

Lehrstuhl für Mikrobiologie
der Technischen Universität München

Development and Integration of Structure Based Visualization Tools in ARB Software Package

Yadhu Kumar

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum
Weihenstephan
für Ernährung, Landnutzung und Umwelt
der Technischen Universität München
zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender : Univ.-Prof. Dr. W. Höll

Prüfer der Dissertation :

1. Univ.-Prof. Dr. K.-H. Schleifer
2. Univ.-Prof. Dr. St. Kramer

Die Dissertation wurde am 11.07.2005 bei der Technischen Universität
München eingereicht und durch die Fakultät Wissenschaftszentrum
Weihenstephan für Ernährung, Landnutzung und Umwelt am
01.12.2005 angenommen.

Contents

1. Introduction	5
1.1. Ribosomal RNA based taxonomy.....	6
1.2. Structural implications of rRNA.....	6
1.3. Conservation of structural features in rRNA.....	7
1.4. Fluorescence <i>in situ</i> hybridization and rRNA Structure.....	8
1.5. Evaluation of multiple sequence alignments.....	9
2. Objectives of the study	11
3. Implementation	13
3.1. Programming language and the Graphical User Interface (GUI).....	13
3.2. Annotations and Sequence Data.....	13
3.3. Supplementary Data.....	14
3.4. Multiple sequence alignment and phylogenetic treeing.....	15
3.5. The positional tree (PT) server.....	15
3.6. Probe design and probe match.....	15
4. SECEDIT: A Tool to Visualize Secondary Structure of ribosomal RNA	17
4.1. Background.....	17
4.2. Description of the tool.....	18
4.3. The interface.....	19

4.4.	Arranging the layout of the structure.....	21
4.5.	Labeling the structure.....	22
4.6.	Configuring structure layout.....	23
4.7.	Mapping other rRNA sequence data.....	23
4.8.	Superimposing oligonucleotide probes.....	24
4.9.	Superimposing sequence associated information.....	25
4.10.	Related work	29
4.11.	Discussion.....	29
5.	RNA3D: A Tool to Visualize Three- Dimensional Structure of rRNA	33
5.1.	Background.....	33
5.2.	Description of the tool.....	34
5.3.	Molecule Display.....	35
5.4.	User Customization.....	36
5.5.	Navigation.....	37
5.6.	Mapping Secondary and Tertiary Structure Interactions of rRNA.....	37
5.7.	Mapping other rRNA sequence data onto <i>E. coli</i> template structure.....	39
5.8.	Mapping oligonucleotide probes onto the molecule.....	43
5.9.	Mapping sequence associated information (SAI) onto the molecule.....	46
5.10.	Related work.....	47
5.11.	Discussion.....	51
6.	SAI viz: A Tool to Visualize Sequence Associated Information (SAI) in ARB Primary Structure Editor	53
6.1.	Background.....	53
6.2.	Glimpse of the tool.....	54
6.3.	Selection of SAI	54

6.4.	Color translation table.....	55
6.5.	Visualization of SAI	55
6.6.	Related work	59
6.7.	Discussion.....	59
7.	SAIprobe: A Tool to Visualize SAI for Oligonucleotide Probes	61
7.1.	Background.....	61
7.2.	Glimpse of the tool.....	62
7.3.	Probe design and match	62
7.4.	Selection of SAI.....	63
7.5.	Visualization of SAI and probes	63
7.6.	Example	64
7.7.	Related work	65
7.8.	Discussion.....	65
8.	CONCAT: A Tool to Concatenate Sequence Data	71
8.1.	Background.....	71
8.2.	Glimpse of the tool.....	72
8.3.	Sequence data.....	72
8.4.	Concatenation.....	73
8.5.	Related work	74
8.6.	Example.....	74
8.7.	Discussion.....	80
9.	Summary	83
	Summary (English).....	83
	Zusammenfassung (German).....	87
	Appendices	91
A.	Acronyms.....	91
B.	Definitions.....	95

Acknowledgements	101
References	103

Chapter 1

Introduction

“What is the use of a book without pictures or conversations?”

- Lewis Carroll (in *Alice's Adventures in Wonderland*)

The alliance of biology and computer science has presented us with a large library of books in the form of databases, text files, spreadsheets and other types of data source. For many of these sources, it is difficult to make sense of the data without the use of graphics. One area of bioinformatics, data visualization, aims to add pictures to the books, and many of the groups involved in such projects are looking at ways to compare genomes, or parts of genomes, graphically. The scope of such comparisons can range from a tiny part of a genome, such as a gene sequence, to an overview of the relative arrangements of gross chromosomal segments throughout the genomes of two or more species. Visualization of molecular data has become more important for many data-rich disciplines. In biology, where data sets are becoming larger and more complex with the structural data, graphical analysis is felt to be ever more pertinent. Although some patterns and trends in data sets may only be determined by sophisticated computational analysis, viewing data by eye can provide the users with an extraordinary amount of information in an instant. Graphical software tools for comparative sequence analysis to visualize the data sets along with the structural data dynamically can provide the researchers with a powerful view of the differences and similarities between the gene sequences and subsequent evaluation *in silico*.

1.1. Ribosomal RNA based taxonomy

With the disclosure of the molecular structure of DNA (Watson and Crick 1953) and the recognition of macromolecules as documents of evolutionary history (Zuckerlandl and Pauling 1965), the microbial systematics has changed dramatically. The use of conserved structures within the microbial cell for identification and phylogenetic classification became more obvious. Ribosomal RNA (rRNA), a key molecule in protein synthesis with its ubiquitous presence and highly conserved nature, soon became the basis for modern molecular taxonomy. Presence of highly conserved and variable regions within the molecule aids in deducing the evolutionary relationships among the rRNA sequences at different levels (from kingdoms down to species). With the availability of hundred thousands of rRNA sequences, rRNA genes have become firmly established as a useful systematic database across the entire breadth of life.

1.2. Structural implications of rRNA

Ribosomal RNA function is largely determined by its structure (Noller 1991), and because the general structure of rRNA is universally conserved across all taxa that have been examined (Woese 1987; Gutell et al. 1994), the secondary structure of rRNA, even when not universally identical across taxa, is more highly conserved than are nucleotides. Since one of the most basic principles of systematics is that only homologous characters can provide meaningful markers of genealogical descent, clearly, the accuracy of a phylogeny from molecular data is critically dependent on the accuracy of sequence alignment. When there is a significant variability between the sequences due to insertions/deletions/mutations, which occurred during the course of evolution, the alignment of such sequences becomes more difficult and problematic. Given that the number and character of positional differences between the aligned sequences are the basis for the inference of relationship, the primary alignments must be evaluated against certain criteria before processing with the treeing algorithms in order to reduce such ambiguities. By using secondary structure to "anchor" homologous positions, many of the inherent problems of aligning rRNA sequences can be reduced.

The RNA component of the ribosome folds, bends, and pairs with itself, forming a complex secondary structure that includes both long range hydrogen-bonded stems and

hairpin stem-loops. They attribute to the formation and function of ribosomes (Noller 1984). Comparative analysis, based on the observation that structural aspects of these molecules are highly conserved even when the sequences have diverged significantly (for example, across kingdoms) “has been the primary instrument for inferring higher order structure” of both large and small subunit rRNAs (Noller et al. 1981; Brimacombe 1981; Woese et al. 1983; Gutell et al. 1994). The unique properties of rRNA demonstrate that the evolution of rRNA genes must be considered based on the structural constraints. Furthermore, it is supported by the role of functional pressure acting upon the structure in preserving the common core of higher order structure, which is evident by the participation of 67% of the rRNA residues in helix formation by intermolecular base pairing. Consequently, the compensatory mutations that occur to maintain the integrity of rRNA secondary structure may not be revealed from the rRNA sequence alone. Furthermore, not all aligned nucleotide positions or all types of substitution changes can be treated equally in terms of phylogenetic relevance because the nucleotides within the rRNA molecule are involved in different kinds of interactions, including both hydrogen bonding to other nucleotides within the molecule and interactions with ribosomal proteins and other RNA molecules (for example, transfer RNAs). Additionally, the information about positional conservation and/or variability has been compiled over the years (Cannone et al. 2002). But such interactions and positional information are difficult to describe without the reference to rRNA secondary structure.

1.3. Conservation of structural features in rRNA

Ribosomal RNA structural motifs (loops, stems, etc.), interactions with ribosomal proteins and other RNA molecules, and conservation profiles, are useful for determining proper weighting (in order to accord different kinds of mutations/substitutions) during phylogenetic reconstruction of rRNA sequences. As first step in evaluating the validity of the various weighting schemes as well as hypotheses of positional homology is the ability of systematists to visualize rRNA structural features. In this regard, the comparative analysis of thousands of rRNA sequences has yielded more reliable RNA structure models (Gutell et al. 2002), which are well established and routinely used in the structure based phylogenetic studies. And with the availability of high-resolution RNA crystal structures for the 30S (Wimberly et al. 2000) and 50S (Ban et al. 2000)

ribosomal subunits, the accuracy of the such models of rRNA were validated with the presence of nearly 98% of the base pairings in the RNA crystal structures (Gutell et al. 2002). Thus, the representation of secondary and tertiary structures of ribosomal RNA in an intuitive way facilitates both simultaneous alignment (of primary and secondary structure) and evaluation by anyone who is familiar with making phylogenetic decisions from the individual characters.

In order to improve the alignment of rRNA sequences, and to improve the ability to describe and analyze molecular sequences, conserved secondary structures can be used as reference models. Additionally, they may also serve as valuable "proofreading" function (de Rijk et al. 1994). Furthermore, structural aspects of rRNA offer a number of opportunities for informed examination of the data for phylogenetic information. Different positions of the molecule may have very rigid evolutionary constraints connected with their function. Features such as base-paired versus non-base-paired nucleotides, compensatory mutations, and types of structure may be examined. Conclusions such as the stem regions are more highly conserved than the single-stranded regions (de Rijk et al. 1994; Maidak et al. 1994) can be better evaluated with the visual representations of the rRNA structures.

Finally, ribosomal RNA is widely studied outside the field of systematics, and as the understanding and integration of the structure and function of these molecules increases, changes in methodology of analysis of rRNA data are likely to occur, along with the biologically sound recommendations for character weighting (Kjer 1995). In this regard, integration and visualization of rRNA structural (secondary and tertiary) features attract due importance with respect to comparative analysis of rRNA sequences demonstrating the need for visualization and description of ribosomal RNA structural features.

1.4. Fluorescence *in situ* hybridization and rRNA structure

Information derived from comparative rRNA sequence analysis has been extensively applied in microbial ecological studies. Presence of highly conserved and variable regions within the rRNA sequences is more often used to identify oligonucleotide target regions unique to phylogenetic entities, for use as taxon-specific hybridization probes or

PCR primers. With the development of the fluorescence *in situ* hybridization (FISH) technique, FISH has become an integral part of the rRNA approach for microbial ecological studies (Amann et al. 1995). The 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. One of the hurdles in carrying out successful hybridization of rRNA sequences is the probe target site accessibility within the cell. This is believed to be the result of the target sequence in the rRNA being inaccessible due to strong interactions with ribosomal proteins and/or highly stable secondary and tertiary structure elements of the rRNA itself (Amann et al. 1995). So, thorough *in silico* evaluation and visualization of such probes and targets are necessary. Particularly, when probe targets are intended to be used for FISH experiments, evaluation with respect to higher-order structure (secondary and tertiary structure of rRNA) is highly important as the hybridization is carried out with nearly native conformations of ribosome molecules.

1.5. Evaluation of multiple sequence alignments

Computational analysis of nucleotide and protein sequences is among the most frequently encountered activities in bioinformatics research. Especially when it comes to sequence-based phylogenetic analyses, compilation and association of information with respect to individual or multiple sequences are of more importance and greatly influence the process of reconstruction and interpretation of phylogenetic trees. The alignment of primary structures (identification and arrangement of homologous positions in common columns) is a critical step in inferring phylogeny of the sequences. When there is a significant variability between the sequences (insertions/deletions/mutations which occurred during the course of evolution), alignment of such sequences becomes a daunting task. So the primary alignments must be evaluated against several criteria before processing with the treeing algorithms. Usually, the sequence alignments are evaluated by including filters, conservation profiles (calculated by simply determining the fraction of the most frequent character), positional variability (the rate of change or the likelihood of a given character state with respect to an underlying tree topology according to parsimony criteria, or by using maximum likelihood approach), maximum base frequency filter and the higher-order (secondary and tertiary) structure

information of ribosomal RNA. All these filters/masks, column statistics, structural information, accessibility and variability maps, and any sequence specific data (either retrieved from literature or designed by experimental observation) must be taken into account during comparative sequence analysis and phylogeny inference.

However, the computational tools for sequence analysis are often specialized in producing only one kind of feature, and frequently in text output format. But the textual representations of multiple sequence alignments are poorly informative. So, the graphical interfaces allowing users to color or shade residues (amino acids or nucleotides) according to various criteria such as physico-chemical properties, degree of conservation within the alignment, secondary and tertiary structure, etc., are highly preferred. The use of colors is very helpful to interpret a multiple sequence alignment. It gives a much more comprehensive view of the information embedded in a multiple alignment than a simple textual representation. Besides, such interfaces propose several interesting facilities such as –

- i) Manual expertise to check or refine alignments. It is necessary to examine the alignment to check that there are no obvious errors. To verify that local similarities detected by pair-wise sequence comparisons are preserved in the multiple alignment. In most cases, several parts of the alignment require further refinement and evaluation. Experienced users are often able to recognize residues that have been misaligned. Such misaligned regions are more frequently revealed by the inclusion of external information (for e.g., secondary and tertiary interactions, three-dimensional structures).
- ii) Annotating alignments and extracting sub-alignments – usage of different colors or shades allows users to annotate alignments (for e.g., to indicate the location of relevant features such as enzyme active sites or RNA secondary structure). Such locations of annotations can be used to define blocks in the alignment and thus to extract sub-alignments (profiles). This feature is particularly useful when building phylogenetic trees where one needs to exclude unreliable parts of alignments (i.e., regions for which the alignment is ambiguous).

Therefore, graphical tools which are capable of achieving intuitive visualization of molecular sequence data dynamically are highly in demand, particularly in data-rich disciplines of biology.

Chapter 2

Objectives of the study

Since biology is a highly visual science, there is always 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. Any publications related to molecular biology contain significant number of illustrations in the form of figures and drawings especially with respect to molecular sequence data. For the reason being, the molecular data in the form of illustrations are easy to interpret and are highly understandable. For example, structural representations of ribosomal RNA sequences are widely used to evaluate multiple alignments of rRNA sequences.

In order to carry out comprehensive analysis of rRNA sequences, the visualization of higher-order structure information of ribosomal RNA is very vital. The tools to evaluate multiple sequence alignments and oligonucleotide probes, intuitively, are always in great demand for comparative sequence analysis studies. So, the following structure based visualization tools were aimed to be developed and integrated into ARB software package.

- A tool to visualize secondary structures of small and large subunit ribosomal RNA: It should be based on the well established comparative RNA structure models and should offer a graphical platform for the researcher to evaluate and analyze the rRNA sequence data against secondary structure features of rRNA.
- A tool to visualize three-dimensional structure of ribosomal RNA: In order to have a system that is capable of displaying three-dimensional structural

information, overlaying any sequence associated information, and also establishing a dynamic interaction between the applications of ARB software in a more intuitive and flexible way.

- A tool to overlay any sequence associated features (both statistical and non-statistical data) onto the linear structure of the individual or multiple sequence alignments in a holistic and more perspective way.
- A tool to provide an intuitive graphical platform for designing, evaluation and visualization of oligonucleotide probes. Using such an interactive graphical user interface, researchers should be able to gain more insights by visually examining the characteristics and criteria of the probe targets and probable binding regions.
- A tool to interactively merge similar sequences and to perform concatenation of molecular sequence data in a more intuitive way. Owing to rapid growing of genome sequence databases, such a tool should allow researchers to carry out multigene studies using the concatenation approach to infer phylogenies within the ARB environment.

Chapter 3

Implementation

3.1. Programming language and the Graphical User Interface (GUI)

Most of the code was written using C++ language to avail the speedy processing. To visualize three-dimensional ribosomal RNA structures, OpenGL graphics library was used. The GUI was implemented using Open Motif and X windows library. The ARB software was developed for UNIX systems and their derivatives (Ludwig et al. 2004). So, the tools which were developed and integrated under this study are currently used within the UNIX/LINUX environments.

3.2. Annotations and Sequence Data

Periodically retrieved raw gene data comprising small subunit rRNA from public databases such as EBI (Kulikova et al. 2004), Genbank (Benson et al. 2004), the RDP (Cole et al. 2005), 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 (<http://www.arb-home.de>). The public release of small subunit rRNA database was taken for the entire study – during development, testing and in example applications. Annotations visualized in all the applications of ARB are according to the public database entries and any additional annotations from the ARB curators and researchers may also be included. Additionally, wherever genome sequence data is employed, the necessary genes were extracted from whole genomes using ARB Genome package (<http://www.arb-home.de>).

