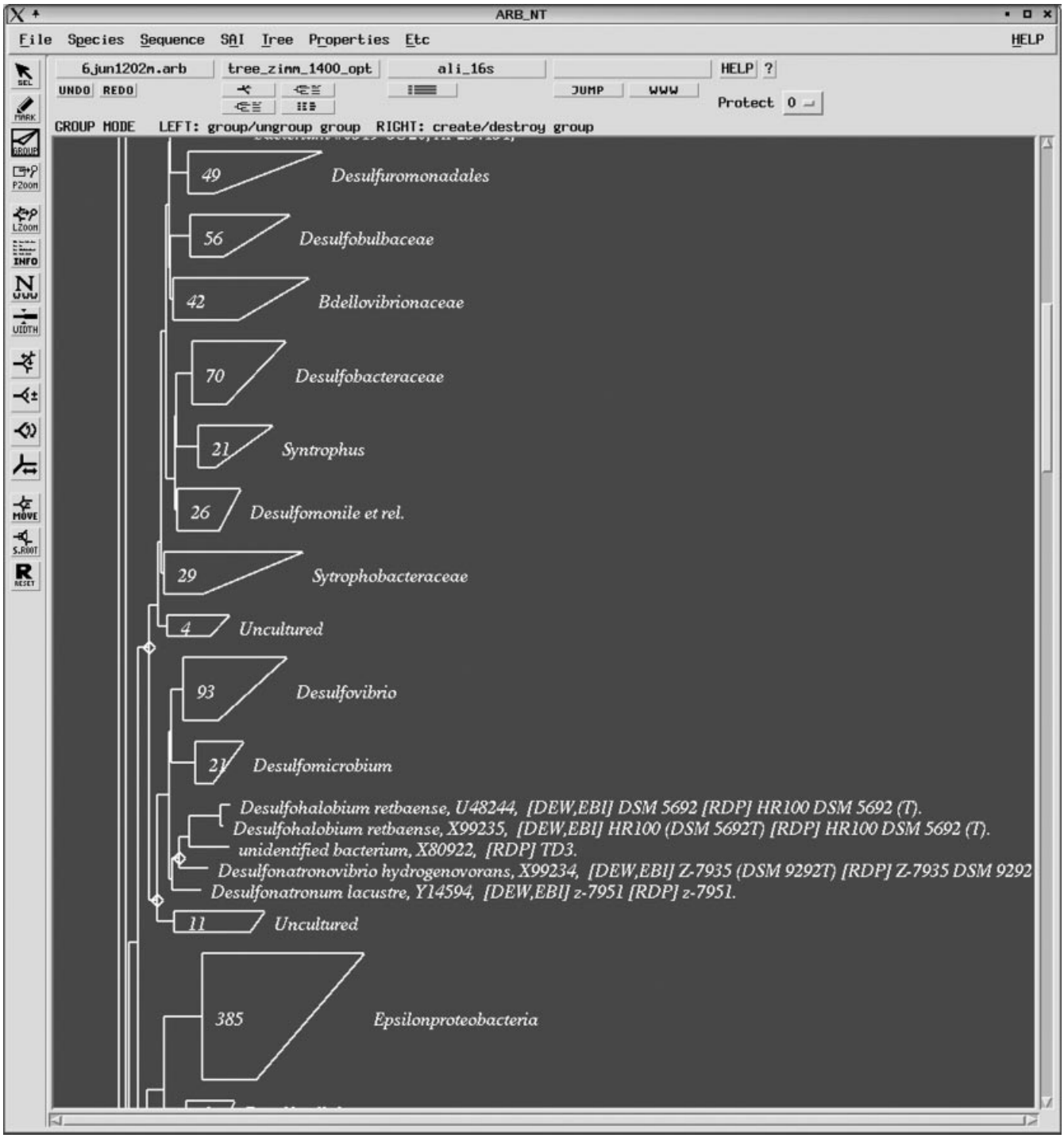

Figure 3. The ARB main window showing part of an ARB parsimony-generated dendrogram. The rectangles represent 'online compressed' monophyletic groups which can be 'unfolded' by mouse click. Database field entries such as taxonomic name, public database accession number and strain designation as reported in EMBL (1), RDP (3) and the European rRNA databases (DEW) (4,5) are visualized at the terminal nodes of the 'unfolded' Desulfohalobiaceae.

