BMC Bioinformatics

Software

Graphical representation of ribosomal RNA probe accessibility data using ARB software package

Yadhu Kumar[1], Ralf Westram[1], Sebastian Behrens[2], Bernhard Fuchs[2], Frank Oliver Glöckner[2,3], Rudolf Amann[2], Harald Meier[4] and Wolfgang Ludwig*[1]

Address: [1]Lehrstuhl für Mikrobiologie, Technische Universität München, D-85350 Freising Germany, [2]Max Plank Institute for Marine Microbiology, D-28359 Bremen, Germany, [3]International University Bremen, D-28759 Bremen, Germany and [4]Lehrstuhl für Rechnertechnik und Rechnerorganisation, Technische Universität München, D-85748 Garching, Germany

Email: Yadhu Kumar - yadhu@mikro.biologie.tu-muenchen.de; Ralf Westram - westram@mikro.biologie.tu-muenchen.de; Sebastian Behrens - sbehrens@stanford.edu; Bernhard Fuchs - bfuchs@mpi-bremen.de; Frank Oliver Glöckner - fog@mpi-bremen.de; Rudolf Amann - ramann@mpi-bremen.de; Harald Meier - meierh@in.tum.de; Wolfgang Ludwig* - ludwig@mikro.biologie.tu-muenchen.de

* Corresponding author

Abstract

Background: Taxon specific hybridization probes in combination with a variety of commonly used hybridization formats nowadays are standard tools in microbial identification. A frequently applied technology, fluorescence *in situ* hybridization (FISH), besides single cell identification, allows the localization and functional studies of the microbial community composition. Careful *in silico* design and evaluation of potential oligonucleotide probe targets is therefore crucial for performing successful hybridization experiments.

Results: The PROBE Design tools of the ARB software package take into consideration several criteria such as number, position and quality of diagnostic sequence differences while designing oligonucleotide probes. Additionally, new visualization tools were developed to enable the user to easily examine further sequence associated criteria such as higher order structure, conservation, G+C content, transition-transversion profiles and *in situ* target accessibility patterns. The different types of sequence associated information (SAI) can be visualized by user defined background colors within the ARB primary and secondary structure editors as well as in the PROBE Match tool.

Conclusion: Using this tool, *in silico* probe design and evaluation can be performed with respect to *in situ* probe accessibility data. The evaluation of proposed probe targets with respect to higher-order rRNA structure is of importance for successful design and performance of *in situ* hybridization experiments. The entire ARB software package along with the probe accessibility data is available from the ARB home page http://www.arb-home.de.

Background

The introduction and use of comparative sequence analysis of appropriate marker genes as a powerful tool in taxonomy has substantially contributed to the rapid growth of molecular sequence databases such as EMBL [1], GenBank [2], and ribosomal RNA (rRNA) databases [3-5].

Evidently, molecular phylogenetic analyses have greatly influenced the restructuring of systematics especially in the case of prokaryotes. Nowadays, identification and classification at the species and higher taxonomic levels mainly relies on a genotypic approach, typically involving an analysis of small, and to a lesser extent, large ribosomal RNA gene (rRNA) structures. The backbone of the current taxonomy of the prokaryotes is almost exclusively based upon a phylogenetic network derived from comparative sequence analysis of the small subunit rRNAs and respective phylogenetic marker genes [6]. As 'living fossils', these molecules at least roughly reflect the evolutionary history of the respective organisms. The mosaic-like primary structures comprising highly variable to highly conserved or invariant regions provide diagnostic information for different levels of phylogenetic relationship. Consequently, this information can be used to identify oligonucleotide target regions unique to phylogenetic entities, for use as taxon-specific hybridization probes or PCR primers. Depending on the target site such oligonucleotide probes or probe combinations can be designed for phylogenetic groupings as diverse as bacterial species or an entire phylum.

Ever since the fluorescence *in situ* hybridization (FISH) technique became an integral part of the rRNA approach to microbial ecology and evolution [7], rRNA-targeted oligonucleotide probes have evolved into a widely used tool for the direct, cultivation-independent identification and enumeration of individual microbial cells or specific groups of bacteria in simple to complex natural environments. In this regard, a good probe design and careful further evaluation *in silico* plays a crucial role to ensure sensitivity and specificity of a potential probe in its practical application. Besides uniqueness of the target sequence, number, character and position of diagnostic residues, comprehensiveness with respect to the inclusion of members of the desired target group (taxon) and exclusion of non-members along with a target molecule or region accessibility in the real hybridization experiment, have to be taken into consideration. Recently, data on *in situ* accessibility of rRNA targets in several microorganisms have become available [8-11].

Since biology is a highly visual science, there is a general demand for tools to visualise the variety of biological knowledge as diagrams, illustrations, two-dimensional and three-dimensional reconstructions, and other types of graphical formats. Hence, the visualization of molecular data in an interactive and intuitive graphical user interface ideally will serve as third eye for a molecular biologist. In this paper, we describe how the ARB software package [3] provides a workbench for designing, evaluation and visualization of oligonucleotide probes in more intuitive way,

using interactive graphical user interface to visually examine characteristics and criteria of target regions.