3.3. Supplementary Data

Secondary and tertiary structure information of small and large sub-unit of ribosomal RNA (comparative structure models) were retrieved from comparative RNA website (<http://www.rna.icmb.utexas.edu>) and are used as template structures for SECEDIT tool (see chapter 4).

The 3D co-ordinates from the homology model of the atomic structure of the *E. coli* 30S ribosomal subunit (PDB entry 1M5G) and the crystal structure of the *Thermus thermophilus* 30S ribosomal subunit (PDB entry 1J5E) are retrieved from the protein data bank (PDB) (<http://www.rcsb.org/pdb/>) and are used as a reference structural co-ordinates for the RNA3D tool (see chapter 5). The secondary and tertiary interactions of rRNA, incorporated into the rRNA 3D models, were retrieved from the comparative RNA website (<http://www.rna.icmb.utexas.edu>).

The RNA-protein interaction data with respect to small sub-unit ribosomal RNA (16S rRNA) were collected from Stern et al. (1988). These cross-link sites were converted to sequence associated information (SAIs) and were used for mapping onto the 2D (SECEDIT) and 3D (RNA3D) ribosomal structure models.

In situ probe accessibility data of 16S rRNA were compiled from Behrens et al. (2003b). A consensus model for the accessibility of small sub-unit rRNA (16S rRNA) to oligonucleotide probes was retrieved and the (probe accessibility) data was converted in to SAIs and used in SECEDIT (see section 4), RNA3D (see chapter 5), SAIviz (see chapter 6) and SAIprobe (see chapter 7) tools.

Additionally, conservation or base composition profiles, filters for including or excluding particular alignment positions and other column statistics, that were included in the testing and example applications, were performed using respective tools of ARB software package (Ludwig et al. 2004).

3.4. Multiple sequence alignment and phylogenetic treeing

Multiple sequence alignments were generated using integrated ClustalW (Thompson et al. 1994) and ARB fast aligner (Ludwig et al. 2004). For phylogenetic treeing purpose, the ARB parsimony, a special treeing tool, was used. Alternative treeing methods were also available in the ARB software package (Ludwig et al. 2004).

3.5. The positional tree (PT) server

The PT-Server (Ludwig et al. 2004) 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. One such PT-server was established using the current release of SSU rRNA database (<http://www.arb-home.de>) for probe design and probe match. The SAIprobe tool (see chapter 7) readily connects to the PT-Server through the probe match tool to retrieve the potential probe targets.

3.6. Probe design and probe match

The small subunit rRNA database 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.

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 (Suggs et al. 1981), 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, physicochemical 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 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

Chapter 4

SECEDIT: A Tool to Visualize Secondary Structure of ribosomal RNA

4.1. Background

In any phylogenetic study, alignment of the less conserved and the variable-length regions is often problematical, even when the sequence divergence is moderate (e.g., between vertebrate orders). For regions of ambiguous alignment, a critical consideration is how these sites should be treated in phylogenetic analyses. Furthermore, automated sequence alignment, as with other steps in phylogenetic reconstruction, requires selection of parameter values whose most appropriate settings can be difficult to determine in advance. Inclusion of ambiguously aligned regions can support erroneous patterns of branching, while removal of such regions can reduce resolution (Gatesy et al. 1993; Wheeler 1994). Although there were methods proposed (Wheeler et al. 1995) to solve such ambiguities, any small changes made to the parameter values of alignment programs will result in misalignment of well-conserved motifs (Hickson et al. 2000).

The implied assumption is that all 16S and 23S rRNAs have the same general secondary and tertiary structures, regardless of the extent of conservation and variation among the linear (primary structure) sequences (Gutell et al. 2002). Additionally, the functional pressures acting upon rRNA secondary structure conservation may differ from the primary structure of the sequence. Functional pressure apparently dictates the evolutionary preservation of a common core of secondary or higher order structure which is manifested by the potential participation of 67% of the

residues in helix formation by intermolecular base pairing (Ludwig and Klenk 2001).

Use of secondary structure information is valuable for helping to determine what criteria generate realistic alignments. For example, consideration of conserved secondary structural elements (motifs) those are common to members of a family. Such criterion helps to generate biologically reasonable alignments which indicate how well a family of rRNA sequences is aligned when compared with the known conserved features in order to resolve the underlying evolutionary relationship among the rRNA sequences. It also provides a way of objectively evaluating the performances of different programs (McClure et al. 1994). Knowledge of secondary structure is very helpful in order to delimit the range of parameter values which preserve the alignment of well-conserved motifs. When the secondary structure features such as loops and paired regions identified in rRNA are considered, the alignment ambiguity can be greatly reduced producing more meaningful alignments of rRNA genes. The studies conducted to evaluate the performance of several multiple sequence alignment programs (CLUSTAL W, Divide & Conquer, Malign, PileUp and TreeAlign) in relation to secondary structure features of ribosomal RNA have shown that the inclusion of secondary structure information greatly increases the accuracy and reliability of phylogenetic reconstruction from the multiple sequence alignments produced by such programs (Morrison and Ellis 1997; Hickson et al. 2000).

The comparative analysis of thousands of rRNA sequences has yielded more reliable RNA structure models (Gutell et al. 2002), which are well established and routinely used in structure based phylogenetic studies. The SECEDIT tool uses such RNA structure models to offer a platform for the evaluation and analysis of rRNA sequence data against secondary structure features of rRNA.

4.2. Description of the tool

The preliminary version of SECEDIT tool of ARB is substantially improved with respect to the following objectives –

- To be able to visualize the secondary structure information of ribosomal RNA.

- To provide user with more flexibility to arrange the structural information interactively.
- To dynamically map other rRNA sequence onto the template structure.
- To map the oligonucleotide probe targets and any other search strings in the structure.
- To superimpose any sequence associated information (SAI) onto the structure.
- To generate a 2D drawings of rRNA secondary structure for publications.

ARB curated database (<http://www.arb-home.de>) includes the secondary structure information from the well established comparative structure models of rRNA (Gutell, 1994; Van De Peer, 1999; De Rijk, 2000). The rRNA structural information is incorporated in the alignment by interspersing the sequence with special symbols denoting the start (" $[$ ") and the end (" $]$ ") of the structural features. The bases participating in helices or stems are denoted by separate symbols (symbols " $<$ " and " $>$ " indicate 5' and 3' base pair position in the helix region, respectively). A special "helix numbering line" contains the names of the helix strands, and indicates which are complementary. Since these forms of structural information in the primary alignment are very difficult to evaluate, SECEDIT generates the classical 2D drawing of the secondary structure, which is easier to grasp.

4.3. The interface

SECEDIT does not try to produce non-overlapping drawings, so the first drawings produced for a new molecule might show a considerable overlap or have a minimal structure. However, this structure can be easily built and arranged interactively according to the user's wishes (see section 4.4). The main window is used fully to display the structure (Figure 4.1). A user customizable menu bar, pop-up dialog boxes and switch-able modes control the application. Several user-definable parameters control how a newly created structure will be drawn. Among others, the general distance between bases in single stranded areas, the distance between bases in a helix and between the bases in a base pair can be set. By default, all base pairs are classified into strong, normal, weak and non-base pairs. Optionally, user can also define additional base pair as "user pair". Each base pair is assigned with a symbol (for e.g., line or dot) and are shown connected using the respective

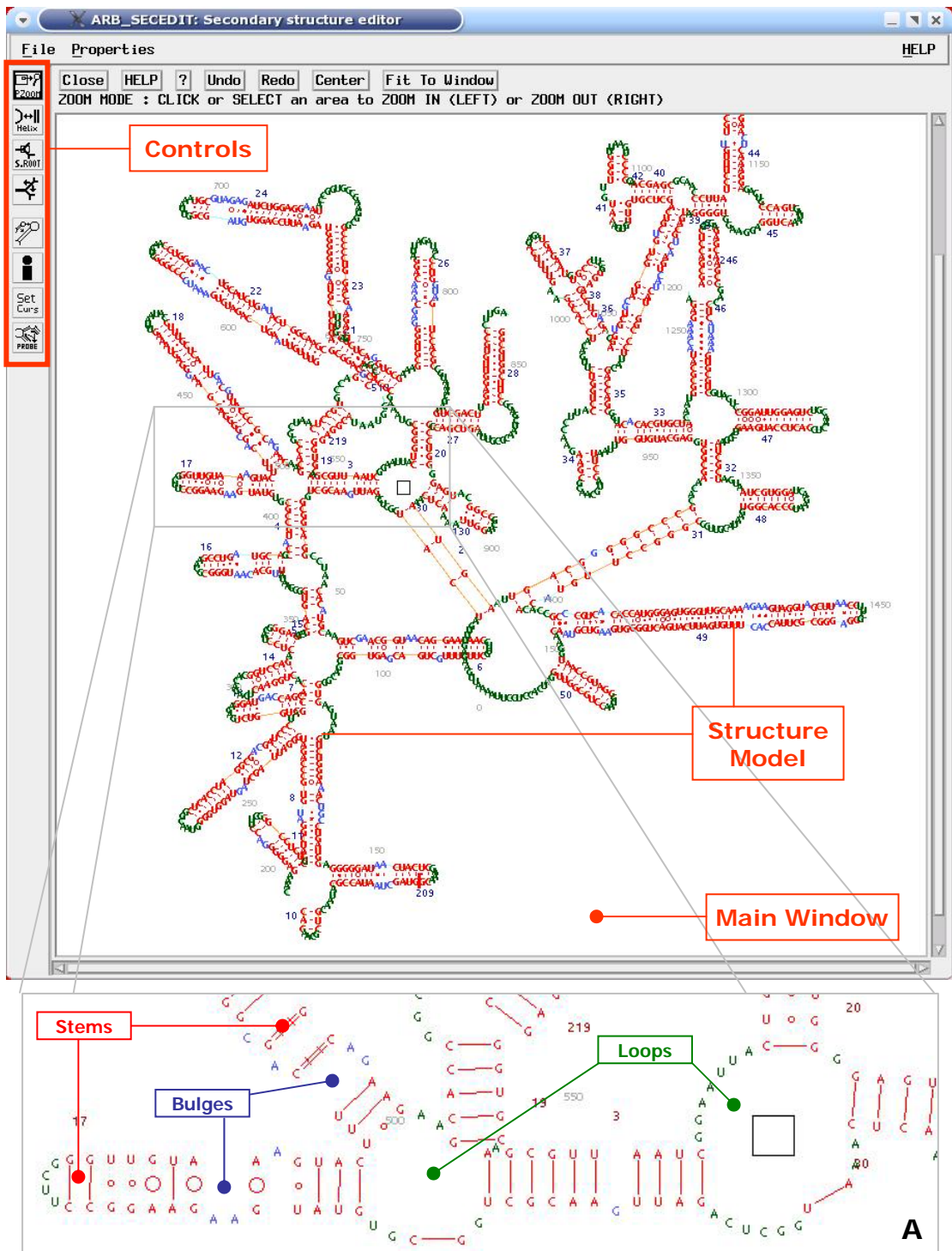


Figure 4.1. Secondary Structure Editor. Screenshot showing the interface of SECEDIT tool. The main window contains the consensus model of rRNA to which any rRNA sequence can be mapped. Here, 16S rRNA secondary structure of *E. coli* is shown. The various structural motifs such as stems (colored red), loops (colored green) and bulges (colored blue) are shown in detail in the inset (A). Numbers displayed in grey and smaller in size denotes the respective nucleotide positions whereas numbers of larger size shown in red reflect the helix numbers according to the ARB numbering scheme. Zooming, rotating and rearrangement of structure can be done using the controls (shown with red box) found on the left panel. Colors and layout settings are done using “Properties” menu.

symbols. Both width and length of the connections between the bases of standard and non-standard base pairs can be changed independently.

The structures can be saved and loaded at any time during the application using the File menu. The entire structure can be exported to external graphic application (Xfig) in order to generate publication ready images.

4.4. Arranging the layout of the structure

The structural information of rRNA is dissected into stems (strands), loops (non-base pairs) and bulges (unmatched bases within the strands). The user can set the root (centre of the molecule) from where the helices and loops ascend. By selecting the "helix mode" displayed at the left pane of the window (Figure 4.1; controls), user can build or collapse the helix strands by clicking on the bases participating in helix formation. Using the mapping feature, one can easily identify such regions in the initial structure, when secondary structure information is superimposed in different colors in the background (see section 4.9). Using "rotate" and "stretch" mode, user can arrange the entire structure interactively according to his/her needs. In "rotate" mode, clicking on a base that belongs to a helix will select the entire helix. And the structure can be rotated by dragging into a different position. With the left mouse button, the entire structure following the selected helix will be rotated and with the right mouse button only the selected helix strand is rotated. Additionally, rotation can also be achieved by clicking and dragging the helix number using left or right mouse button. The center of rotation is indicated by a red box in the loop region of the structure. It can be repositioned any time by clicking other loop regions in the structure by using the left or right mouse button. With the "stretch" feature, user can stretch the structure by expanding or contracting the helix and loop regions of the structure. This feature helps the user to view the specific regions more clearly and distinctly. At any given time, the entire structure can be moved in 2D space by holding the middle mouse button and dragging. This enables the user to scroll through the structure when it falls outside the visible area of the window. Scaling (zoom) of the structure is achieved by switching to "zoom" mode. Zooming in to the molecule is performed by left mouse clicks and zooming out of the molecule is done using right mouse clicks. Alternatively, user can also specify the area (drawn as

rectangular box) by clicking and dragging the mouse buttons to expose the different sections of the molecule (Figure 4.1).

When mapping a rRNA sequence onto the consensus secondary structure model, different sizes (length) of helix strand are also taken into account in fitting the selected rRNA sequence to the consensus model. Helix strands are resized (by expanding and contracting) based on the base pairs (in underlying rRNA sequence) participating in helix formation.

The SECEDIT tool directly communicates with the ARB primary structure editor (Figure 6.1) and the underlying database. Any change or shifting in alignments done in the primary structure is immediately updated in the secondary structure of rRNA. Additionally, one can set the cursor position by clicking any base in the structure using "set cursor" mode. This allows the user to locate the exact position in the linear structure of the sequence alignments (in ARB primary structure editor) and to enable the user to quickly inspect the region he/she is looking at, in one glance.

4.5. Labeling the structure

All helices are automatically labeled with their helix name. ARB helix numbering scheme is followed in the entire application. For 16S rRNA, helices are numbered from 1 to 50. Another type of label is the base numbering. Base numbers with respect to the primary sequence data can be displayed with the intervals of 50 base positions. Optionally, when the set of search strings are mapped onto the structure, respective labels can be displayed at the site of location. At any given time, the detailed information of current species mapped to the structure can be displayed by clicking on the "species info" mode. Species info mode opens a pop-up window and the entire annotation of the respective species, contained in the ARB database, is displayed.

4.6. Configuring structure layout

Each object (stem, loop, bulge, bond, label, etc.) has properties such as font, size, and color. The "Properties" menu opens the respective pop-up dialogs to configure such objects and the layout of the structure. Users can either change all the properties globally affecting the entire structure or can limit their settings to specific type of objects such as bases in the helix strand or in loops, base numbers, helix names or labels. The new settings, defined by the user will be updated instantly. Additionally, user can map the color settings with those of ARB primary structure editor settings. This enables the user to achieve synchronization of colors used in both primary and secondary structure editors. This feature is more useful when mapping search strings or any sequence associated information derived from sequence alignments onto the structure (see section 4.9). Using "Change display options" dialog, found under "Properties" menu, user can toggle on and off any structural features displayed. This option allows the user to display structural information which is relevant to his/her needs by emphasizing the details in the structure. Alternatively, the entire structure can be drawn as a line diagram without the bases, which is referred as "skeleton structure". Thickness and the color of the skeleton are customizable. This feature has more importance to see the distribution of any column statistics performed on the sequence alignments or probe accessibility data, when mapped onto the structure (see section 4.9).

4.7. Mapping other rRNA sequence data

The SECEDIT tool is interconnected with the ARB primary structure editor (Figure 6.1). Importing and aligning of rRNA sequences are described previously in Implementation chapter (see chapter 3). Generally, 16S rRNA sequences are aligned against the master sequence (for e.g., *E. coli* or any reference sequence) along with the 16S rRNA secondary structure model as a reference (in case of small-subunit rRNA database). Any rRNA sequence contained in the multiple sequence alignments of ARB primary structure editor can be overlaid onto the rRNA secondary structure consensus model. The user can select the rRNA sequence he/she wish to map onto the structure by the left mouse button in the multiple sequence alignment and the same will be instantly mapped onto the master structure. The

insertions, deletions, base substitutions (mutations) are readily mapped onto the consensus structure based on the structural information included in the primary alignment. The user can, at any time, edit the basic structural information (secondary structure model) to include any additional structural information, either retrieved from the external databases (<http://www.rna.icmb.utexas.edu>) or derived from his/her own experience on comparative sequence analysis. Such additional structural information to the original secondary structure model is immediately incorporated into the two-dimensional representations of rRNA sequence data in SECEDIT window. Mapping rRNA sequence data onto the consensus model helps the user to gain insight into the sequence data with respect to the structural conformations of rRNA. Since the accuracy of the phylogenetic tree is directly dependent on the proper juxtapositioning of the sequences in the alignment (Gutell et al. 2002), SECEDIT enables the user to approximate the best juxtapositioning of sequences that represent similar placement of nucleotides in their fitted structural conformation with respect to the consensus structure. Thus, enabling the researcher to reduce the alignment ambiguity and generate more biologically accurate sequence alignments for phylogeny reconstruction.