Profiles, masks and filters

Conservation or base composition profiles, higher-order structure masks and filters including or excluding particular alignment positions are important tools for sequence data analyses, especially for phylogenetic inference (15). The ARB package provides tools for determining such profiles based upon the full database or user-defined subsets. 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 sequence associated information (SAI) and can be visualized and modified by the primary structure editor. The filter selection tool allows not only choice of sets of particular filters but also

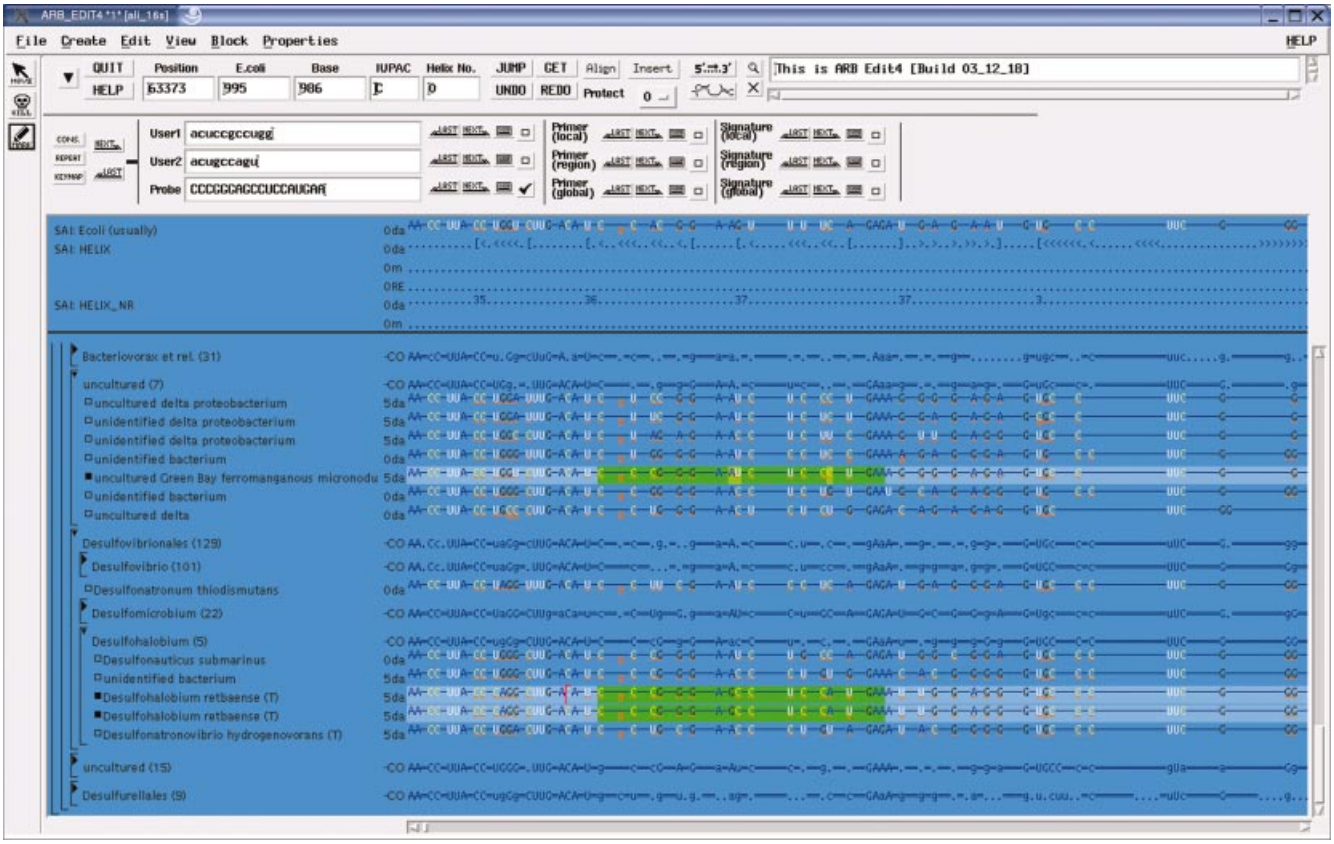

Figure 4. The ARB primary structure editor. As an example for highlighting a search string a probe target site is shown by background color. Perfect and mismatched pairing is color coded as well.

performance of fine tuning with respect to the inclusion or exclusion of alignment positions in the case of multiple character filters.