Implementation
Sequence data
Periodically retrieved raw gene data comprising small subunit rRNA from public databases such as EBI [1], Genbank [2], the RDP[4], and the sequence data determined in our laboratory and other partner groups are imported into the ARB database, processed according to a variety of criteria and finally provided as curated databases at the ARB projects web-site [13]. The current public release of small subunit rRNA database [3] containing only complete sequences was taken for designing, evaluation and visualization of probes and targets, respectively. Partial sequences are avoided as they greatly limit the probe design by reducing the number of potential target regions and also give no hint about the specificity of existing probes that target to non-sequenced regions of the respective rRNAs.

The positional tree (PT) server
The PT-Server [3] is a suffix tree server implemented in the ARB software which is used for indexing all sequence data represented in the underlying ARB sequence database. Once established, the particular PT-Server allows rapid and exact searching for target regions with respect to sequence identity or uniqueness.

Probe design and probe match
Probe design is carried out using the PROBE Design tool (PDT) of ARB software involving following steps:

1. The user selects the target group or a species of interest.

2. The parameters such as size of the probe and the probable physico-chemical characteristics like %GC content, melting temperature (T_m) according to the 4°C GC, 2°C AT rule [14], and self-complementarity (hair-pin bonds) are specified. Optionally, a range of allowed target positions within the sequence alignment of the respective database can be defined.

3. Potential probe candidates are searched involving the respective PT-Server. Both, target and probe sequence are displayed in a result list. Ranking within this list follows estimated probe quality according to criteria defined for probe design such as number, character and position of diagnostic residues, coverage of the target group, physico-chemical demands, which are displayed in separate probe results window along with relevant information.

4. Once the user selects the desired probe in the result list, it can be evaluated against the entire database by using the PROBE Match tool (PMT) of ARB. PMT, by default
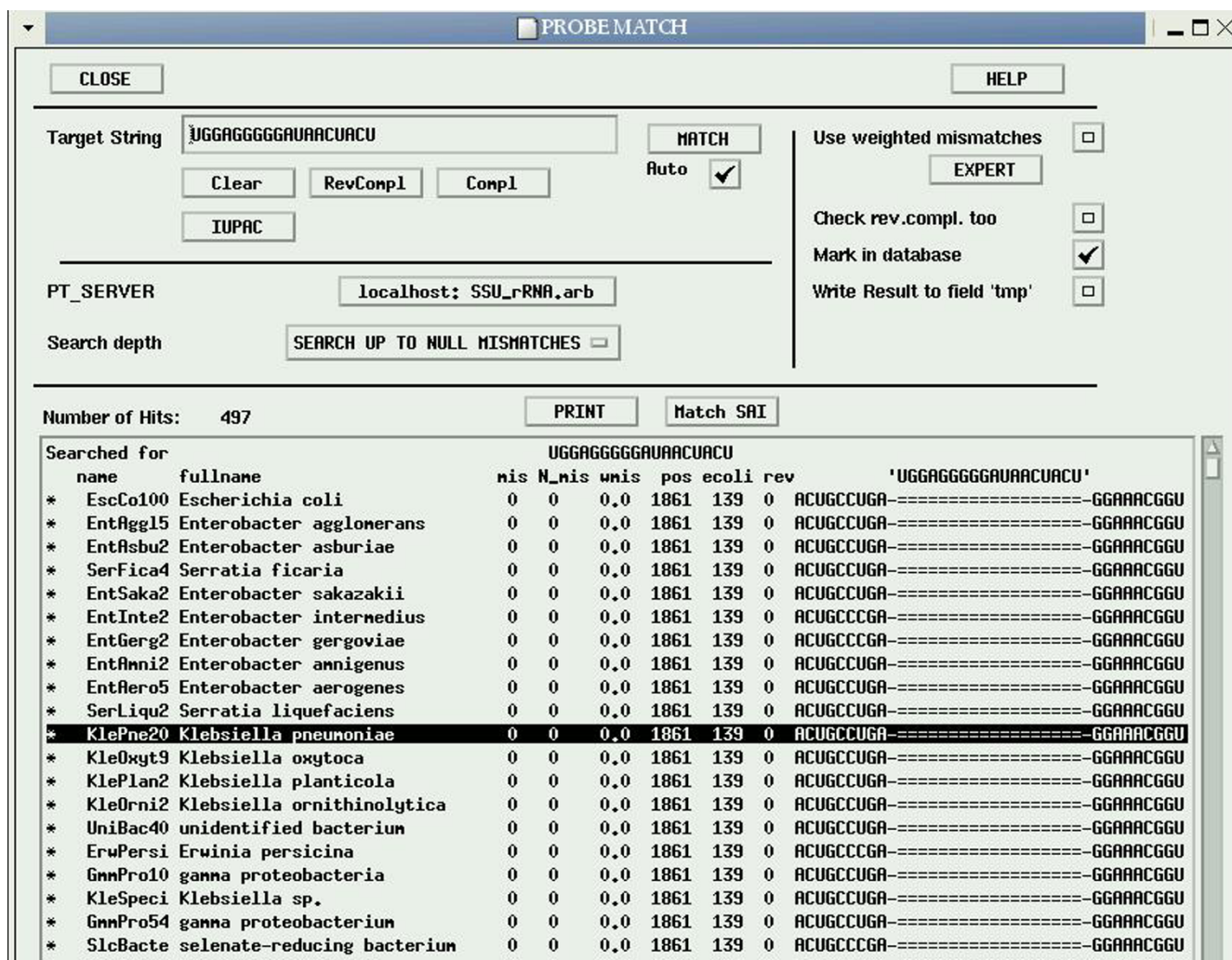
Figure 1
Probe match window. Probe Match Window showing results of *in silico* evaluation of the probe candidate (5'-UGGAG-GGGGAUAACUACU-3') designed as part of a probe combination for the enterobacteria group. It hits 497 members of the group containing 947 species. Additional information such as number of mismatches, *E. coli* position, region up- and downstream of the probe target in the actual nucleotide sequence along with the local alignment are also shown.