4.8. Superimposing oligonucleotide probes

Oligonucleotide probes are designed using integrated Probe Design and Probe Match tools of ARB (see chapter 3 for further details). The localization of the proposed oligonucleotide probe targets can be visualized in customizable background colors in the rRNA secondary structure model. The users have an opportunity to evaluate the probe targets with respect to structural conformations of RNA and add more confidence to the proposed probe before considering for the real experiments. This feature, coupled with RNA3D tool (see section 3.2), is profoundly useful when the probes are intended to be used in the fluorescence *in situ* hybridization (FISH) experiments, where the hybridizations are carried out with the nearly native conformations of ribosomes in permeabilized cells. Optionally, multiple probes designed by using ARB Multiprobe design tool can also be superimposed in different background colors. Sets of search strings that are defined and visualized in the linear sequences of ARB primary structure editor are, as well be mapped onto the consensus model. Additionally, the user can also evaluate the proposed

probe candidates with respect to the variability or accessibility maps (see section 4.9).

4.9. Superimposing sequence associated information (SAI)

Various column statistics like sequence consensus, base frequency, positional variability based on parsimony method and any other user defined column statistics, are performed on the sequence alignments using the integrated tools of ARB package. Any information derived from performing such column statistics on the multiple sequence alignments can be overlaid onto the consensus model using different background colors. Please see section 6.4 for details regarding SAI color translation information. With the availability of consensus model for the accessibility of the small subunit rRNA to oligonucleotide probes (Behrens et al. 2003b), probe targets can be evaluated by superimposing the accessibility maps on the rRNA secondary structure model. This feature is of more importance to the researchers who are designing FISH experiments. Secondary structure of target and neighboring regions are important for hybridization behavior and hybrid stability. Any intramolecular secondary structure formation may interfere or hinder probe-target hybridization. It is also important for designing helper probes, which (usually unlabeled) 'weaken' the secondary structure and may help to expose the inaccessible target sites (Figure 4.2). Thus, visual evaluation of oligonucleotide probes *in silico* with respect to the higher-order structure and experimental accessibility data, suggests the researcher regarding the performance of the selected probe candidates during the hybridization experiments (for additional information see section 5.9).

Any statistical and non-statistical data with respect to individual columns in the primary alignments can be readily overlaid onto the consensus secondary structure model of ribosomal RNA and the respective structural maps can be obtained using SECEDIT tool. Figures 4.3 and 4.4 displays 'positional variability map' and 'RNA-Protein interaction map' of 16S rRNA model generated by SECEDIT tool, respectively.

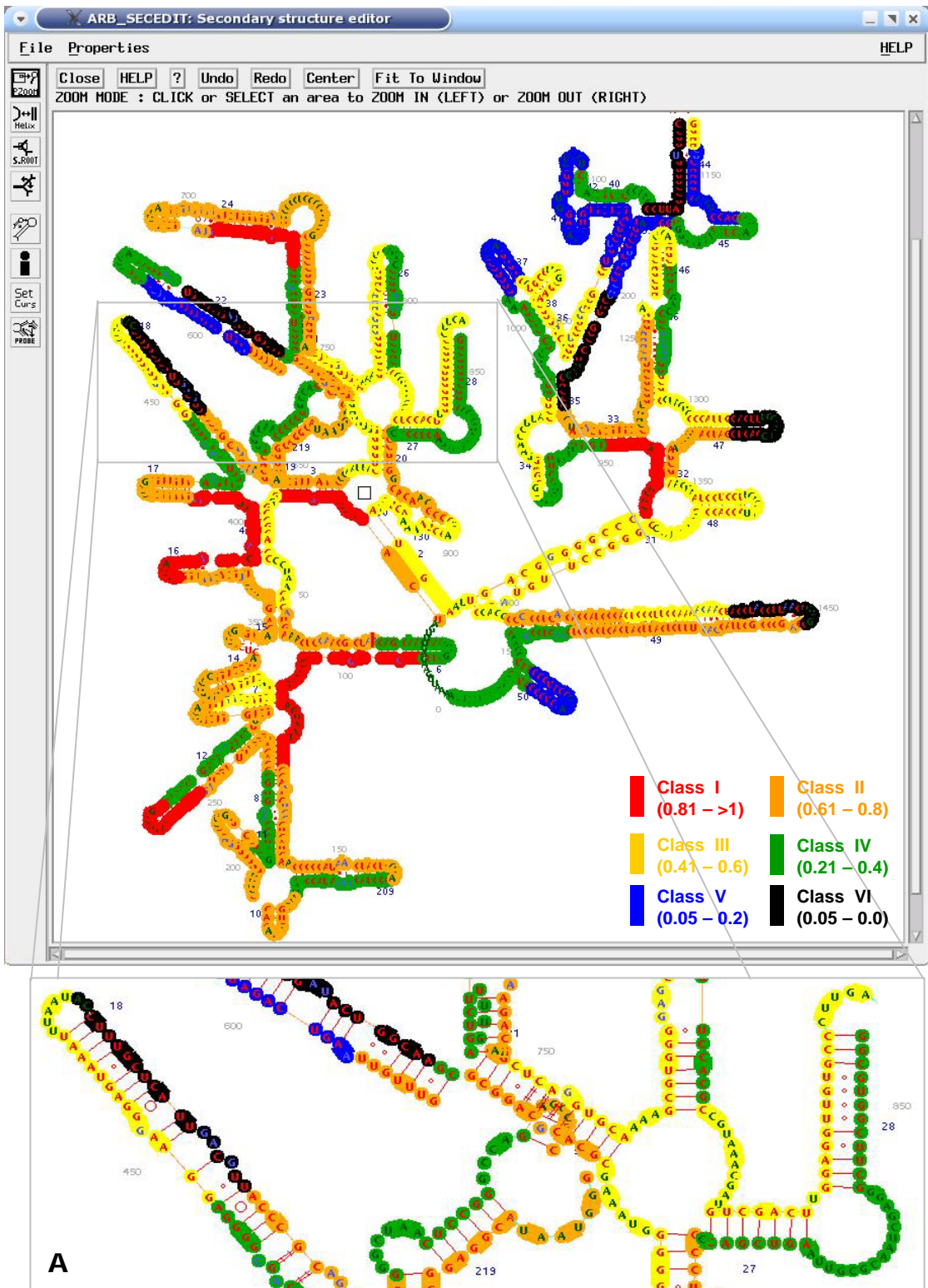


Figure 4.2. Probe Accessibility Map. Screenshot generated by SECEDIT tool showing the distribution of relative fluorescence hybridization intensities of oligonucleotide probes targeting 16S rRNA of *E. coli*. Probe accessibility data is taken from the published work (Behrens et al. 2003b). The different background colors indicate brightness range of different classes (classes I through VI) with respect to the observed fluorescence intensities. Numbers displayed in grey and smaller in size denotes the respective nucleotide positions where as numbers of larger size shown in blue reflect the helix numbers according to the ARB numbering scheme. Detail of the structure is shown in the inset (A).

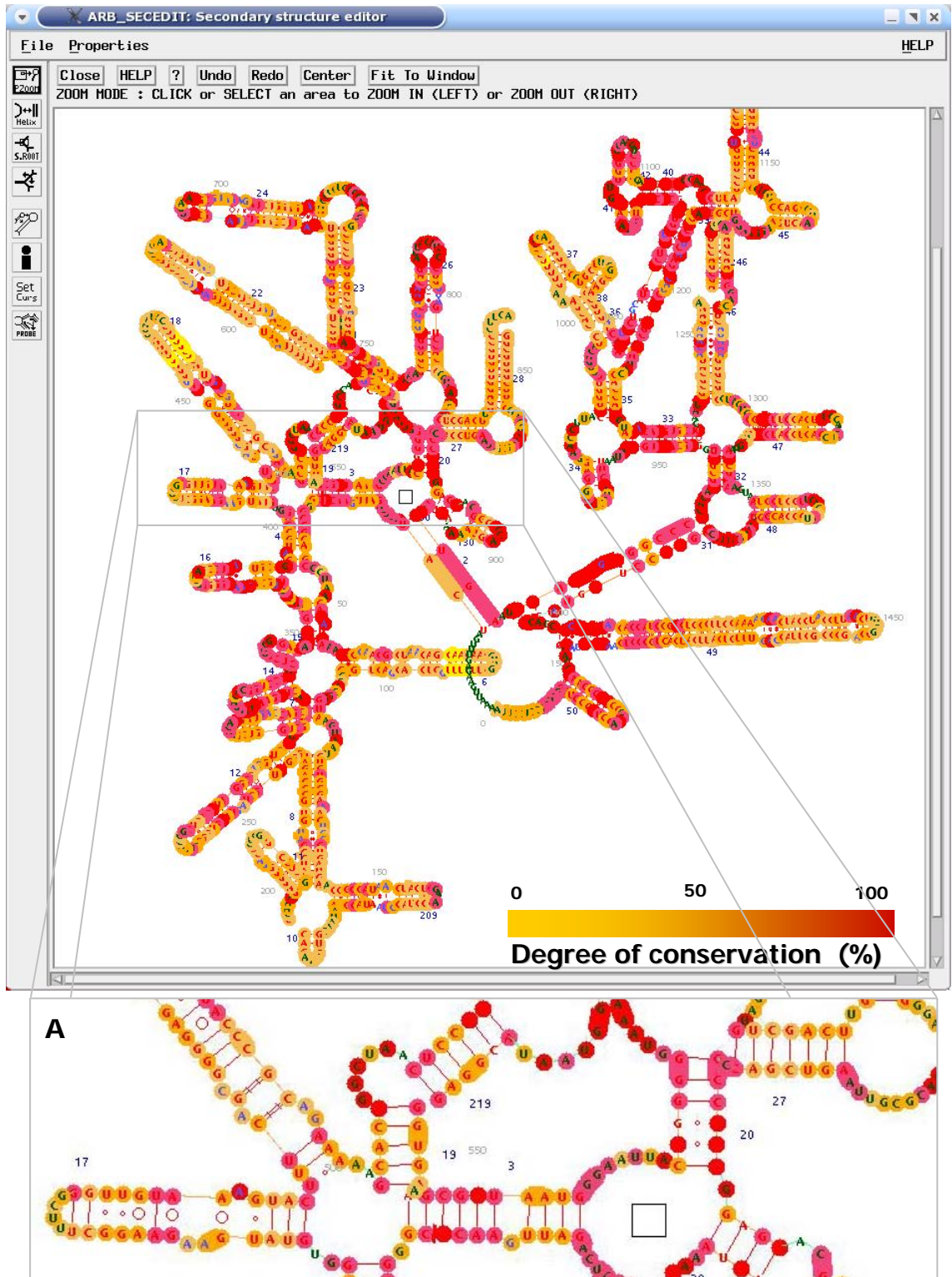


Figure 4.3. Positional Variability Map. Screenshot generated by SECEDIT tool showing the positional variability map superimposed onto the consensus model of small subunit ribosomal RNA. Column statistics are performed on multiple alignments using parsimony method and minimum number of mutations for each site is determined. The positional variability values are then overlaid onto the 16S rRNA secondary structure model base-by-base to generate positional variability maps. Bases with background inclining towards yellow are highly variable and the bases with background inclining towards red are highly conserved positions (see inset A for clear view). Positional variability values are calculated with the data set containing 1000 rRNA sequences.

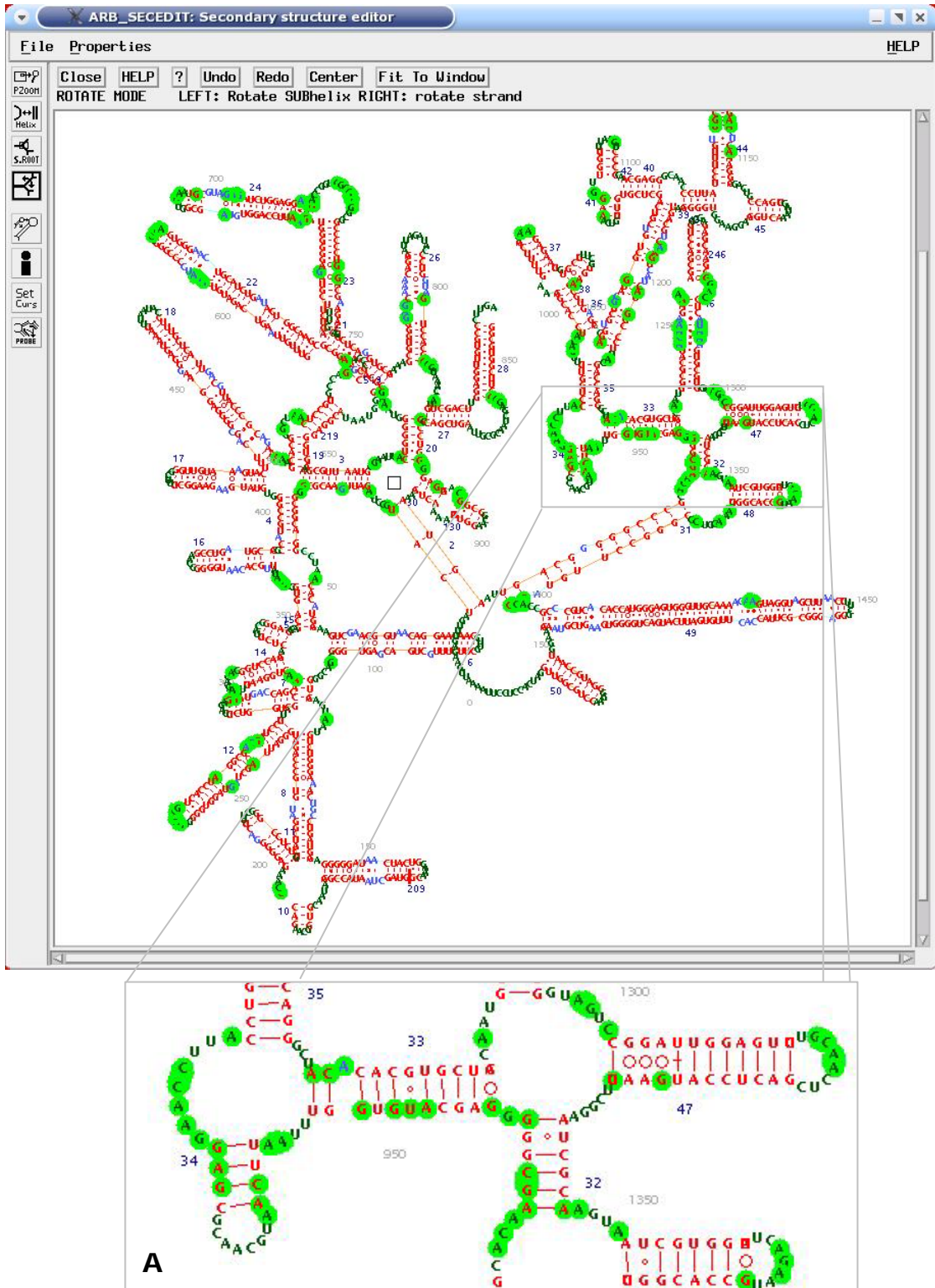


Figure 4.4. RNA-Protein Interactions Map. Screenshot generated by SECEDIT tool showing the RNA-Protein interactions. Here, the regions of 16S rRNA consensus model contacted by ribosomal proteins (S2 to S21) are highlighted in green. The RNA-Protein interaction data is fetched from the published literature (Stern et al. 1988). Detailed view of the structure with precise protein contact sites (green background) is shown in the inset (A).

4.10. Related Work

Several programs (Shapiro et al. 1984; Yamamoto et al. 1987; Cedergren et al. 1988; Martinez 1988; Gautheret et al. 1990) have been developed in order to produce two-dimensional structure drawings of rRNA. But they share some of the following problems: Most of the programs employ computationally intensive methods limiting their use for smaller molecules and are restricted only for workstations. Another problem is the lack of post-production editing; the drawings produced cannot be easily changed or annotated. This makes it difficult to emphasize structural similarities between different molecules, or to indicate peculiar areas where insertions or deletions are encountered. Programs such as CARD (Winnepenninckx et al. 1995) and XRNA (Wieser and Noller 1995) took a different approach to circumvent such problems but they were highly labor intensive. In the former, the sequences for every structural element have to be typed in where as in the latter, arranging large structures is too laborious, making them difficult to use. Although the program RNAviz (De Rijk 1997; De Rijk et al. 2003) solves these problems, it is restricted to produce only publication quality 2D drawings rather than to evaluate rRNA sequence alignments. Furthermore, it does not allow mapping of any additional information (for e.g. probe accessibility data) on to the models that are displayed, dynamically. The SECEDIT tool described in this thesis, overcomes all the drawbacks. It can be used for evaluating sequence alignments, mapping sequence associated information and also to produce publication quality 2D drawings of rRNA.