Phylogenetic treeing

As mentioned in Materials and Methods, 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 on the parsimony-generated tree topology. These branch lengths reflect the significance of the individual 'tetra-furcations' by expressing the difference of the most parsimonious and the two least parsimonious solutions when performing nearest-neighbor interchange (NNI) of adjacent branches or subtrees. These relative distances are standardized according to a distance matrix deduced from primary structure comparison. Thus branch lengths in ARB-parsimony generated trees in the first instance visualize the significance of topologies, and in the second instance reflect a degree of estimated sequence divergence. A prominent feature of ARB-parsimony is the possibility of adding sequences to an existing tree without allowing any changes in the initial tree. This enables the user to reconstruct and optimize 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 perturbing the initial tree topology. The second peculiarity of the treeing software concerns the tree optimization performing cycles of NNI and Kernigham–Lin (KL) (16) tree modifications. This optimization can not only be applied to the complete tree but also confined to user-selected subtrees. Thus tree optimization is possible applying the appropriate filters for the respective phylogenetic levels and groups. In this context, it is of interest that, while performing stepwise optimizations, 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 visualization in the ARB main window.

The positional tree server

The ARB positional tree (PT) server, once established, allows rapid finding of sequence identity or peculiarity. Thus it is the central tool for fast search of closest relatives for automated sequence alignment or to define diagnostic sequence stretches for primer and probe design. Establishing a positional tree server of any oligonucleotide sequence up to 20mers occurring in the underlying database and assignment of the individual oligonucleotides to the sequences or organisms containing