evaluates the targets for the sequence (strand) stored in the database. Optionally, the complementary sequence (opposite strand) can be evaluated as well. Members of the target group are displayed in a separate PROBE Match window along with other information such as number of mismatches, weighted mismatches, *E. coli* positions, reverse complementarity and local alignment of probe targets (Figure 1).

Results and discussion
As the demand for oligonucleotide probes that can identify and quantify bacteria by nucleic acid hybridization is permanently increasing, *in silico* evaluation and visualiza-

tion of such probes and targets are necessary, particularly, when used for FISH experiments. Target accessibility is among the crucial criteria to be evaluated with respect to experimental success of the respective probe based identification and detection system [7-12]. To facilitate this evaluation procedure, new functionalities were added to the ARB software package providing a more intuitive graphical environment. As an example, oligonucleotide probes were designed for the enterobacteria group represented by 947 database entries. The 5'-UGGAG-GGGGAUAACUACU-3' probe was selected from the list of potential probes and evaluated against the background of the full dataset of complete and partial small subunit
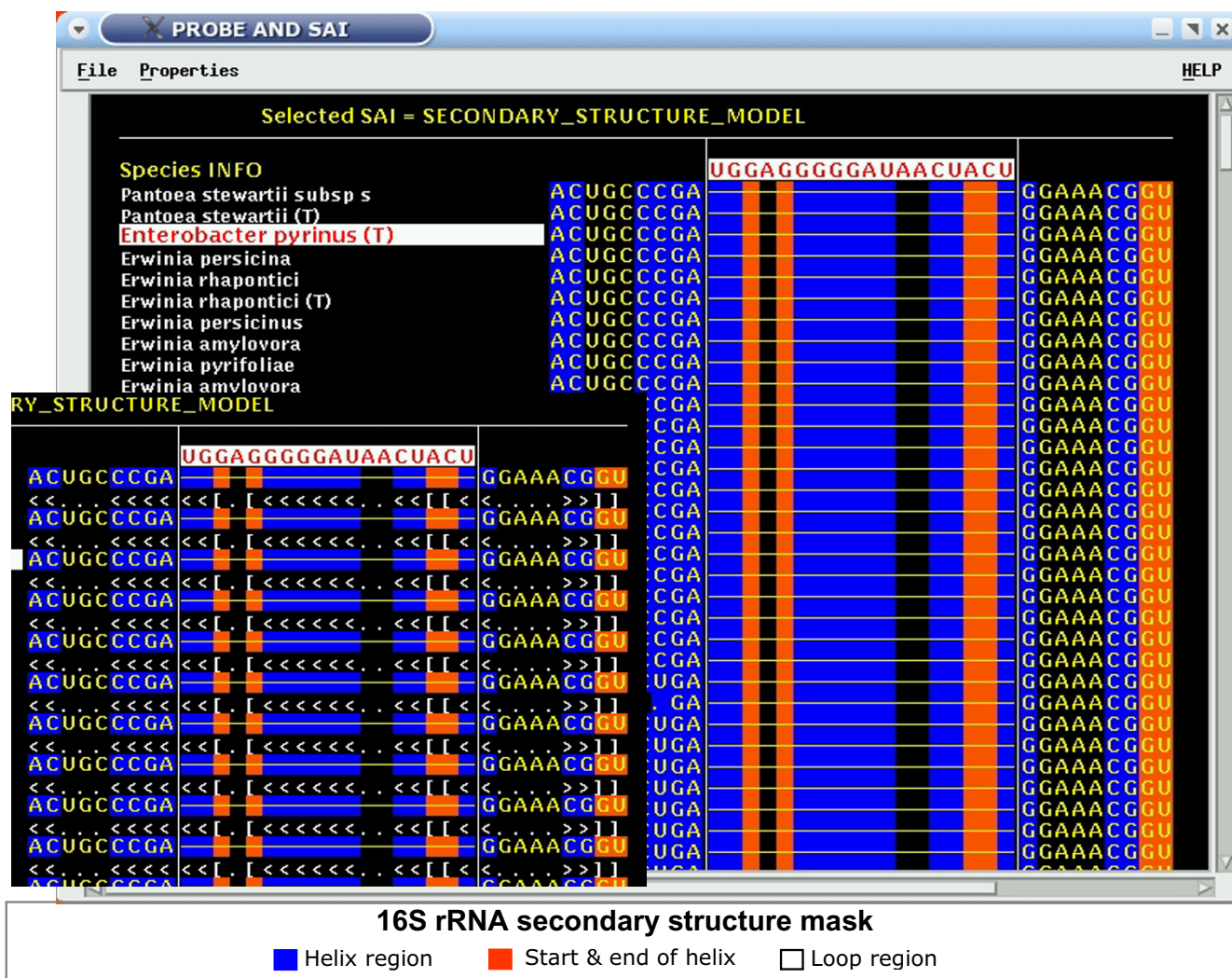
Figure 2
Screenshot displaying 16S rRNA secondary structure model where helix region is colored in blue, starting and ending positions of helix halves are in red and bases without background represent commonly non base paired positions.