4.11. Discussion

Sequence alignment should not be a one-step process and, as with other steps in phylogenetic reconstruction, the assumptions of the alignment program must be considered and the biological accuracy should be evaluated. The identification of conserved motifs (Gutell et al. 1992; Hickson et al. 1996; Lee et al. 2003; Lee and Gutell 2004; Hoerter et al. 2004) in rRNA molecules offers an easy way to assess how an alignment reflects biological accuracy, and the visual inspection of the data can also provide the investigator with added insight into how taxa vary. Thus, the available secondary-structure models for both large and small subunit rRNAs (Gutell, 1994; Van De Peer, 1999; De

Rijk, 2000) should be used to compare and fine tune the sequence alignments prior to further phylogenetic analyses of the data.

Analysis of the patterns of sequence conservation and variation present in RNA sequence alignments can reveal phylogenetic relationships and the same has been used to predict the RNA structures (Gutell et al. 1986). Hence, the accuracy of the phylogenetic tree and the predicted RNA structure is directly dependent on the proper juxtapositioning of the sequences in the alignment. These alignments are an attempt to approximate the best juxtapositioning of sequences that represent similar placement of nucleotides in their three-dimensional structure. For sequences that are very similar, the proper juxtapositioning or alignment of sequences can be achieved simply by aligning the obviously similar or identical subsequences with one another. However, when there is a significant amount of variation between the sequences, it is not possible to align sequences accurately or with confidence based on sequence information alone. In such situations, one can juxtapose those sequences that form the same secondary and tertiary structure by aligning the positions that form the same components of the similar structure elements (for e.g., aligning the positions that form the base of the helix, the hairpin loops, etc.). Given the accurate prediction of the 16S and 23S rRNA secondary structures from the comparative analysis of rRNA sequences (Gutell et al. 2002), one can gain more confidence in the accuracy of the positioning of the sequence positions in the alignments using SECEDIT tool.

Additionally, careful observation during phylogenetic reconstruction using rRNA sequences have led to different structural inferences (Cannone et al. 2002). There are positions that are conserved within one phylogenetic group and different at the same level in the other phylogenetic groups. For example, bacterial rRNAs have positions that are conserved within all members of their group, but different from the Archaea and the Eucarya. These types of patterns of conservation and variation transcend all levels of the phylogenetic groups at each level of the phylogenetic tree (e.g., phyla Gamma-, Alpha-, Beta-, Delta, and Epsilonproteobacteria). Observation of different rates of evolution at the positions in the rRNA directed to identification of highly variable and conserved regions in rRNA sequence (Woese 1987). In addition to the different rates of evolution, many of the positions in the rRNA are

dependent on one another (positional covariation), which have led to development of secondary structure models for rRNA (Gutell 1994). Using integrated SECEDIT tool, one can use such underlying dynamics in the evolution and positional dependency of the RNA to facilitate the alignment and structural analysis of the rRNA sequences.

Studies have revealed that even small differences in sequence alignment can result in quite different phylogenies (Kjer 1995; Morrison and Ellis 1997) stressing the need for generating accurate and more reliable sequence alignments for phylogenetic reconstruction of rRNA sequences. The utilization of secondary structure information of rRNA is vital to help identify the basis for aligning and comparing rRNA sequences, and to give the researcher a better understanding of the sequences being analyzed in order to generate accurate and more reliable sequence alignments. In addition to improving the alignment of rRNA sequences, and improving the ability to describe and analyze molecular sequences, conserved secondary structures may serve as valuable "proofreading" function (De Rijk *et al.* 1994). There are many potential sources of error between the initial amplification of the DNA and the final publication of the sequences. Some stems in rRNA are so universally conserved that any sequence that is unable to fold into one of these stems should alert the investigator to the possibility of an error. For example, any incompatibility of the sequence data with the stem 18, which is found in all genomes of eukaryotes and prokaryotes sequenced to date, indicates the presence of such errors.

Furthermore, studies have shown that there are differences in higher-order structures that influence target site accessibility to oligonucleotide probes even though the small-subunit rRNA is a highly conserved molecule (Behrens *et al.* 2003a). Secondary structure of target and neighboring regions are important for hybridization behavior and hybrid stability. Any intramolecular secondary structure formation may interfere or hinder probe-target hybridization. A strong conformational effect of oligonucleotide hybridization is demonstrated in the study of Fuchs *et al.* (2000) in which unlabeled helper oligonucleotides were successfully used to increase probe-conferred fluorescence signals. Using SECEDIT tool, one can identify such probable conformational effects and design additional helper probes (using ARB probe design tool). Use of helper probes are likely to open up inaccessible target sites by introducing conformational changes during

hybridization. Thus, a thorough *in silico* evaluation of oligonucleotide probe targets with respect to rRNA secondary structure and the experimental accessibility data offers the researcher to carefully design more reliable probes in order to use them in FISH experiments.

SECEDIT offers visualization of statistical and non-statistical positional features such as number of different characters “evolutionary allowed” at the particular position, positions involved in ribosomal proteins, mRNA, elongation factor binding or interaction, antibiotic resistance, mutation or modification, etc., with respect to known secondary structure elements (motifs) of rRNA. Generally, any information derived from the full background data set can be visualized with the individual sequence in SECEDIT window. Such a feature is more crucial for gaining more insight into the sequence data during comparative sequence analysis and drawing meaningful biological interpretations.

Finally, the tools which are aiming to produce 2D drawings of secondary structure of rRNA are limited to stand-alone applications, which require manual loading of sequence alignments, structural information for able to produce such drawings. The integration of SECEDIT tool with in the ARB software package, offers an integrated platform for performing comparative sequence analysis of rRNA sequences, more comprehensively.

Chapter 5

RNA3D: A Tool to Visualize Three- Dimensional Structure of rRNA

5.1. Background

The molecular data in the form of illustrations are highly understandable to interpret the potential evolutionary and functional relationships among the underlying sequences. As more knowledge is gained with respect to rRNA higher order structure through the availability of thousands of SSU rRNA (Van De Peer, 2000; Wuyts, 2002) and LSU rRNA (De Rijk, 2000; Wuyts, 2001) sequences coupled with the high-resolution RNA crystal structures, has led to a breakthrough in the insight into evolutionary relationships between bacterial phyla (Woese 1987) and between the major eukaryotic kingdoms and protist taxa (Van De Peer and De Wachter 1997).

Methods for crystallizing and derivatizing RNA molecules like X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM) have led to the determination of several RNA structures. With the availability of high-resolution RNA crystal structures for the 30S (Wimberly et al. 2000) and 50S (Ban et al. 2000) ribosomal subunits, the accuracy of the comparative secondary structure models of rRNA were validated with the presence of nearly 98% of the base pairings in the rRNA crystal structures (Gutell et al. 2002). Each of the structure determination approaches involves painstaking preparation of rRNA samples and often months of work to obtain and interpret structural information. And the bottlenecks for rRNA structure determination are typically identification of stable, well-behaved constructs and preparation of highly

pure samples. These issues are currently addressed empirically for each rRNA to be studied, through preparation and screening of dozens of samples. In cases where one of the sequences has a known three-dimensional structure it can be more informative to compare the alignment with the solved structure, to better understand how the local environment of the nucleotides relates to conservation. It would thus be more appropriate to take the all-atom structure of ribosomal RNA of *E. coli* (Tung et al. 2002), deduced from the crystal structure of 30S ribosomal subunit of *Thermus thermophilus* (Wimberly et al. 2000), as a template structure to map the ribosomal RNA sequence of other organisms for which a crystal structure has not yet been deduced. Thorough knowledge of the three-dimensional structure coupled with the secondary structure information of rRNA is more often necessary to generate more reliable and dependable alignments in order to use rRNA sequences for the reconstruction of phylogenetic trees. With this background, the RNA3D tool has been developed and integrated to the ARB software package (Ludwig et al. 2004) which combines the sequence alignment information with the three-dimensional structures of resolved 30S and 50S bacterial ribosomal subunits.

5.2. Description of the tool

The RNA3D tool was developed keeping in mind to have a system that is capable of displaying structural information as well as interacting between the applications of ARB software in a more intuitive and flexible way. The main purposes of the RNA3D tool are to achieve the following capabilities –

- To merge the structural and phylogenetic information of underlying sequences with the structural data of rRNA.
- To allow dynamic mapping of rRNA sequence data.
- To interact with the primary structure of rRNA sequences (ARB Primary Structure Editor) in order to facilitate the fine tuning of alignments.
- To visualize any column statistics performed on the rRNA sequences contained in the ARB databases.
- To be capable of overlaying the mutation, insertion and deletion information with respect to the master sequence onto the molecule in real time.
- To combine the secondary structure information of rRNA (in the form of loops and helices) with the three-dimensional structure of rRNA.

- To be able to display the oligonucleotide probe targets designed (using ARB Probe Design tool) or imported, in the three-dimensional spatial conformation of rRNA molecule.
- To visualize actual nucleotide bases in the form of letters A (Adenine), G (Guanine), C (Cytosine) and U (Uracil) instead of the chemical structure, which is more informative and readable.
- To display base positions with respect to the master sequence within the molecule.
- To provide the user with more customizable tool which can be used according to his/her needs.

5.3. Molecule Display

The annotation of RNA three-dimensional structures consists of a preprocessing of the information embedded in their 3-D coordinates. Since the structure of *Escherichia coli* ribosome has not yet been resolved, the homology model of the atomic structure of the *E. coli* 30S ribosomal subunit which is determined (Tung et al. 2002) using the crystal structure of the *Thermus thermophilus* 30S ribosomal subunit (Wimberly et al. 2000) as the template (PDB entry 1J5E), is retrieved from the protein data bank (PDB) (PDB entry 1M5G) and used as a template structure for the RNA3D tool. In order to objectively represent the structural knowledge of three-dimensional ribosomal RNA structure, the respective 3-D coordinates were extracted from the PDB file (1M5G) and used for further structural analysis and searches.

The RNA3D tool uses the popular OpenGL graphics library combined with Open Motif user interface for achieving more intuitive rendering and manipulation of the rRNA molecule within the ARB environment. It processes PDB structural information stored in PDB file (1M5G) into the annotated structures and renders (draws) them into the virtual space using OpenGL routines. In order to provide user with the more detailed perspective of 16S rRNA structure, structural information corresponding to the ribosomal proteins were not included in the annotation. The extracted structural information is then fed to OpenGL engine, where it is further transformed into a hierarchy of OpenGL objects, which encode molecule chains, residues and base positions. At this stage, further processing may occur, for example when the user requests the mapping of secondary structure

of the RNA onto the molecule in the form of loops and stems (see section 5.6).

To achieve more performance and dynamic overlay of any sequence associated information, rendering (drawing) was simplified to chain display with the capacity to display residues in the form nucleotides – Adenosine (A), Guanine (G), Cytosine (C) and Uracil (U) at the respective coordinates in the molecule. Most of the applications which are intended to display three-dimensional structures, display the entire chemical structure of the molecule. Viewing the entire chemical structure in the molecule's 3D structure is less readable for the user. Additionally, base positions can be displayed at the respective coordinates or at intervals specified by the user. Displaying of base positions help the users i) to locate probe binding sites within the molecule, ii) to refine the sequence alignments according to the molecule structure, and also iii) to identify the exact position in the primary sequence, where insertions, deletions and base substitutions occur with respect to the template sequence when a different rRNA sequence is mapped onto the master structure (see section 5.7) (Figure 5.1).

5.4. User Customization

Since the user customization is an important consideration in graphical user interface (GUI) design, RNA3D tool provides the individual users with more possibilities to customize the interface to suit their particular purpose and preferences. As a first step toward enhancing the user customization capability of RNA3D tool, any form of annotation and information overlay can be toggled on and off. This feature allows users to focus on annotations they consider important without being distracted by information irrelevant to their particular needs. Viewing all of the structural and overlay information at once result in overwhelming displays. The feature, to toggle on and off the information displayed, becomes more essential for viewing the molecule more clearly. Additionally, the user is provided with more customization capabilities in the form of specifying different colors, shapes, letters, and the thickness of the objects rendered onto the scene at any time using Color Palette, Bases, Helix, Molecule and Mapping buttons of RNA3D tool. For example, the user can colorize the entire molecule based on the residues that are participating in loop or stem formation in the accepted

secondary structure model of 16S rRNA. Additionally, a range of colors is included in order to overlay any sequence associated information onto the molecule (see section 5.9) and also a separate range of colors to visualize the series of search strings or probe targets within the molecule (see section 5.8).

5.5. Navigation

The entire set of visualized objects can easily be rotated, translated and scaled at the user's wish. Navigation through the molecule is basically bound to the standard mouse buttons and mapped to simple keys on the keyboard. Rotation of the molecule is achieved by moving the mouse whilst holding the left mouse button. The molecule can be rotated in any desired direction (360 degrees). Additionally, the molecule can be made to rotate automatically by pressing the space bar on the keyboard. Pressing again the space bar will stop the automatic rotation of the molecule. Translation (horizontal and vertical movement of the molecule in RNA3D Window) of the molecule is performed using left, right, up and down arrow keys on the keyboard. Molecule can be scaled using the zoom function of RNA3D tool. For easy navigation, the zoom function is bound to wheel of the mouse. The user can zoom in to or out of the molecule by performing upward or downward motion of the wheel, respectively. Easy rotation, translation and scaling of the molecule enable the user to observe the buried and exposed molecule sections. This feature is more useful when the user wishes to see the binding sites of rRNA targeted oligonucleotide probes in the molecule (see section 5.8). Additionally, the user can display the current cursor position in the linear structure of sequence alignments (ARB Primary structure editor) in the molecule. Using this feature, the user can quickly locate the region of rRNA sequence he/she is examining in the both primary and three-dimensional structure of the rRNA molecule.

5.6. Mapping Secondary and Tertiary Structure Interactions of rRNA

Secondary and tertiary structure interactions of the well established comparative structure models of rRNA (Woese et al. 1983; Gutell et al. 1985; Gutell et al. 1986; Gutell et al. 1992; Gutell 1994; Gautheret et al. 1995) were

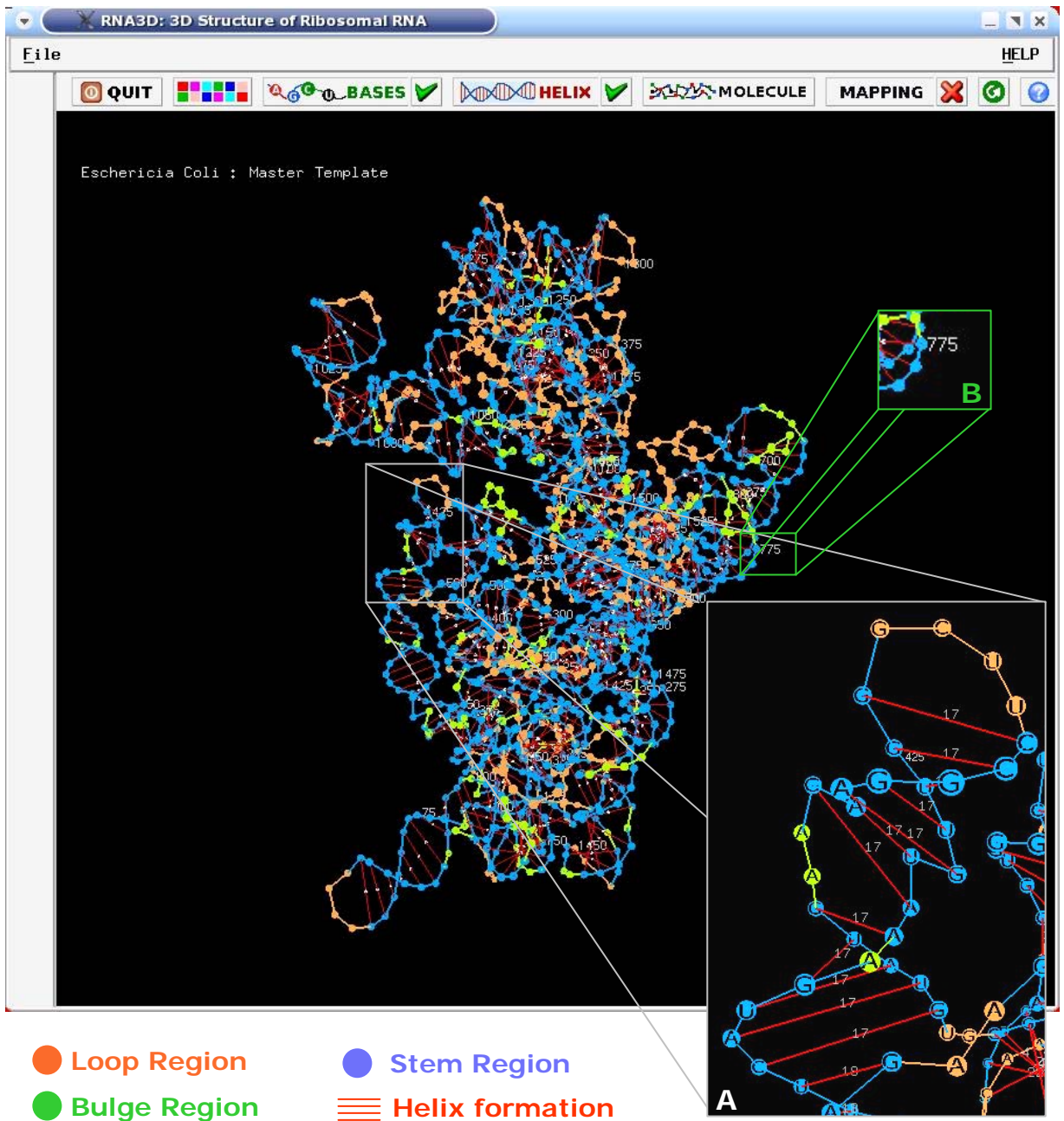


Figure 5.1. Three-dimensional Structure of 16S rRNA. Screenshot showing the interface of RNA3D tool. The entire display area is used for rendering the three-dimensional structure in OpenGL 3D environment. The molecule can be rotated, scaled (zoom) and translated (moved) by moving mouse and arrow keys. Any feature of the structure and respective settings can be toggled on or off using the informative buttons located on the top of the application. Color settings can be altered using color palate found on the top-left corner. In this screenshot, secondary structural features are combined with the three-dimensional structure of *E.coli* 16S rRNA. Residues representing loop regions and bulge regions of the 16S rRNA model are colored **orange** and **green**, respectively. And the residues participating in helix formation are colored **blue**. Part of the structure along with secondary structure interactions is shown in more detail in the inset (A). Letters A, G, C and U denote the actual residues in the 16S rRNA sequence. The numbers shown on the helices (colored **red**) represent the respective helix numbers in the secondary structure model. Respective nucleotide positions in the 16S rRNA sequence are displayed in grey (inset B).