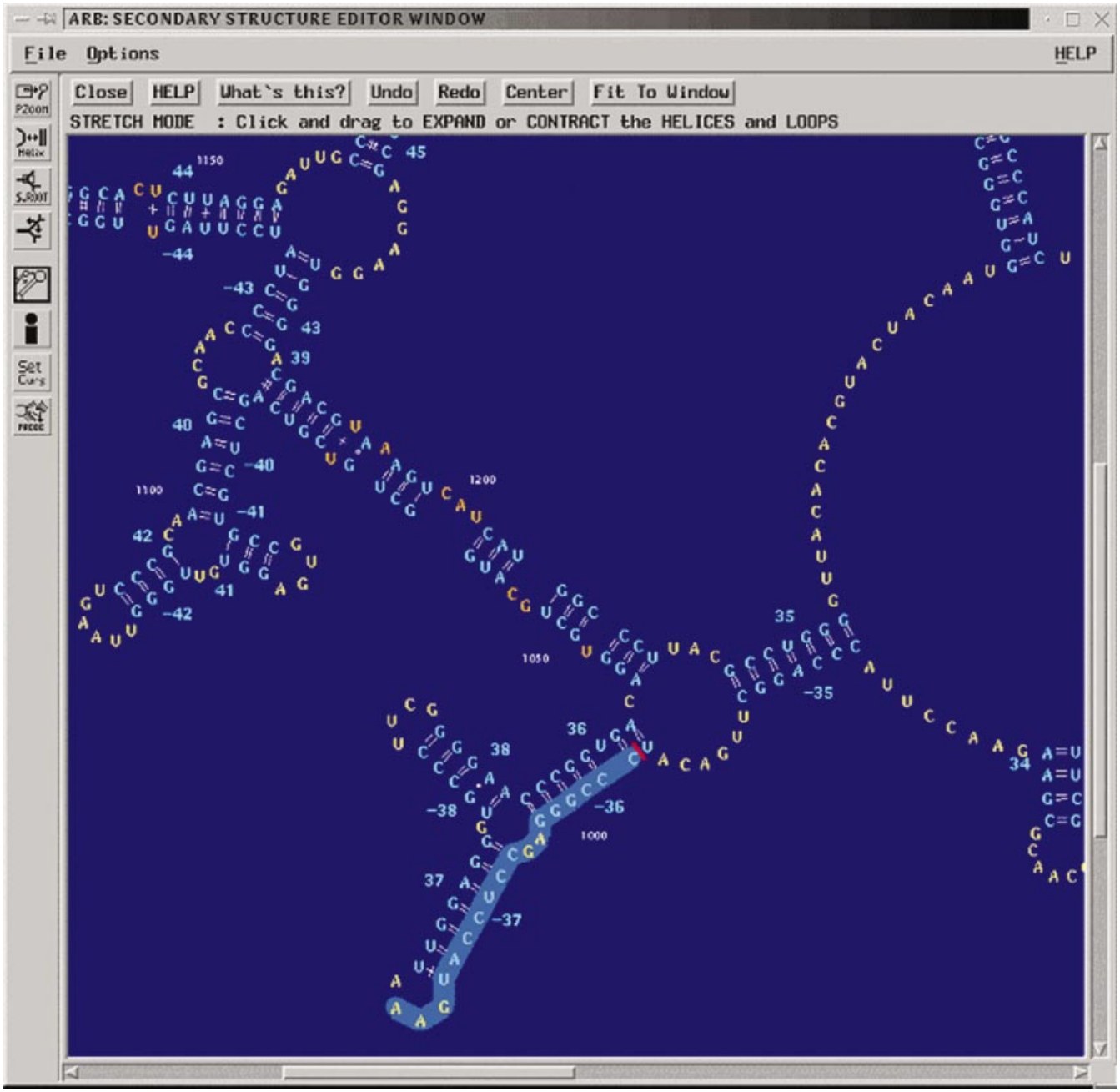

Figure 5. Secondary structure editor. The sequence selected in the primary structure editor (Fig. 4) is automatically fitted into a consensus secondary structure model.

them is the basis for these procedures. PT-server-based analyses do not rely upon aligned sequences. The PT server is not provided with the ARB program package or ARB database. It has to be established for the respective database locally. The PT server can be used by multiple users on the local machine or via network. The computing time for generating the respective files depends strongly on the size and structure of the individual database as well as the performance of the machine used. The advantages of these logarithmic algorithms over linear ones such as Blast (17) or Fasta (18) are effectiveness and rapidity.

Sequence alignment

As mentioned in Materials and Methods, for *de novo* generation of a nucleic or amino acid sequence alignment ClustalW (19) 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.