rRNA sequences. The selected probe perfectly matches the respective target of 497 members of the enterobacteria group (Figure 1). The same probe has been visualized in all the screenshots presented in the paper.

Although a phylogenetic probe is primarily judged in terms of its taxonomic range to identify the members of its intended target taxon to the exclusion of non-target bacteria, for a practical consideration it must also fulfil certain other criteria with respect to its applicability depending on the particular hybridization format. In case of the fluorescence *in situ* hybridization approach the results of the

accessibility studies conducted by Fuchs and co-workers on the 16S and 23S rRNA of *Escherichia coli* and other organisms are among such criteria. They showed that some regions of *E. coli* ribosome are virtually inaccessible for oligonucleotide probes when FISH is performed [8,9]. They proposed a color code assigned to six intensity classes of *in situ* hybridization signals. Within the ARB program, these classes are coded in respective SAIs (so called Sequence Associated Information) and optionally visualized as background colors of the sequences in primary structure (ARB_Edit4), secondary structure (SEC_Edit), and probe visualization windows (PROBE
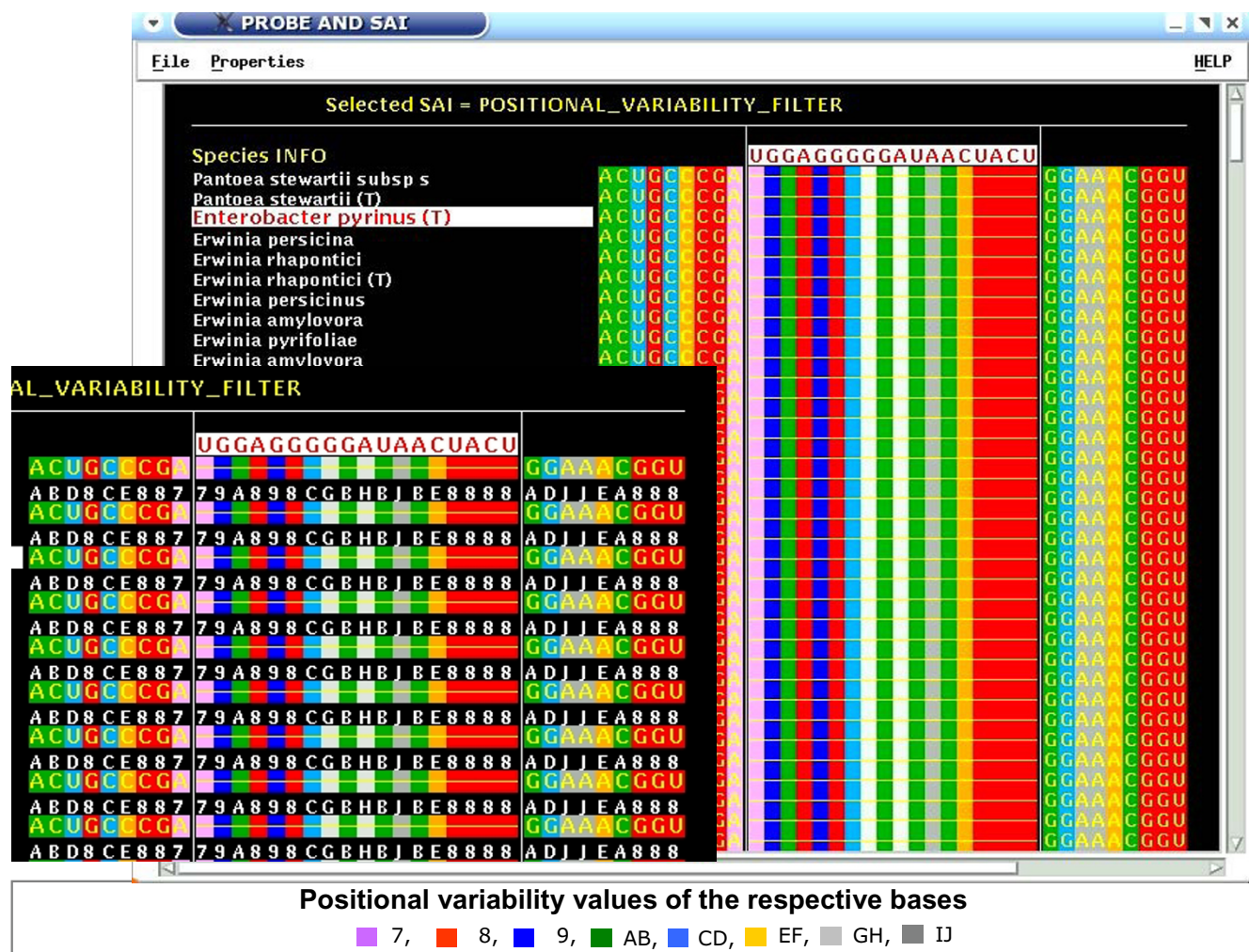
Positional variability values of the respective bases