retrieved from comparative RNA website (<http://www.rna.icmb.utexas.edu>) and fitted to the three-dimensional structure of *E. coli* master sequence. The preprocessing was done on the retrieved secondary structure interactions in order to differentiate stems (helices), bulges (unpaired bases in the helix region) and loops in the rRNA crystal structure (Figure 5.1). When variability maps are overlaid onto the structure, this feature enables the user to identify the conserved and variable regions in the small subunit of the ribosomal RNA. Mutations of single nucleotide with respect to loop and stem regions of the rRNA structure can be seen when mutation information (calculated for the overall sequences in the database) is superimposed. Optionally, the user can choose the number of stems (helix regions) to be displayed in the crystal structure along with the helix numbers (1-50 for 16S rRNA). Helix numbers are displayed according to ARB numbering scheme (Ludwig et al. 2004). Tertiary interactions observed (Gautheret et al. 1995) in the small subunit of rRNA can also be displayed in the rRNA structure. Different colors can be defined for all of the secondary and tertiary interactions, which help the user to immediately see the differences and distribution of different interactions in the small subunit of rRNA crystal structure. Additionally, intermolecular contacts i.e., between 16S rRNA bases and ribosome protein residues, which are important to stabilize the tertiary fold of the rRNA, as well as the complex formation of the ribosome (Mueller and Brimacombe 1997), can be visualized in the crystal structure. A more recent study on testing the accuracy of ribosomal RNA comparative structure models (Gutell et al. 2002) confirmed that nearly all of the predicted covariation-based base pairs, including the regular base pairs and helices, and the irregular base pairs and tertiary interactions, were present in the 30S and 50S rRNA crystal structures. Thus, the deeper insight into the rRNA crystal structure along with the secondary and tertiary interactions will have the potential to assist the user in refining the multiple sequence alignment itself when a large number of datasets is included.

5.7. Mapping other rRNA sequence data onto *E. coli* template structure

The RNA3D tool is well integrated into the ARB software and establishes the communication with the other tools and functions of ARB at any time. The rRNA sequences which are in the underlying ARB database are aligned with

multiple sequence alignment programs such as ClustalW (Thompson et al. 1994), FastAligner (Ludwig et al., 2004). Optionally, multiple sequence alignment can be checked manually in the ARB primary structure editor (Figure 6.1). Generally, 16S rRNA sequences are aligned against the master sequence (for e.g., *E. coli* or any reference sequence) along with the 16S rRNA secondary structure model as a reference. Any rRNA sequence contained in the multiple sequence alignments of ARB primary structure editor can be overlaid onto the structure of *E. coli* in RNA3D tool. The user can select the rRNA sequence he/she wishes to map onto the structure by the left mouse button in the multiple sequence alignment and it will be instantly mapped onto the master structure. The display of full name and short name (ARB ID) within the RNA3D window, in addition to the crystal structure, immediately displays the current species (rRNA sequence) he/she is mapping onto the master structure. This feature is more useful when a large number of rRNA sequences are included in the multiple sequence alignments. Also the feature is later very handy, when the user walks through the constructed phylogenetic tree selecting species in different taxa. And the selected species in the tree is dynamically mapped to the molecule in the RNA3D window. The selected rRNA sequence is annotated with mutation (base substitutions), insertion and deletion information at each site as compared to the master sequence (*E. coli*). For the regions where the sequences are aligned without deletion or insertion, direct base substitution (mutation) is applied. Because the C'---C' distance is essentially the same (~10.2 Å) in all Watson-Crick base pairs (Watson and Crick, 1953), this simple procedure preserves the base pairing and the double helical structure while substituting the bases. Although there do exist the requirement of structural adjustments for non-Watson-Crick base pairs, currently, simple base substitutions are kept because the development of new models to achieve the necessary structural adjustments is out of the scope of the RNA3D tool. In the regions where the alignment (of selected rRNA sequence) involves insertions, the respective insertion points to corresponding *E. coli* base position in the alignment are shown as down arrows in the crystal structure (Figure 5.2). The number of insertions and the participating nucleotides can also be displayed at the insertion points. In the case of regions, where deletions are observed in the alignment corresponding to the master sequence (*E. coli*), respective sites in the crystal structure are indicated as deleted, using ⚡ symbol. Additionally, the bases which are presumed to be

missing in the rRNA sequence alignments when comparing with the consensus model and/or during manual curation, are also visualized in the 3D structure. Public release of curated ARB databases (<http://www.arb-home.de>) does contain these missing bases in its multiple sequence alignments and the same can be seen in the ARB primary structure editor (Figure 6.1) as dots ("."). These anomalies are as well mapped onto the crystal structure as question marks ("?"), which more often attributed to errors during sequencing.

Overlaying of mutation, deletion and insertion information at each site of the sequence alignment when coupled with the secondary and tertiary interactions of rRNA (see section 5.6), gives the user an over all view of the current species (rRNA sequence) with respect to the resolved crystal structure. Since the accuracy of the phylogenetic tree is directly dependent on the proper juxtapositioning of the sequences in the alignment (Gutell et al. 2002), RNA3D tool enables the user to approximate the best juxtapositioning of sequences that represent similar placement of nucleotides in their fitted structural conformation with respect to the master structure. When coupled with ARB secondary structure editor (Figure 4.1; see chapter 4), more accuracy can be achieved in aligning the sequences, when a significant amount of variation exists between the sequences. For such cases, sequences that form the same secondary and tertiary structure can be juxtaposed by aligning the positions that form the same components of the similar structure elements (for example, aligning the positions that form the base of the helix, the hairpin loop, etc.). Additionally, the possibility of overlaying information derived from phylogeny, column statistics, etc., (see section 5.9), helps the user to fine tune the alignment and also provides more insight into the mapped species (rRNA sequence). Secondary structure models of ribosomal RNA were basically developed based on the comparative paradigms that the different RNA sequences can fold into the same secondary and tertiary structures and the unique structure and function of an RNA molecule are maintained through the evolutionary process of mutation and selection (Woese et al. 1983; Gutell et al. 1985). The same assumption can be extended to the three-dimensional crystal structures of ribosomal RNA as there are, at present, very few rRNA crystal structures deposited in the PDB.

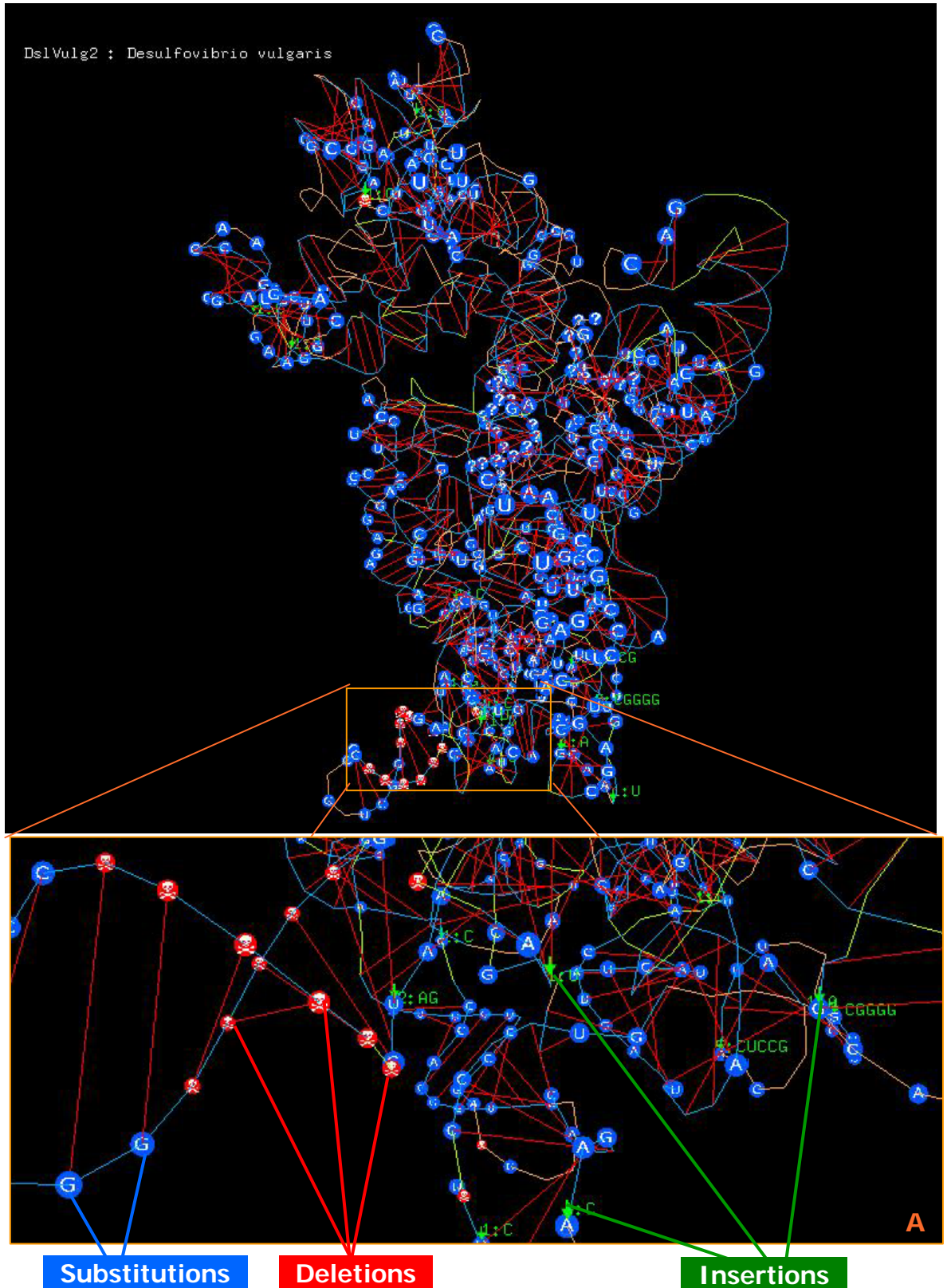


Figure 5.2. Mapping 16S rRNA sequence data. Screenshot generated by RNA3D tool showing mapping rRNA sequence data from different organism onto the three-dimensional structure of *E.coli* small subunit rRNA. In this screenshot the 16S rRNA sequence of *Desulfovibrio vulgaris* is initially aligned in reference to *E.coli* rRNA sequence and then mapped onto the *E.coli* tertiary structure. **Blue** residues represents **substitutions or mutations** where as **red** regions denotes **deletions** in reference to *E.coli* rRNA sequence. **Green** arrows with residues indicates the positions and number of plausible **insertions**. Part of the structure is shown in more detail in the inset (A).

Figure 5.3 represents the three-dimensional structures that are mapped with the 16S rRNA sequences of organisms from different taxa. Information with respect to substitutions (mutations), deletions and insertions in reference to small subunit rRNA of *E.coli* are overlaid onto the tertiary structure. By the figure, it is more evident that the closely related organism (*Klebsiella pneumoniae*) has fewer changes and the more distantly related archaea (*Halobacterium salinarum*) has greater variation compared to the reference organism (*E.coli*). Observing the entire rRNA sequence along with the differences (mutations, deletions and insertions) in one glance helps the researcher to visually inspect the quality of alignment and also hints the need for further refinement of rRNA sequence alignments.

5.8. Mapping oligonucleotide probes onto the molecule

Oligonucleotide probes are designed using integrated Probe Design and Probe Match tools of ARB (see chapter 3 for further details). The localization of the proposed oligonucleotide probe targets can be visualized in customizable background colors within the rRNA crystal structure. Using the navigation capabilities of RNA3D tool (see section 5.5), the user can get an idea about the probable binding site of the proposed probe with respect to the structural conformation of rRNA (Figure 5.4:A). Along with the secondary structural interactions (see section 5.6) the user has an opportunity to evaluate the probe targets with more confidence before considering for the real experiments (Figure 5.4:B). This feature is profoundly useful when the probes are intended to be used in the fluorescence *in situ* hybridization (FISH) experiments, where the hybridizations are carried out with the nearly native conformations of ribosomes in permeabilized cells. Using the special feature of ARB (Multiprobe Design Tool), the user can design up to five probes that can identify the target group (Ludwig et al. 1998). More often these probe sets are used for multiple FISH experiments. Visualization of the multiple probes simultaneously in the rRNA crystal structure in different background colors provides the researcher a deeper insight into the proposed probe targets and offers him/her a careful and thorough evaluation of the probe candidates *in silico* before making any decision on the selection of probes. Additionally, the user can also evaluate the proposed probe candidates with respect to the

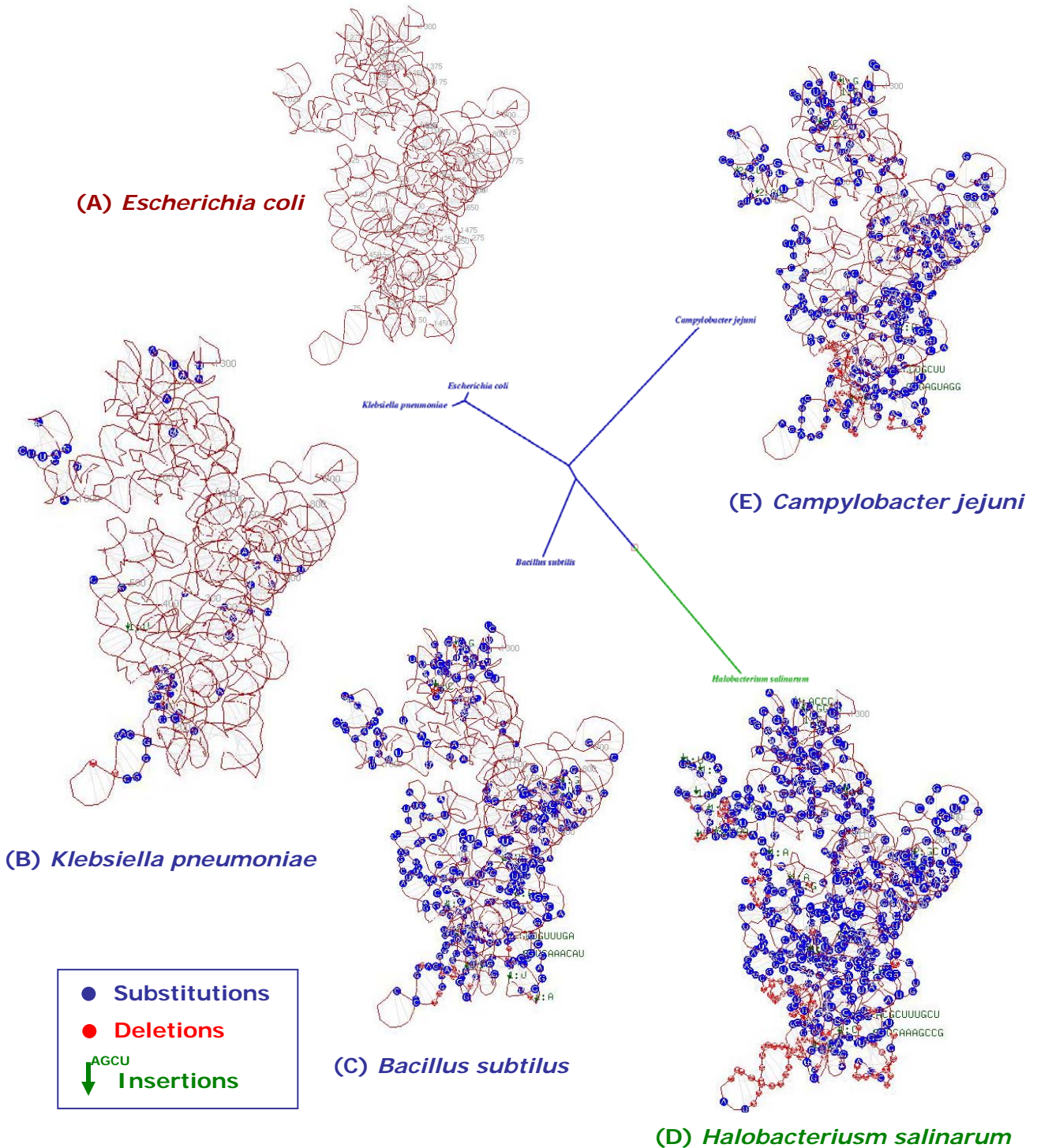


Figure 5.3. Mapping 16S rRNA sequence data. Screenshots generated by RNA3D tool showing mapping rRNA sequence data from different organisms onto the three-dimensional structure of *E.coli* small subunit rRNA. In this composite figure, the 16S rRNA sequence from different taxa are initially aligned in reference to *E.coli* rRNA sequence and then mapped onto the *E.coli* tertiary structure. **Blue** residues represents **substitutions or mutations** where as **red** regions denotes **deletions** in reference to *E.coli* rRNA sequence. **Green** arrows with residues indicates the positions and number of plausible **insertions**. By a glance, it is more evident that an organism (B) that is closely related to the reference structure (A) has fewer differences than the organisms (D and E) that are distantly related.