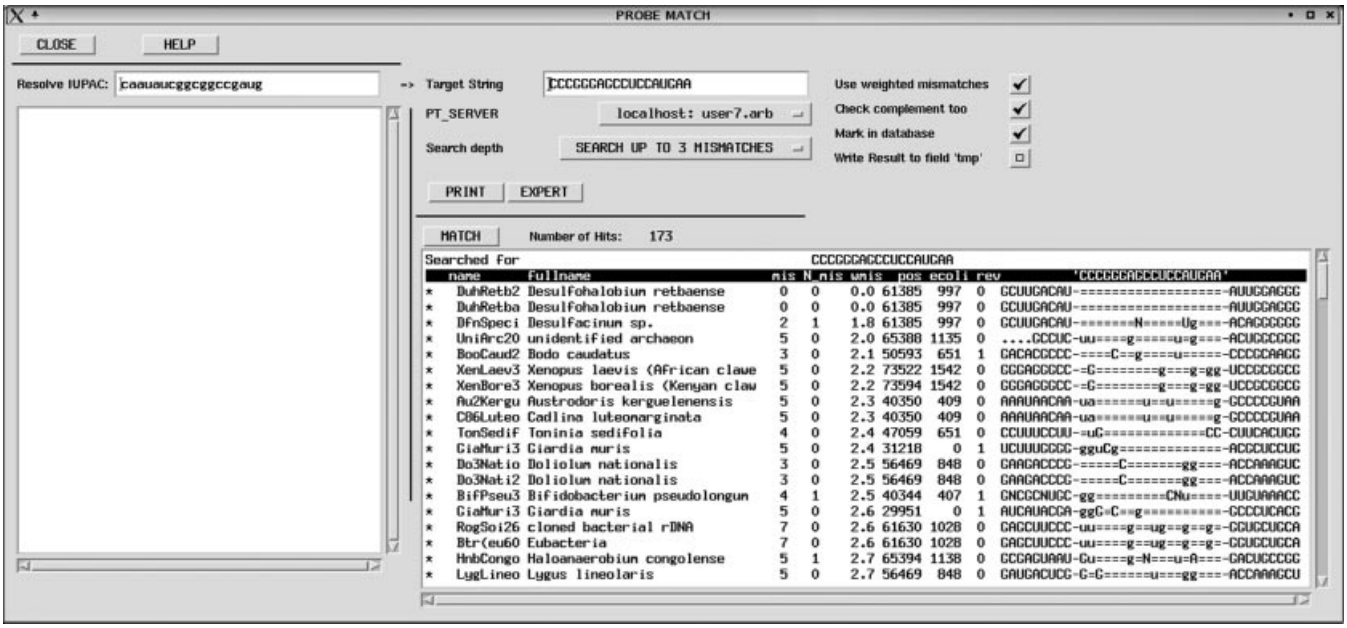

Figure 6. Results of probe design and evaluation. Part of the primary structure alignment containing the probe target site is shown for the target organism *Desulfohalobium retbaense* and the non-target organisms containing the most similar sequence stretches.

sequences or consensus defined by the user or automatically determined by PT-server-based search for most similar reference sequences, are used as the template.

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

Probe design and evaluation

Currently, taxon- or gene-specific probe design certainly plays a central role in many molecular biological research and analysis projects, for example the identification and detection of organisms in complex environmental samples or expression studies within the scope of genome projects. 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 the 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. First, the user selects the organism or a group of organisms for which he or she 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 thermodynamic criteria (12,20). 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 no to five mismatches) in the respective database (Fig. 6). Furthermore, these sequence strings can automatically be visualized in the primary and secondary structure editors. The latter information is of particular importance when designing probes for *in situ* cell hybridization. A tool for visualization of accessibility maps (13,14) in the primary and secondary structure editors is under development.

A special advance is the ARB multiprobe software component. It determines sets of up to five probes optimally identifying the target group (21). These probe sets can be used for multiple fluorescence *in situ* hybridization experiments.

Data import and export

The sequence as well as additional data can be imported and exported in commonly used flat file formats. The parsing from and to tagged flat files can be customized by advanced users. There is also a tool for automated completion of database submission forms for those users determining sequences on their own.

Availability and documentation

Although so far not officially published, previous versions of the software package and databases have been available for several years and the software has been used worldwide. The ARB package provides a comprehensive set of tools to support the user's work. However, depending upon the user's interests, some knowledge of the basics of sequence alignment, phylogenetic analyses or probe hybridization is needed. Some familiarity with UNIX operating systems is advantageous. During installation, environment variables, paths,

Table 1. Run time studies for PT server generation, automated alignment and parsimony-based treeing

No. of sequences	10	100	1000	10 000	25 743
PT server	–	5 s	22 s	3 min	7 min 30s
Add to alignment	4 s	38 s	6 min		
Add to tree	1 min	9 min 15 s	2 h		

Datasets varying with respect to the number of sequence entries were used for PT server generation. The most comprehensive of these datasets comprising 25 743 entries and 40 000 alignment positions was used as a template for inserting the specified numbers of sequences into the database alignment or tree. For treeing, 2141 alignment columns were included.

permissions and aliases have to be defined. Instructions for installation can be downloaded from http://www.arb-home.de/download/ARB/documentation/ARB_install.pdf. Self-installing versions of the recent program releases are currently available for Linux systems only. The binaries, source code and some documentation are available at the download area of the ARB web site http://www.arb-home.de/download/. An HTTP Browser is required as ftp connection is not accepted. Furthermore, there is an email forum of the worldwide ARB users community. Subscription is needed for those interested in joining (subscribe@arb-home.de). Although a comprehensive formal handbook is not yet available, manuals, instructions and problem solutions are available from the ARB homepage and by contacting the ARB staff and user community via the email forum. ARB sequence databases are currently available for 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/downloads/databases/.