■ 7, ■ 8, ■ 9, ■ AB, ■ CD, ■ EF, ■ GH, ■ IJ

Figure 3
Screenshot showing positional variability filter generated by the respective ARB tool in various background colors (light purple – 7; red – 8; dark blue – 9; green – AB; blue – CD; yellow – EF; light grey – GH; grey – IJ; increasing numbers followed by the alphabetical order of letters indicate increasing degree of sequence conservation).

Match) of ARB. All the displays produced by the ARB software are interconnected and any changes in one window are automatically updated in other windows as well. Simultaneous visualization and evaluation of oligonucleotide probes in different levels allows the user to look carefully and closely into the proposed probe candidates *in silico*, before carrying out further *in situ* or *in vivo* studies. More importantly, the user can perform a variety of sequence related operations such as importing sequence data, aligning, treeing, designing, evaluation and visualization of probes, performing statistical calculations and many other functions using interoperating and user friendly tools controlled from a common graphical platform within the ARB software package.

Visualization of potential probe candidates and the sequence associated information (SAI) such as higher order structure, conservation, G+C content, transition-transversion profiles and *in situ* target accessibility patterns, is possible at three different levels: the local alignment (PROBE Match tool), global alignment (ARB Primary Structure Editor) and secondary structure levels (Secondary Structure Editor).

Visualization of SAI in probe match window
Visualization of probe candidates in a local alignment along with additional sequence associated information (SAI) can be managed with the PROBE Match SAI window. The neighboring region up to nine nucleotides on
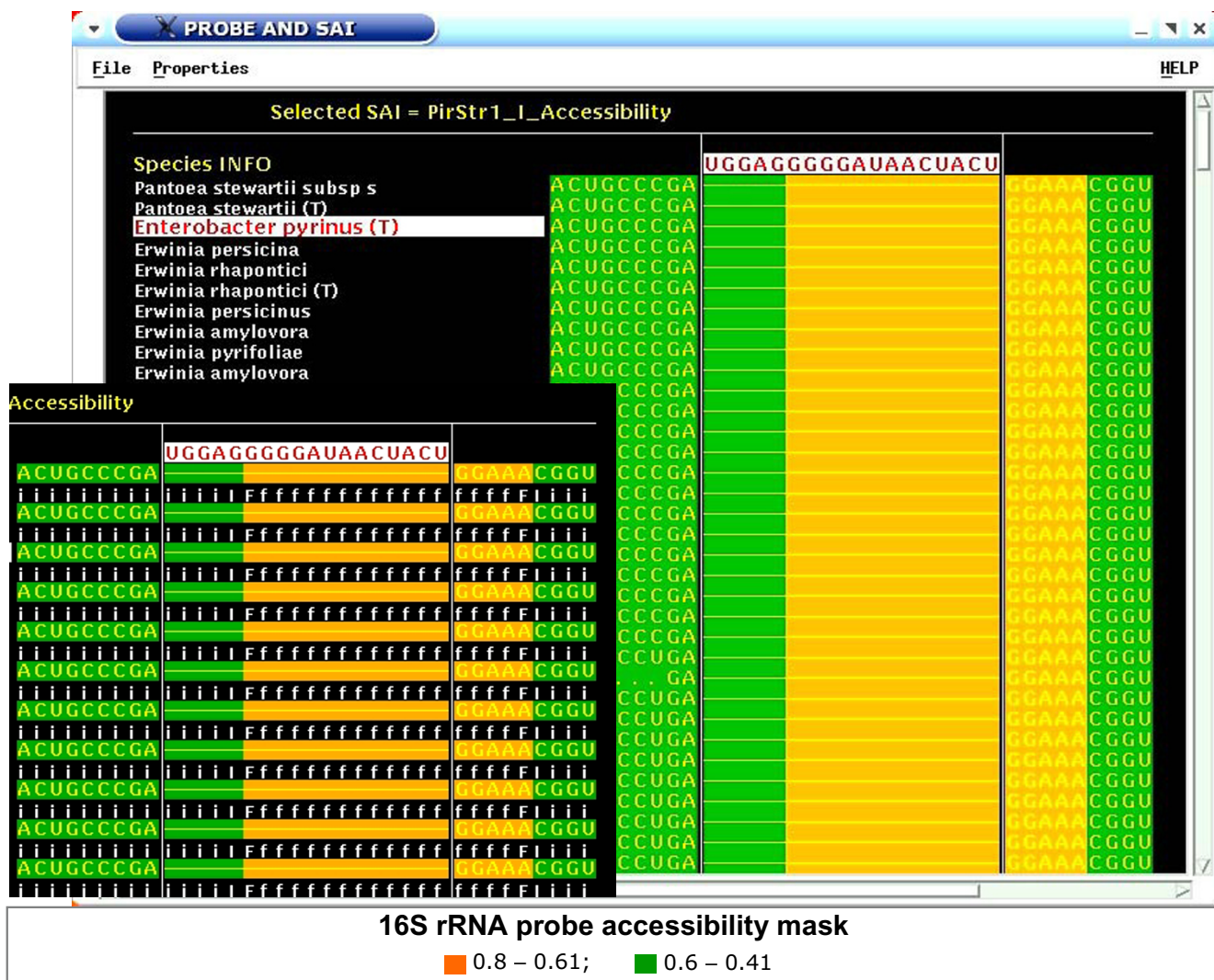
Figure 4
Probe visualization window. Probe visualization window displaying local alignments of probe target regions for members of the target group. Column statistics performed on the sequence alignments and structure masks are visualized in different background colors in the probe visualization window. This screenshot shows 16S rRNA accessibility map, where the experimentally determined relative fluorescence intensities are visualized in different colors (orange (0.8 – 0.61); green (0.6 – 0.41) [11]).