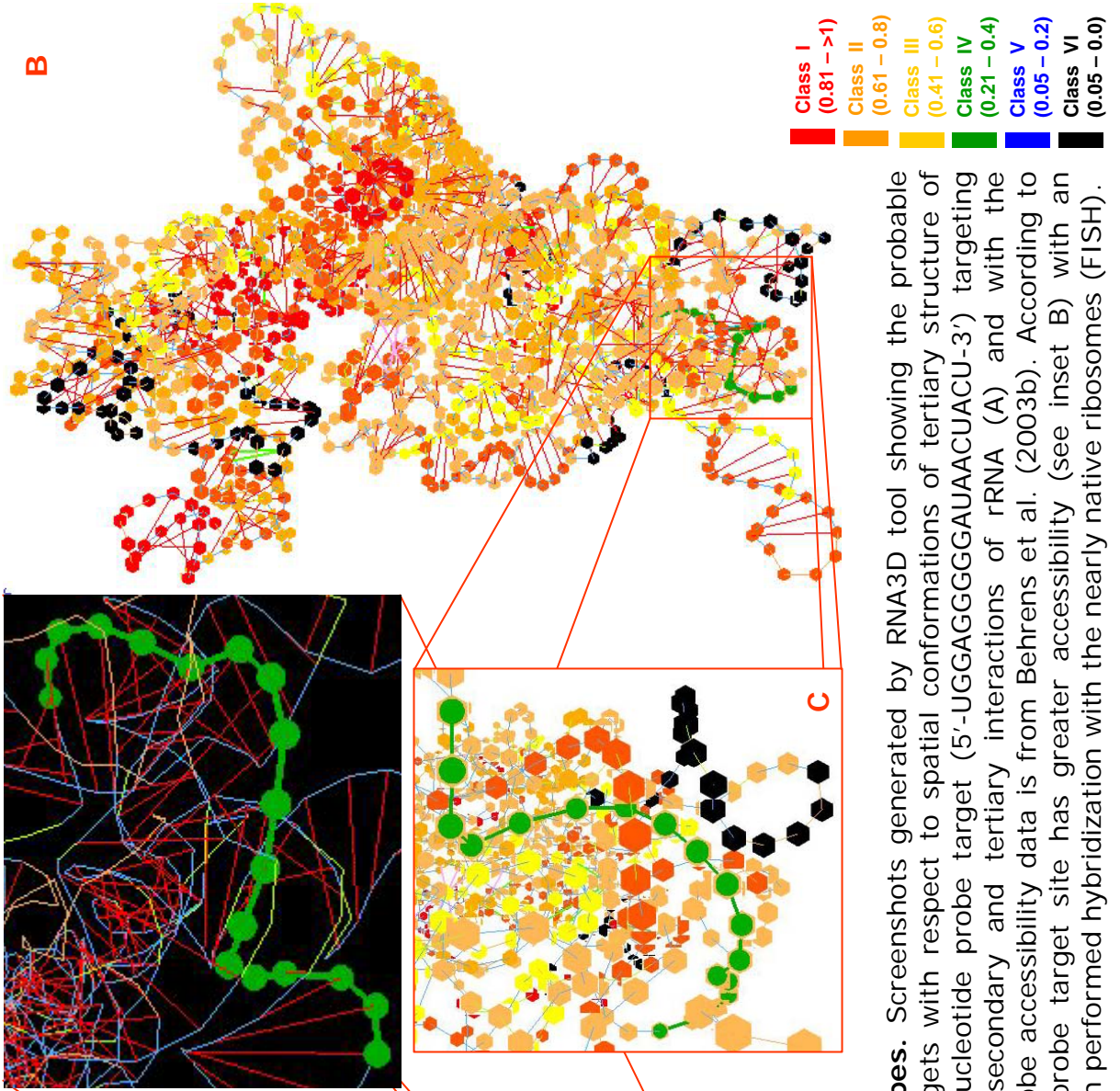


Figure 5.4. Mapping oligonucleotide probes. Screenshots generated by RNA3D tool showing the probable binding site of the oligonucleotide probe targets with respect to spatial conformations of tertiary structure of small subunit rRNA. In this example, oligonucleotide probe target (5'-UGGAGGGGUAUACUCU-3') targeting enterobacterial group is evaluated against secondary and tertiary interactions of rRNA (A) and with the experimental probe accessibility data (B). Probe accessibility data is from Behrens et al. (2003b). According to the probe accessibility data, the proposed probe target site has greater accessibility (see inset B) with an expected fluorescence intensity up to 0.8 when performed hybridization with the nearly native ribosomes (FISH).

consensus models, accessibility maps, etc., (see section 5.9), which adds the confidence to the selected probes. Optionally, sets of search strings that are defined and visualized in the linear sequences of ARB primary structure editor (Figure 6.1; see chapter 6) can as well be mapped onto the 3D structure of rRNA with different background colors (see section 5.4).

5.9. Mapping sequence associated information (SAI) onto the molecule

Various column statistics like sequence consensus, base frequency, positional variability based on parsimony method and any other user defined column statistics, are performed on the sequence alignments using the integrated tools of ARB package. Any information derived from performing such column statistics on the multiple sequence alignments is readily overlaid onto the master structure. Once the column statistics are performed, the user can define the color translation table for the chosen SAI in the ARB primary structure editor (see chapter 6). Different colors (up to 10 colors) can be set to the values or characters stored in the SAI to visualize in the molecular structure. The molecule can be re-colored using new settings anytime by clicking the color palate button (Figure 5.1). Superimposition of information derived from the underlying multiple sequence alignment on to the molecule dynamically is a powerful feature, which a researcher would be interested in, to observe the phylogenetic forces or any other models/masks (eg., consensus model) in a real time virtual environment. It also gives a deeper insight into the fitting of other rRNA sequence onto the template structure (see section 5.7). This feature is important for the researchers who are designing FISH experiments. The importance of visualization of probes in the crystal structure has been mentioned before in the "mapping oligonucleotide probes onto the molecule" section. Target accessibility is among the crucial criteria to be evaluated with respect to the experimental success of the respective probe based identification and detection system (Amann et al. 1995; Fuchs et al. 1998). Accessibility studies on 16S and 23S rRNA of *E. coli* and other organisms with respect to FISH experiments (Fuchs et al. 1998; Fuchs et al. 2001; Inacio et al. 2003; Behrens et al. 2003b) revealed that some regions of *E. coli* ribosome are virtually inaccessible for oligonucleotide probes when FISH is performed. Availability of accessibility data on members of the domains *Bacteria*, *Archaea* and *Eukarya* led to a development of a consensus

model for the accessibility of the small subunit rRNA to oligonucleotide probes (Behrens et al. 2003b). The proposed consensus model and individual accessibility maps can be very well superimposed onto the 3D structure of rRNA. The *in situ* hybridization signals, which are classified into six classes based on the intensities, can be assigned with different colors during mapping onto the molecule (Figure 5.5). The *in silico* evaluation of oligonucleotide probes with respect to the higher-order structure and experimental accessibility data should help to design more successful hybridization experiments (Figure 5.4).

Any statistical and non-statistical data with respect to individual columns in the primary alignments can be readily overlaid onto the tertiary structure of small subunit rRNA and the respective structural maps can be obtained using RNA3D tool. Figures 5.6 and 5.7 displays 'positional variability map' and 'RNA-Protein interaction map' of the three-dimensional structure of small subunit rRNA generated by RNA3D tool, respectively.

5.10. Related work

Several programs exist to visualize three-dimensional structures from PDB files (for e.g. RasMol: <http://www.bernstein-plus-sons.com/software/rasmol/>, Raster3D: <http://www.bmsc.washington.edu/raster3d/raster3d.html>, WebMol: <http://www.cmpharm.ucsf.edu/~walther/webmol.html>, Cn3D: <http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>). Most of them are general applications and mainly targeted to visualize three-dimensional protein structures. Some of the applications are web-based necessitating an internet connection for their use and some need special plug-ins (for e.g. Chime plug-in) for rendering 3D view of molecules. Furthermore, programs that are achieving overlay of information derived from multiple alignments onto the molecule (Stothard 2001; Glaser et al 2003) are limited to static displays and are restricted to protein molecules. In this regard, RNA3D tool, using OpenGL for 3D rendering, overcomes such limitations with its dynamic capabilities and offers a special platform to carry out in-depth structural analysis with respect to ribosomal RNA. With the capabilities like overlaying the phylogenetic information and other sequence associated features onto the molecule dynamically, RNA3D helps in validating multiple alignments of rRNA genes. Additionally, it also allows evaluating oligonucleotide probes with respect to 3D conformations of

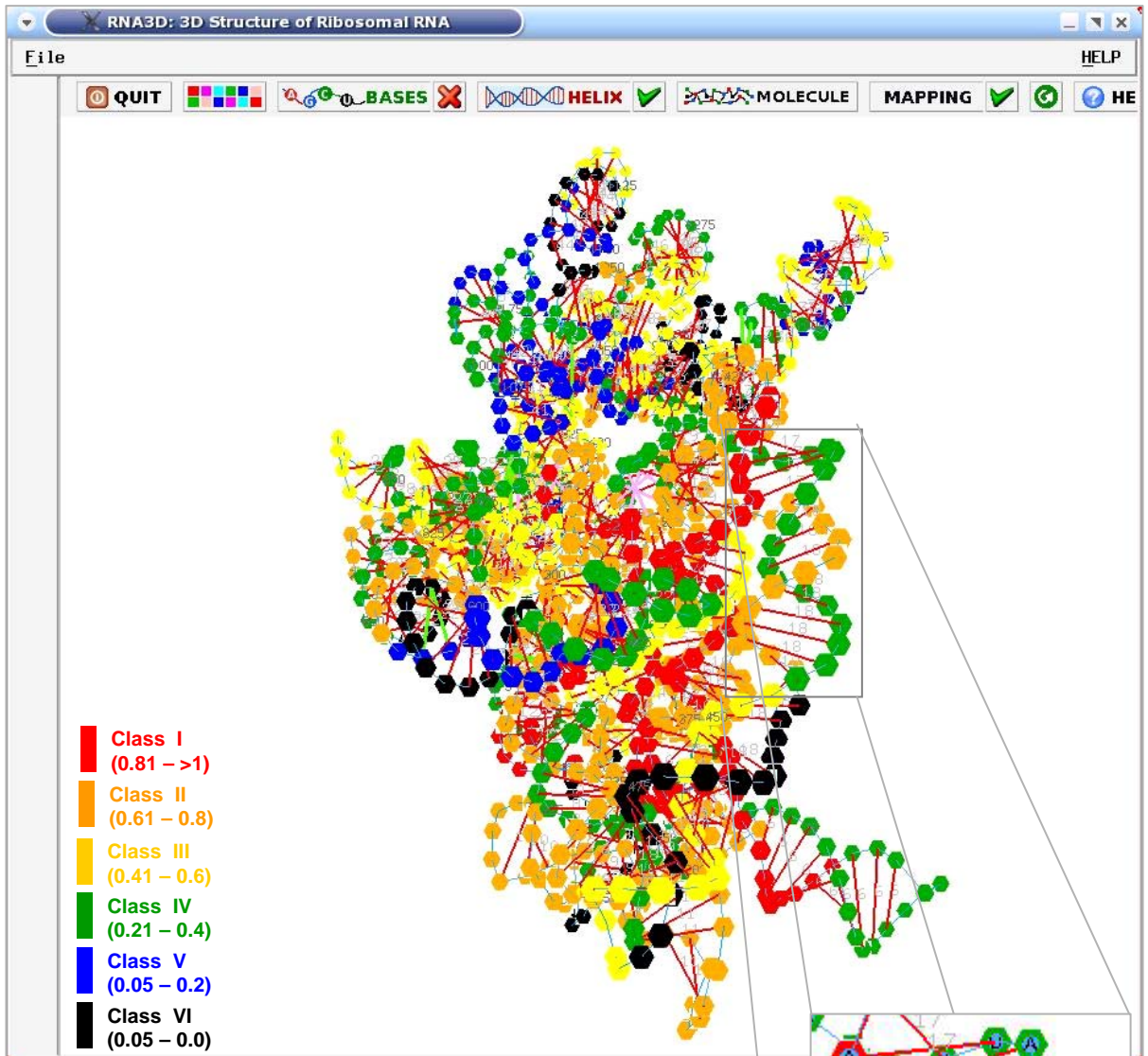
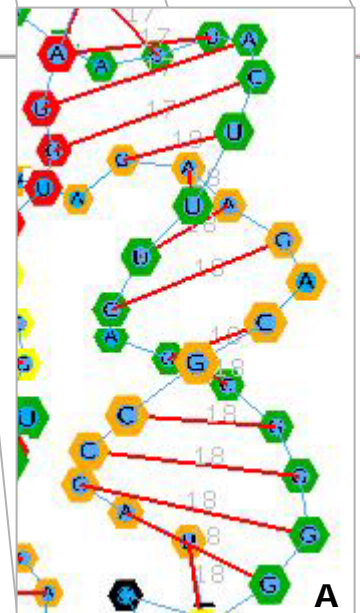


Figure 5.5. Probe Accessibility Map. Screenshot generated by RNA3D tool showing the distribution of relative fluorescence hybridization intensities of oligonucleotide probes targeting 16S rRNA of *E.coli*. Probe accessibility data is taken from the published work (Behrens et al. 2003). Probe accessibility information is mapped onto the three-dimensional structure of *E.coli* 16S rRNA. The different background colors indicate brightness range of different classes (classes I through VI) with respect to the observed fluorescence intensities. Residues colored **red** are readily accessible where as residues colored **black** are virtually inaccessible for the probes. Detail of the structure along with the respective residues in 16S rRNA sequence is shown in the inset (A).