Systems, hardware and processing time requirements

The ARB group provides tested versions for SuSE LINUX and Sun Solaris systems. According to information provided by users, the LINUX version also runs 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, mainframe memory is more important than processor performance. The users among the wet laboratory partners of the ARB group are performing their analyses on dual Pentium III PCs with 1 Gb memory and 1 Gb swap space. The background storage requirements depend mainly upon the number and size of the user ARB databases and PT servers. The sizes of the installed program package, the current small subunit rRNA database and the respective PT server files are about 25 Mb, 80 Mb and 350 Mb, respectively. Twenty-one-inch monitors at 1600 × 1200 are recommended. However, ARB is also routinely used on laptops or older PCs and workstations with less memory and monitors with lower resolution.

Table 1 gives some information on processing times for some of the major functions in ARB: generation of the PT server, automated sequence alignment and phylogenetic treeing. Run-time measurements were performed on a dual-processor (Intel® Xeon™, 2.6 GHz) PC equipped with a 2 Gb RAM running SuSE Linux 8.2. The ARB databases ssu_jan03.arb (25 743 almost complete small subunit rRNA sequences, 40 000 alignment positions) and ssu_10k.arb, ssu_1k.arb, ssu_100.arb (subsets comprising 10 000, 1000 and 100 sequences) used for these PT server building studies are available at http://www.arb-home.de. For adding 10, 100 and 1000 sequences to the database alignment or tree, ssu_jan03.arb was used as basis. The ARB aligner was used in combination with a PT-server-directed search for most similar reference sequences. For phylogenetic treeing the respective sequences/organisms were added to a tree comprising all entries of the database applying the ARB parsimony tool.

Future developments

The ongoing developments are focusing on two major tasks. First, a web tool providing all potential probe target sites which can be derived from the current database version and should phylogenetically (taxonomically) make sense. Users can not only search for hierarchical and multiple probes submitting names, strain designations or accession numbers of organisms as search strings, but can also send their own probe sequences for *in silico* evaluation. Second, 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 visualization 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.

ACKNOWLEDGEMENTS

The authors highly acknowledge F. O. Glöckner and R. Amann (Max Planck Institute for Marine Microbiology, Bremen, Germany) for redistributing the ARB database and software and answering user queries, as well as hosting and organizing ARB workshops. The authors thank Laura Schulz (Ludwig Maximilian University, Munich, Germany) for critical reading of the manuscript. ARB software development and database maintenance was partly supported by the European Union within the HRAMI project, by the German Research Foundation and by the German Ministry of Research and Education BIOLOG project.

REFERENCES

1. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al*. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **30**, 21–26.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
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. and 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. (2001) The European Large Subunit Ribosomal RNA Database. *Nucleic Acids Res.*, **29**, 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, Tokyo.
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. *Proc. Supercomputing Conference (SC2002)*, *Baltimore*, *MD*, Nov. 2002, IEEE Computer Society.
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. *et al*. (2002). The comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron and other RNAs. *BioMed Central Bioinform.*, **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. Fuchs,B.M., Wallner,G., Beisker,W., Schwippl,L., Ludwig,W. and Amann,R. (1998) Flow cytrometric analysis of the *in situ* accessibility of *Escherichia coli* 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, **64**, 4973–4982.
14. Fuchs,B.M., Syutsubo,K., Ludwig,W. and Amann,R. (2001) *In situ* accessibility of the *Escherichia coli* 23S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, **67**, 961–968.
15. 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* (2nd edn). Springer, New York, pp. 49–65.
16. Kernigham,B.W. and Lin,S. (1970) An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.*, **49**, 291–307.
17. 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.
18. Pearson,W.R. and Lipman,D.C. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
19. Thompson,J.D., Higgins,D.G. and Gibson,D.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment. *Comput. Appl. Biosci.*, **8**, 189–191.
20. Amann,R. and Ludwig,W. (2000) Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS Microbiol. Rev.*, **24**, 555–565.
21. 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 Soil*, **204**, 1–9.