either terminus of the potential probe target is retrieved from the database. A local alignment of the extracted rRNA sequence is established and displayed along with the respective unique identifier such as ARB short_name, accession number, or any other underlying database fields (eg., Full Name, Group) (Figure 2, 3, 4). The user can select any information that is associated with the sequences (SAI) such as secondary structure masks (Figure 2) or any statistical calculations performed on the

sequence level like sequence consensus, positional variability using parsimony method (Figure 3) or any other user defined models, filters or statistics as well as *in situ* accessibility maps for visualization (Figure 4). Different background colors can be assigned to characters and values or character groups and ranges of values of the particular SAIs, respectively. Optionally, the real characters or values contained in such SAIs can directly be visualized below the individual sequences. This offers a researcher a
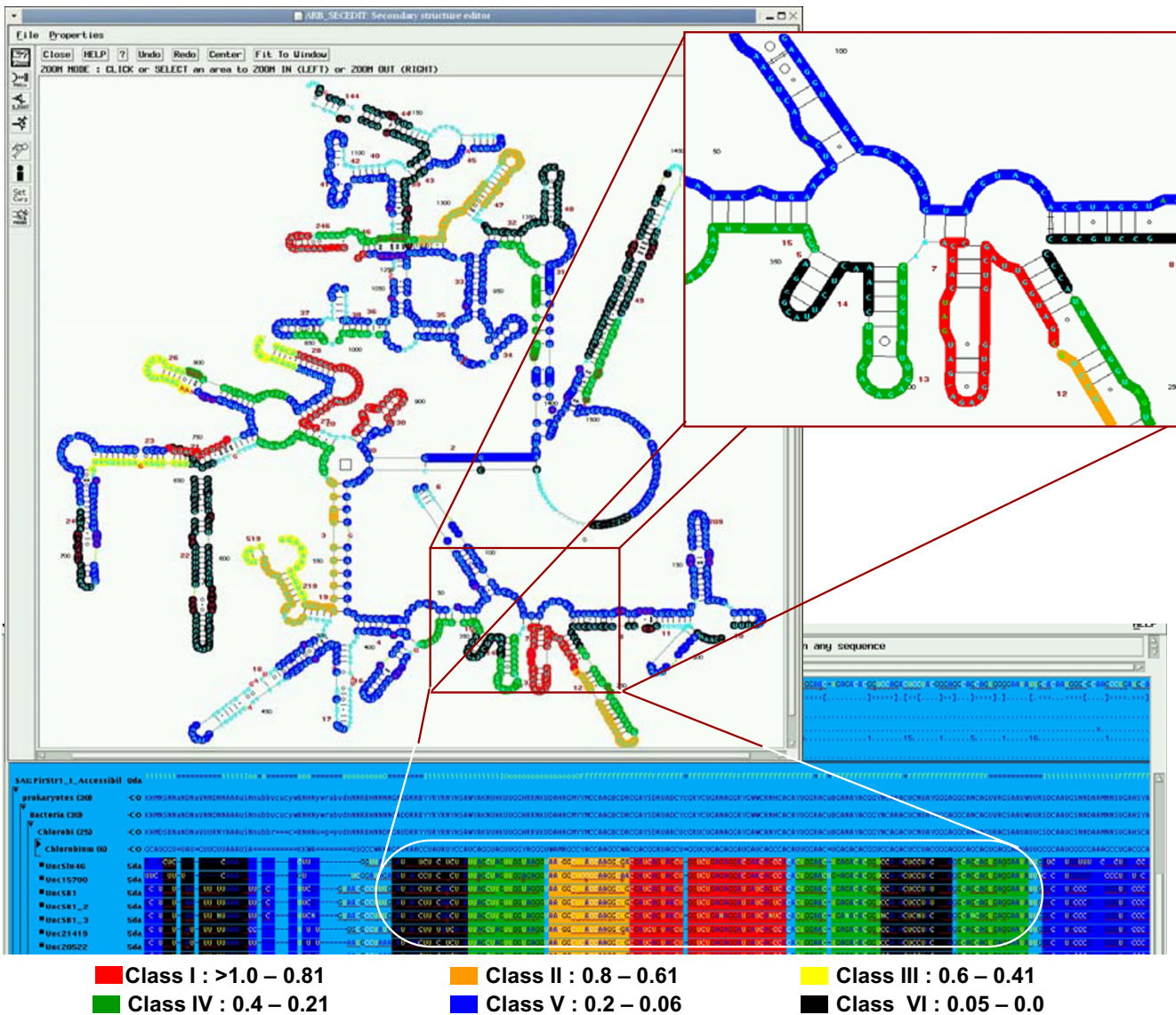
🟥 **Class I : >1.0 – 0.81**	🟧 **Class II : 0.8 – 0.61**	🟨 **Class III : 0.6 – 0.41**
🟩 **Class IV : 0.4 – 0.21**	🟦 **Class V : 0.2 – 0.06**	⬛ **Class VI : 0.05 – 0.0**