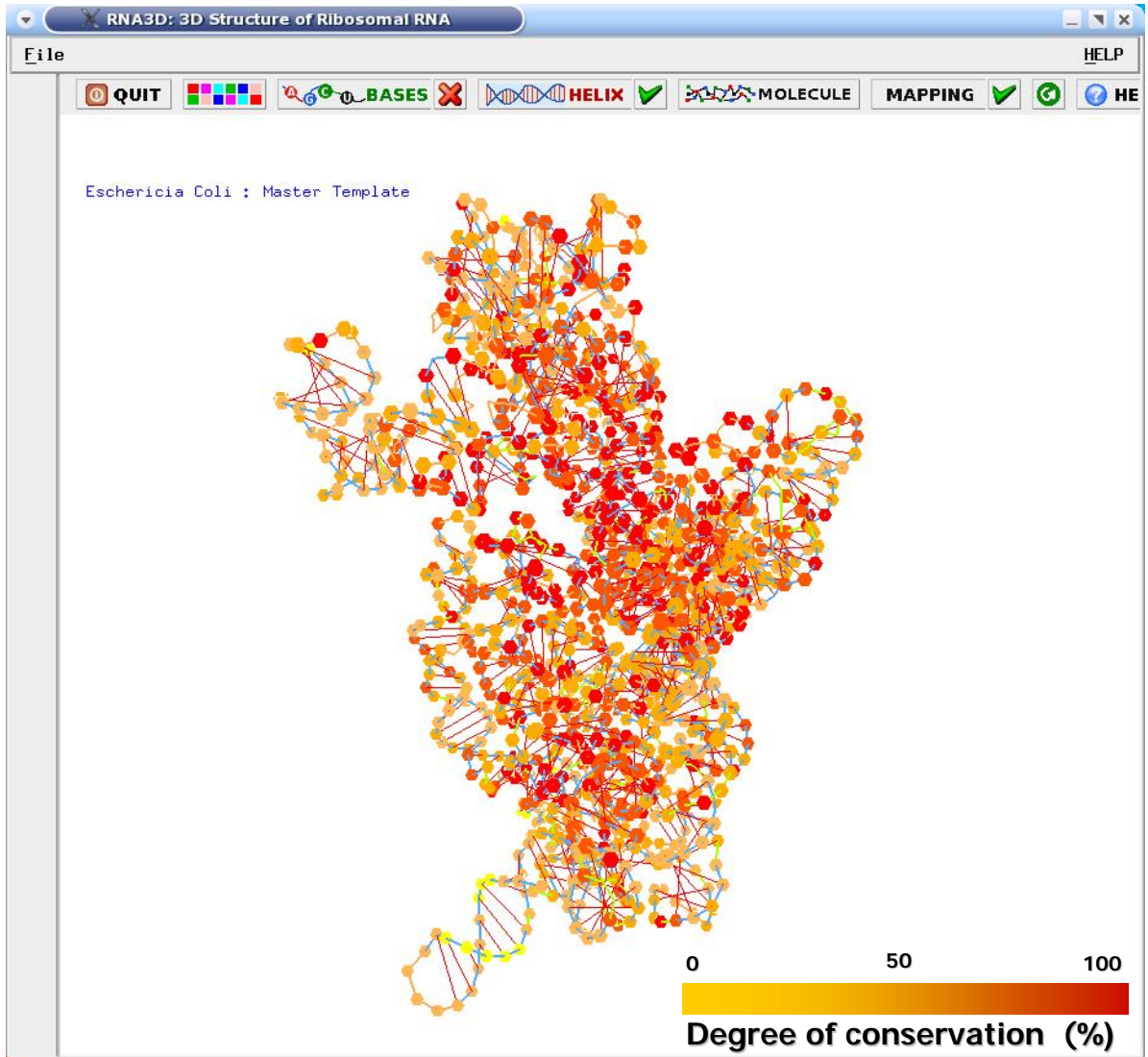


Figure 5.6. Positional Variability Map. Screenshot generated by RNA3D tool showing the positional variability information superimposed onto the three-dimensional structure of small subunit ribosomal RNA. Column statistics are performed on the multiple alignments using parsimony method and minimum number of mutations for each site is determined. The positional variability values are then overlaid onto the tertiary structure of 16S rRNA of *E.coli* residue-by-residue to generate 3D positional variability maps. Residues inclining towards yellow are highly variable whereas the residues inclining towards red are highly conserved positions. Positional variability values are calculated with the data set containing 1000 16S rRNA sequences.

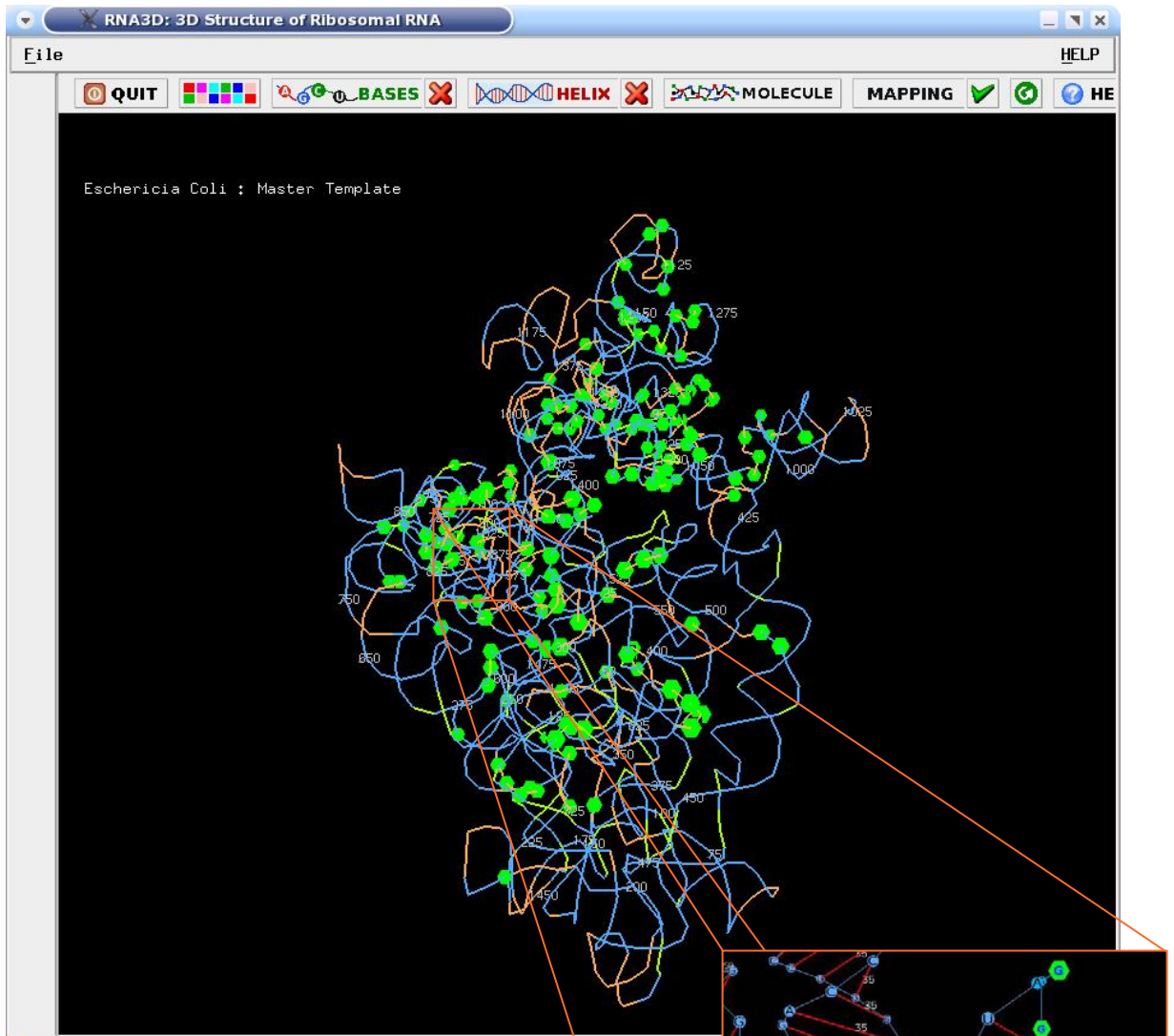


Figure 5.7. RNA-Protein Interactions Map.

Screenshot generated by RNA3D tool showing the RNA-Protein interactions with in the three-dimensional spatial conformations of rRNA. The ribosomal proteins (S2 to S21) contact sites with in the tertiary structure of small subunit rRNA are highlighted in green. The RNA-Protein interaction data is retrieved from the published literature (Stern et al. 1988). Detailed view of the structure with precise protein contact sites (green background) along with the participating residues and secondary structural interactions of rRNA is shown in the inset (A).

The inset (A) shows a detailed view of the structure with precise protein contact sites (green background) along with the participating residues and secondary structural interactions of rRNA. The inset displays a network of nodes and edges, where nodes represent residues and edges represent interactions. The nodes are color-coded by base pair (A, C, G, U) and are connected by red lines, indicating interactions. The background is green, highlighting the protein contact sites.

ribosomal RNA, which is highly valuable in FISH studies. Such unique features of RNA3D tool are seldom found in the existing tools, which are more specialized to visualize the molecules deposited in the protein data bank.

5.11. Discussion

Visualization of the three-dimensional structure of ribosomal RNA in an intuitive display environment within the ARB environment provides the molecular systematists with greater possibilities to carry out structure based phylogenetic analysis. The RNA3D tool is based on the tacit assumption that all the molecules within a family have a common three-dimensional shape which is supported by a common secondary structure that allows key functional groups to adopt similar spatial positions. Therefore, the atomic structure model of the *E. coli* 30S ribosomal subunit (Tung et al. 2002) is taken as a reference structure to evaluate other rRNA sequences and is further substantiated with the availability of very few rRNA crystal structures. Furthermore, the studies conducted by Gutell and coworkers have confirmed the accuracy of the covariation-based secondary structure models of rRNA with the crystal structures of ribosomal subunits (Gutell et al. 2002). Such studies support the inclusion and usage of three-dimensional structures of rRNA for carrying out rRNA based studies. The accuracy of the 16S and 23S rRNA comparative structure models as confirmed by three-dimensional structures not only augments the credibility of the comparative approach, but it also aids in validating the sequence alignments that have been initiated, refined and expanded over the past two decades. By mapping individual rRNA sequence onto the template structure the user can visually inspect the quality of the local alignment and identify the regions that may need any manual checking for further refinement of sequence alignments. Additionally, the entire sequence cannot be viewed at once in primary sequence alignments, so by superimposing the sequence onto the 3D structure the user can get a complete view on the entire sequence in one glance. Observations such as presence of intramolecular interactions (rRNA tertiary interactions) in the loop regions of the structure can be evaluated with respect to three-dimensional conformations of ribosomal RNA in real time. The tertiary interactions are attributed to their role in stabilizing tertiary fold of rRNA (Tung et al. 2002) and hence they are highly conserved. Such information is very useful during evaluation of probe

targets with respect to 3D conformations of rRNA for the intended use in FISH studies (Behrens et al. 2003a).

The RNA3D tool, with its dynamic capabilities and interaction with several tools of the ARB package, allows the changing of display parameters while the molecule is being displayed without compromising with the performance, which is very important to observe any inference drawn with the underlying sequences in the real-time environment. With the additional capabilities like mapping other rRNA sequence data, phylogenetic information, and other sequence associated information, onto the master structure dynamically, provides the researcher with more possibilities to observe the phylogenetic forces (for e.g. positional variability according to parsimony criterion) in a real time virtual reality environment. By mapping oligonucleotide probes and the probe accessibility maps, one can virtually observe the probable hybridization behavior of the prospective probe *in silico*.

The integration of RNA3D tool into the powerful and widely used ARB software package enables the communication at any time with other tools housed in the ARB software package (Ludwig et al. 2004). Thus, giving the researcher an all-in-one software platform to carry out thorough sequence analysis with much deeper perspective, which is seldom found for his/her disposal. In the future, such virtual reality display environments will become more important as tools for bioinformatics, as they provide much higher possibilities to integrate molecular sequence data, structure data and analysis data on one platform.

Chapter 6

SAI viz: A tool to visualize sequence associated information in ARB primary structure editor

6.1. Background

The alignment of primary structures i.e. identification and arrangement of homologous positions in common columns, is a critical step in inferring phylogeny of the sequences. When there is a significant variability between the sequences, alignment of such sequences becomes a daunting task. Additionally, the homologous character of positions in variable regions is not necessarily indicated by the sequence identity or similarity and hence can often not be reliably recognized. Since the number and character of positional differences between aligned sequences are the basis for the inference of relationship, more obviously, the primary alignments are evaluated against several criteria before processing with the treeing algorithms. Criteria such as conservation profiles, positional variability, and maximum base frequency filter are commonly employed for evaluation of sequence alignments.

Generally, computational tools for sequence analysis are often specialized in producing only one kind of feature, and frequently in text output format. More often such sequence associated information (SAI) or features cannot be visualized in the primary structure alignments. It is extremely difficult for a plain text format of sequence in a primary alignment to reveal all the associated information of a sequence to the user in a more comprehensible way. Although the ARB primary structure editor does display the residues in different colors, a more intuitive graphical

display of the features associated to individual or group of sequences was missing. So, SALviz, a visualization tool, was developed to facilitate visualization of such sequence associated features by superimposing on the linear structure of the individual or multiple sequences in the ARB primary structure editor, in a more perspective way.

6.2. Glimpse of the tool

The SALviz tool presents a holistic, graphical view of features associated with nucleotide or protein sequences. The SALviz tool is integrated into the ARB primary structure editor and can be invoked from the "View" menu by selecting "Visualize SAI" sub-menu. The tool consists of a pop-up dialog and offers the options for necessary settings to visualize sequence associated features on primary structure alignments.

6.3. Selection of SAI

SAIs are calculated by using the respective tools of ARB software package and are stored along with the alignments in the ARB database. Additionally, structural information (for e.g., secondary and tertiary structure of small- and large-subunit of rRNA), accessibility data (experimental *in situ* probe accessibility data (Behrens et al. 2003b)) and any other information can be imported and converted to SAI (using ARB import functions). All SAIs are stored in ARB database for future use. Users can select the desired SAI by clicking "Select SAI" button. SALviz initiates the connection with the ARB database and retrieves the SAIs stored in the current database. Retrieved SAIs are presented in the form of scrollable table dialog. Alternatively, the desired SAI can be selected in the primary structure editor (Figure 6.1) by clicking on the respective SAI line. To select a SAI directly in primary structure editor, "Autoselect SAI" toggle in SALviz window has to be enabled. By walking through the different SAIs in the primary structure editor, user can easily navigate and select different SAI to visualize on sequence alignments. This feature offers more flexibility for users to quickly overlay the information on sequence alignments as well as to switch between different SAIs at a mouse click.

6.4. Color translation table

Color translation tables contain color definitions for the respective values or characters found in SAI. By default, ARB database does not contain any color translation tables. Users can define color translation tables for the respective SAIs by using "Create" and "Edit" buttons. Color translation table consists of a set of colors (usually ten colors) where the user has to fill the values or characters (contained in SAI) in the corresponding input fields. The underlying algorithm allocates the specific colors for the corresponding values or characters of SAI. The translation colors are defined in the "Properties" menu of primary structure editor. By default, a range of gray colors is included in the color options to help the user to generate informative black and white background mapping. But the colors can be changed at any time according to the user's preference. And new changes are immediately updated in all the application windows. Normally, one has to define separate color translation table for each of the SAI contained in the underlying database. Users can define multiple color definitions (color translation tables) for the same SAI. Special feature implemented in SAlviz tool enables the synchronization of color definitions with the respective SAI. For example, when the user selects a different SAI the respective color definition is selected automatically. Furthermore, color definitions or color translation tables can be modified, copied and deleted using the edit, copy and delete buttons, respectively.

Generally, color translation tables are not stored in the ARB database. The color definitions are stored locally in the ARB properties database and are loaded along with the main database. This separation from the main database helps the researchers to share their color definitions among the group without actually sharing the entire working database.

6.5. Visualization of SAI

The sequence associated information is visualized in different background colors on the primary structure of individual sequences or multiple sequence sets. SAlviz tool converts the color definitions into series of background colors, which are overlaid onto each sequence base-by-base or residue-by-residue in the primary structure editor. Using the "Auto select SAI" feature of SAlviz tool (see section

6.3), one can quickly switch between different SAIs by selecting the desired SAI using mouse and overlay the respective associated information on sequence alignments, dynamically. In addition, users can change the color settings by navigating "Properties" menu and the display will be updated with new settings immediately. Optionally, one can restrict the visualization of SAIs to the marked species, by setting "Visualize SAI for" toggle button to "Marked Species". Toggling to "All Species" will apply visualization to the entire set of sequences displayed in the primary structure editor. This feature helps the researcher to inspect the particular sequence or sequence sets of his/her interest more clearly, by switching off the visualization for rest of the sequences.

The entire visualization of sequence associated information can be switched on and off any time during the application run by toggling "Enable Visualization" button to on and off, respectively.

Any statistical and non-statistical data that are calculated and retrieved from the literature can be readily overlaid onto the primary alignment. This helps the user to visually inspect and evaluate the primary alignments against several criteria defined by his/her needs. Figures 6.1 and 6.2 demonstrate the possibility to evaluate the primary alignments against certain column statistics performed on the entire data set using SAIviz tool. Figure 6.1 shows the overlay of conservation profile onto the primary alignments of 1000 rRNA sequences. Conservation profiles are calculated using parsimony method and the positional variability for each column is determined. Figure 6.2 shows the rRNA sequence alignments overlaid with the consensus sequence information.