Figure 5
ARB Primary and secondary structure windows. ARB primary and secondary structure windows showing the distribution of relative fluorescence intensities of oligonucleotide probes targeting 16S rRNA structure of *M. sedula* [11]. The different background colors indicate brightness range of different classes (classes I through VI) with respect to the observed fluorescence intensities. Numbers displayed in black and lower case denotes the respective nucleotide positions and the numbers showing in red represent helix numbers in the model.

deeper insight in to the proposed oligonucleotide probe targets for careful examination of probe candidates *in silico* before making any decision on the selection of probe.

Visualization of SAI in ARB primary structure editor
On the global alignment level, the user selected oligonucleotide probe is visualized in different background colors in the primary structure editor window of ARB [3]. The primary structure editor (Figure 5) of ARB displays multi-

ple sequence alignments generated by the respective ARB software tools [3] of the selected sequences from the underlying database in the user-defined colors and symbols.

As already described for the local alignment level, any type of SAI can be visualized by the user defined background colors for the individual alignment columns. Customized color selections can be assigned to the different types of

SAIs mentioned before. By scrolling the mouse or the use of ARB search tools, the user gets an easy access to the information for any range or the selection of sequences. In the context of probe evaluation for *in situ* hybridization experiments, mapping of experimentally derived *in situ* accessibility patterns onto the primary structures of interest certainly provides valuable support to the users for probe evaluation. Part of a multiple 16S rRNA sequence alignment is shown in the figure 5. The brightness classes defined for 16S rRNA structural model of *Methanosaeta sedula* [11] are mapped on the aligned sequences and indicated by background colors according to Behrens *et al* [11].

Visualisation of SAI in ARB secondary structure editor
Theoretically as well as experimentally derived secondary structure information of SSU rRNA [15-17] is used more profoundly in sequence alignment refinement and probe design and evaluation. The tertiary structure of the SSU rRNA of the bacterium *Thermus thermophilus* which had been elucidated with atomic resolution by X-ray diffraction crystallography of ribosomal subunit [17] allows evaluating the exactness of the secondary structure model. The secondary structure of SSU rRNA has a crucial role in evaluating the proposed probe candidates prior to the actual experimentation. The ARB Secondary structure editor (Figure 5) provides the user with more intuitive graphical display of the secondary structure model of SSU rRNA. The user can visualize the entire SSU rRNA sequence of any organism in the respective database which fits into the common consensus model. The localization of proposed oligonucleotide probe targets can be visualized in customizable background colors.

Conclusion
The evaluation of proposed probe target position with respect to higher-order rRNA structure is of more importance especially when probes are intended to be used for *in situ* hybridizations [7-12]. Albeit there have been several software programs developed for the design of rRNA targeted oligonucleotide probes [18,19], the criteria taken to design the probes are generally restricted to certain parameters such as size, nucleotide composition, specificity definition, and the general hybridisation behavior. None of the software described [18,19] takes into account the special requirements of rRNA targeted probes that are destined for FISH applications which is, the structure dependant probe accessibility of the ribosomal RNA. This feature has been developed and implemented in ARB. Using this tool, *in silico* probe design and evaluation can be performed with respect to *in situ* probe accessibility data. By identifying and excluding the probes targeting sites with a poor accessibility the number of time consuming empirical tests can be reduced.

Availability and requirements
The entire ARB software and the periodic updates of well aligned and annotated ribosomal RNA databases are made freely available for the scientific community via World Wide Web [13]. Currently, the ARB Software is available for PCs running LINUX operating systems and SUN SOLARIS systems.

Authors' contributions
YK developed and implemented the tool and drafted the manuscript. RW participated in design and implementation. SB and BF provided the accessibility data and revised the manuscript. FOG, RA and HM critically revised the manuscript. WL initiated the development of the tool and supervised the ARB project.

Acknowledgements

References
1. Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, Harte N, Kanz C, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Stoehr P, Stoesser G, Tuli A, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R: **The EMBL nucleotide sequence database.** *Nucleic Acids Res* 2004, **32:**D27-30.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **Genbank: update.** *Nucleic Acids Res* 2004, **32:**D23-26.
3. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar , Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüßmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH: **ARB: a software environment for sequence data.** *Nucleic Acids Res* 2004, **32:**1363-1371.
4. Maidak BL, Cole JR, Lilburn TG, Parker CT, Saxman PR Jr, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM: **The RDP-II (Ribosomal Database Project).** *Nucleic Acids Res* 2001, **29:**173-174.
5. Wuyts J, PerrieÁre G, Van de Peer Y: **The European ribosomal RNA database.** *Nucleic Acids Res* 2004, **32:**D101-103.
6. Ludwig W, Klenk HP: **Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics.** In *Bergey's Manual of Systematic Bacteriology* 2nd edition. Edited by: Garrity G. New York: Springer; 2001:49-65.
7. Amann R, Ludwig W, Schleifer KH: **Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation.** *Microbiol Rev* 1995, **59:**143-169.
8. Fuchs BM, Wallner G, Beisker W, Schwippl I, Ludwig W, Amann R: **Flow cytometric analysis of the *in situ* accessibility of Escherichia coli 16S rRNA for fluorescently labeled oligonucleotide probes.** *Appl Environ Microbiol* 1998, **64:**4973-4982.
9. Fuchs BM, Syutsubo K, Ludwig W, Amann R: ***In situ* accessibility of the Escherichia coli 23S rRNA for fluorescently labeled oligonucleotide probes.** *Appl Environ Microbiol* 2001, **67:**961-968.
10. Inàcio J, Behrens S, Fuchs BM, Fonseca I, Spencer-Martins I, Amann R: ***In situ* Accessibility of *Saccharomyces cerevisiae* 26S rRNA to Cy3-Labeled Oligonucleotide Probes Comprising the D1 and D2 Domains.** *Appl Environ Microbiol* 2003, **69:**2899-2905.
11. Behrens S, Ruehland C, Inàcio J, Huber H, Fonseca A, Spencer-Martins S, Fuchs BM, Amann R: ***In Situ* accessibility of small-subunit rRNA of members of the domains Bacteria, Archaea and Eucarya to Cy3-Labeled oligonucleotide probes.** *Appl Environ Microbiol* 2003, **69:**1748-1758.
12. Behrens S, Fuchs BM, Mueller F, Amann R: **Is the *In situ* Accessibility of the 16S rRNA of Escherichia coli for Cy3-Labeled Oligonucleotide Probes Predicted by a Three-Dimensional**

Structure Model of the 30S Ribosomal Subunit? *Appl Environ Microbiol* 2003, **69**:4935-4941.

13. **The ARB project** [http://www.arb-home.de]

14. Suggs SV, Hirose T, Miyake T, Kawashima EH, Johnson MJ, Itakura K, Wallace RB: **Use of synthetic oligodeoxyribonucleotides for the isolation of specific cloned DNA sequences.** In *Developmental biology using purified genes* Edited by: Brown D, Fox CF. New York: Academic Press Inc; 1981:683-693.

15. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, MuÈller KM, Pande N, Shang Z, Yu N, Gutell RR: **The comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron and other RNAs.** *BMC Bioinformatics* 2002, **3**:15.

16. Gutell RR: **Collection of small subunit (16S- and 16S-like) ribosomal RNA structures.** *Nucleic Acids Res* 1993, **21**:3051-3054.

17. Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V: **Structure of the 30S ribosomal subunit.** *Nature* 2000, **407**:327-339.

18. Ashelford KE, Weightman AJ, Fry JC: **PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database.** *Nucleic Acids Res* 2002, **30**:3481-3489.

19. Pozhitkov AE, Tautz D: **An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification.** *BMC Bioinformatics* 2002, **3**:9.