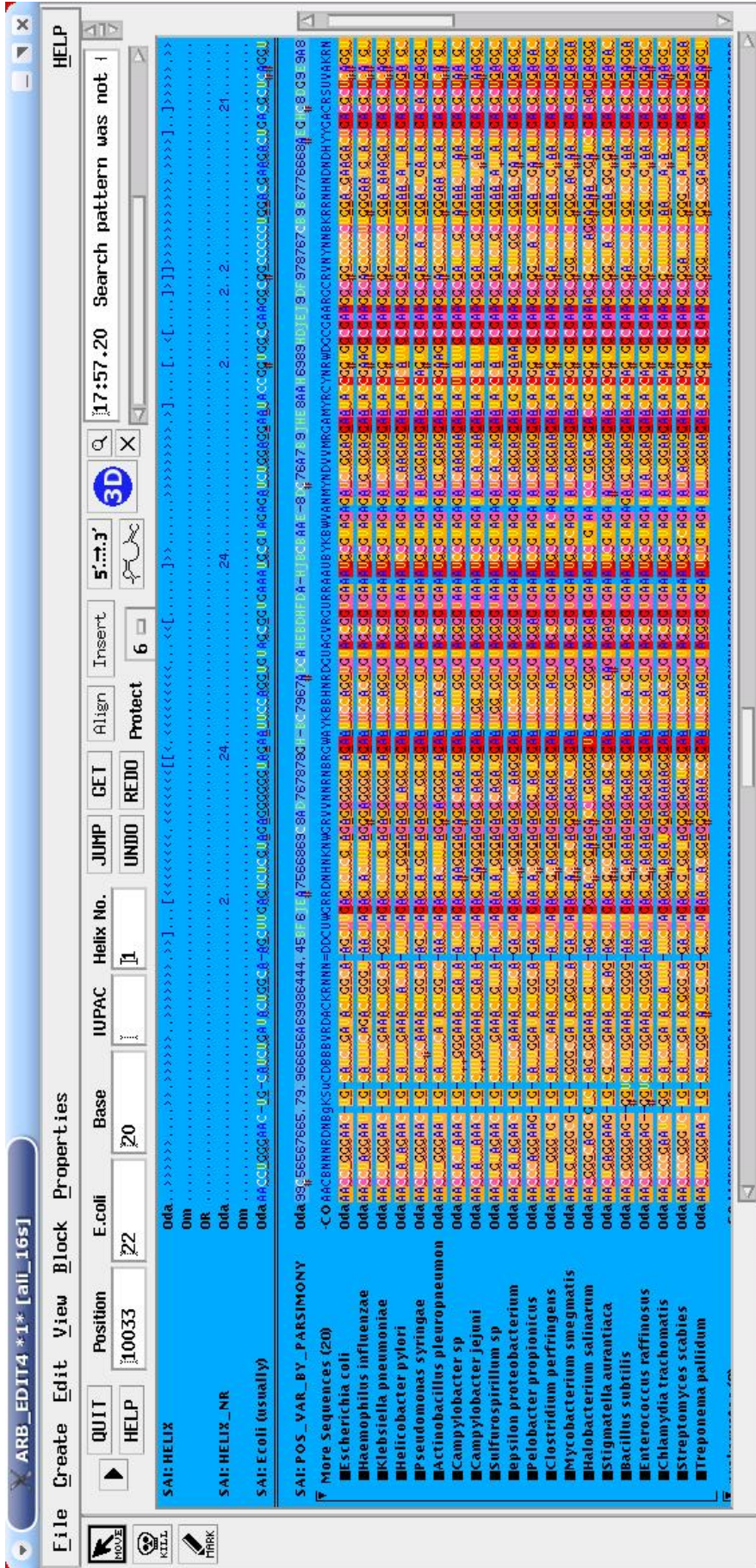


Figure 6.1. Positional Variability Mask. Screenshot generated by SAIviz tool showing the positional variability overlaid onto the primary alignments. Column statistics are performed on multiple alignment of 16S rRNA sequences using parsimony method and the minimum number of mutations for each site is determined. The positional variability values are then overlaid onto the primary alignments by translating the values in to different colors based on the positional conservation. Columns with background inclining towards yellow are highly variable where as the bases with background inclining towards red are highly conserved positions. Positional variability values are calculated with the data set containing 1000 rRNA sequences.

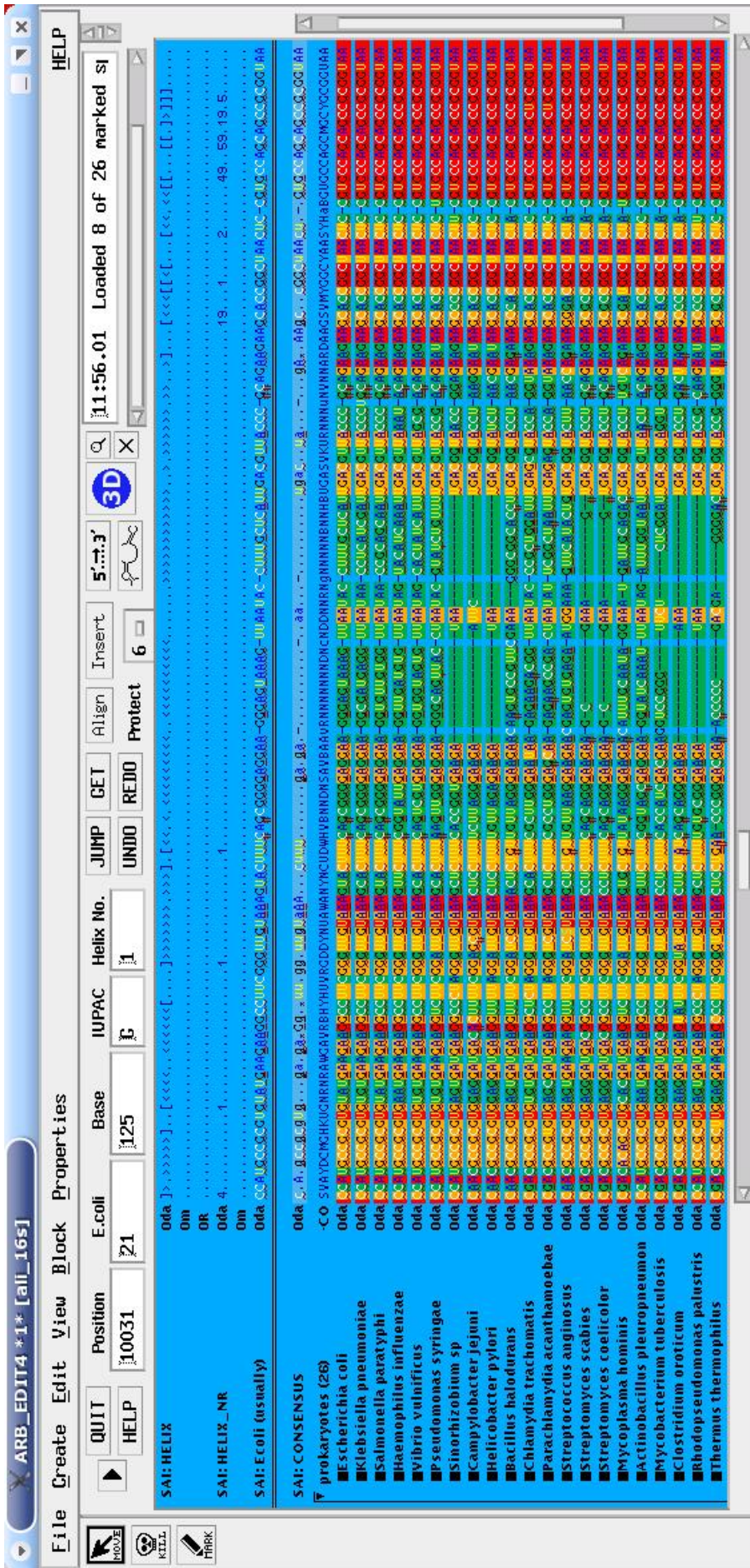





Figure 6.2. Consensus Sequence Mask. Screenshot generated by SAlviz tool showing the consensus sequence overlaid onto the primary alignments. Consensus sequence is determined by performing column statistics on multiple alignment containing 1000 prokaryotic 16S rRNA sequences. The columns with **red** background are present in more than **90%** of the rRNA genes where as the columns with **yellow** background are present in **70-90%** of the rRNA genes. Bases which are present in less than **70%** are shown in **green** background. Columns containing >60% gaps are shown in **blue** background.

-  >90% Consensus
-  >70% Consensus
-  <70% Consensus

Positional Consensus of Small Subunit rRNA

6.6. Related work

Software tools achieving intuitive and more informative displays of local and global alignments have been developed in the recent years. Among them, PipMaker (Schwartz et al. 2000), MultiPipMaker (Schwartz et al. 2003) and VISTA (Mayor et al. 2000) are found to be more useful. But, to use these tools for comparative sequence analysis, user has to provide sequence files along with the files to mask repeats, annotation of genes, and highlighting of other sequence features. More often the additional files have to be generated using different programs. SAIviz tool, developed and integrated into ARB software package, circumvents the problem of need to provide several files by establishing the connection with the underlying database and the integrated tools, to extract the necessary sequence associated features (for e.g. Conservation profiles). With SIAviz tool, users can generate more informative sequence alignments at ease. Furthermore, such informative sequence alignments are highly useful for identifying and resolving alignment ambiguities in multiple alignments which are not possible with the programs mentioned above.

6.7. Discussion

Visualization of SAI in ARB primary structure editor presents a holistic, graphical view of features annotated on nucleotide or protein sequences. The interactive tool highlights the residues in the sequence that corresponds to features chosen by the user, and allows easy searching for sequence motifs or extraction of particular subsequences. The results obtained from diverse sequence analysis tools, which are integrated in ARB software package, can be visualized in a more integrated fashion using the SAIviz tool.

Since the number and character of positional differences between aligned sequences are the basis for the inference of relationship, the primary alignments must be evaluated against several criteria before processing with the treeing algorithms. Criteria such as conservation profiles, positional variability, maximum base frequency filter, higher-order (secondary and tertiary) structure information of ribosomal RNA, accessibility and variability maps, and any sequence specific statistical or non-statistical data (retrieved from literature or designed by experimental observation) should

be taken into account during evaluating primary sequence alignments. The integrated SALviz tool allows insights to be gained by viewing the sequence alignments along with such sequence associated information or features, which may be more revealing than the linear structure of nucleotide or protein sequences.

SALviz tool focuses on the detailed base-by-base or residue-by-residue level of a sequence and its annotations, and visualization in different set of colors. It enables the user to grasp the most salient features of a sequence at a glance, and inspect the corresponding bases or residues precisely. Together with the ARB primary structure editor, SALviz tool establishes a general platform for visualizing the results of sequence analysis programs, as well as for comparing these results to the annotations of the same sequence.

Since the display of sequence associated information is linked directly and dynamically to the sequence itself, the user gets more insights into the individual sequences, thus, allowing a thorough evaluation of the primary sequence alignment data. Furthermore, such display of information along with sequence data can be used for highlighting certain features during presentations and publications (using snapshots), which are more easily understandable. Finally, visualization of SAIs in sequence alignments allows creating a thoroughly descriptive picture of DNA/RNA conservation that is well suited for comparative sequence analysis.

Along with the SAIprobe tool (see section 3.4), this work has been published in BMC Bioinformatics Journal with the title "Graphical representation of ribosomal RNA probe accessibility data using ARB software package" (Kumar et al. 2005).

Chapter 7

SAI probe: A tool to visualize sequence associated information for oligonucleotide probes

7.1. Background

The introduction and use of comparative sequence analysis of appropriate marker genes as a powerful tool in taxonomy have substantially contributed to the rapid growth of molecular sequence databases such as EMBL (Kulikova et al. 2004), GenBank (Benson et al. 2004), and ribosomal RNA (rRNA) databases (Maidak, 2001; De Rijk, 2000; Ludwig, 2004). 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 (Ludwig and Klenk 2001). 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 (Amann et al. 1995), 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 region accessibility in the real hybridization experiment, have to be taken into consideration.

Along with the probe design and probe match tools of ARB software package, SAIprobe provides an intuitive graphical platform for designing, evaluation and visualization of oligonucleotide probes. Using such an interactive graphical user interface, users can gain more insights by visually examining the characteristics and criteria of target regions.

7.2. Glimpse of the tool

SAIprobe tool presents an intuitive graphical view of oligonucleotide probe candidates along with the annotation, local alignment and associated sequence information (SAI). SAIprobe is integrated to ARB probe match tool and can be invoked by clicking "Match SAI" button found in probe match dialog, under "Probes/Match Probes" menu of ARB main window.

7.3. Probe design and match

Oligonucleotide probes are designed using "probe design" and "probe match" tools of ARB (see chapter 3 for details). Along with the probe targets, neighboring region up to nine nucleotides on either terminus of the potential probe target is retrieved from the database. Retrieved probe targets and neighboring region are further processed by SAIprobe tool to generate an intuitive display.

7.4. Selection of SAI

SAIs are calculated by using the respective tools of ARB software package and are stored along with the alignments in the ARB database. Additionally, structural information (for e.g., secondary and tertiary structure of small- and large-subunit of rRNA), accessibility data (experimental *in situ* probe accessibility data (Behrens et al. 2003b)) and any other information can be imported and converted to SAI (using ARB import functions). All SAIs are stored in ARB database for future use. Users can select the desired SAI by selecting "Select SAI" menu under "Properties" menu of SAIviz tool. SAIviz initiates the connection with the ARB database and retrieves the SAIs stored in the current database. Retrieved SAIs are presented in the form of scrollable table dialog to the user for selection.

7.5. Visualization of SAI and probes

The SAIprobe tool retrieves the oligonucleotide probe target sets along with the short stretch of neighboring region from the underlying PT server (see chapter 3). Retrieved sequence data is then, presented in a scrollable tabular form in SAIprobe window. A local alignment is established from the extracted neighboring region (up to nine nucleotides on either terminus of the potential probe target retrieved from the underlying database). Oligonucleotide probe targets and a local alignment of the extracted rRNA sequence is displayed along with the respective annotation (unique identifier such as ARB ID, accession number, or any annotation/feature associated with the sequence data (e.g., Full Name, Group)) (Figures 7.1, 7.2 and 7.3). In any given time, the entire annotation contained in the database of the respective species target, can be browsed in a pop-up dialog (Species info dialog). Selection of target species is done by dragging the mouse and clicking on the desired species. Entire annotation associated with the selected species target is retrieved from the database and is displayed in the "Species Info" dialog. Additionally, user can also perform certain calculation on sequence data, on the fly, using ARB Command Interpreter (ACI) tool and display the result in SAIprobe window. For example, the mol% GC content of the target rRNA species can be displayed. The selected SAI is then, superimposed on the display in different background colors. The characters and values or character groups and ranges of values of the particular SAIs

are assigned different colors, respectively (color translation table – see section 6.4). Colors and fonts for the respective displays can be changed any time by the user using “Properties/Set Colors & Fonts” menu. Optionally, the real characters or values contained in such SAIs can directly be visualized below the individual sequences. This feature offers a deeper insight into the proposed oligonucleotide probe targets for careful examination of probe candidates *in silico* before making any decision on the selection of a probe.

7.6. Example

As an example, oligonucleotide probes were designed for the enterobacteria group represented by 947 database entries. The 5'-UGGAGGGGGGAUAACUACU-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 rRNA sequences. The selected probe perfectly matches the respective target of 497 members of the enterobacteria group. For column statistics, positional variability, the rate of change of a given character state with respect to an underlying tree topology according to parsimony criteria was calculated for each nucleotide column. For non-statistical columns, structural information of 16S rRNA sequence data was used. Additionally, published accessibility maps (Behrens et al. 2003b) are converted to SAI and included in probe visualization and evaluation. Finally, the list of oligonucleotide probe targets along with respective annotation (full name and ARB ID), local alignment and associated information is displayed in different background colors. The same probe has been visualized in all the screenshots presented. In Figure 7.1, the screenshot shows the oligonucleotide probe targets overlaid with 16S rRNA accessibility map (Behrens et al. 2003b), where the experimentally determined relative fluorescence intensities are visualized in different background colors (orange (0.8 – 0.61); green (0.6 – 0.41)). See figures in the section 3.1 for full map of probe accessibility. Figure 7.2 contains a screenshot displaying the superimposition of 16S rRNA secondary structure model on probe targets. 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. Figure 7.3 includes a screenshot presenting an overlay of positional variability information on probe targets 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.

7.7. Related Work

Albeit there have been several software programs developed for the design of rRNA targeted oligonucleotide probes (Ashelford, 2002; Pozhitkov, 2002), the criteria taken to design the probes are generally restricted to the certain parameters such as size, nucleotide composition, specificity definition, and the general hybridisation behavior. None of the software described (Ashelford, 2002; Pozhitkov, 2002) take 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. The SAIprobe tool developed and integrated into ARB software package allows to evaluate the oligonucleotide probes *in silico* with respect to several criteria particularly the probe accessibility data.

7.8. Discussion

As the demand for taxa-specific oligonucleotide probes is permanently increasing, *in silico* evaluation and visualization of such probes and targets are necessary, particularly, when they are 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 (Amann et al. 1995; Behrens et al. 2003b; Behrens et al. 2003a; Fuchs et al. 1998; Fuchs et al. 2001; Inacio et al. 2003; Ludwig et al. 1998).

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

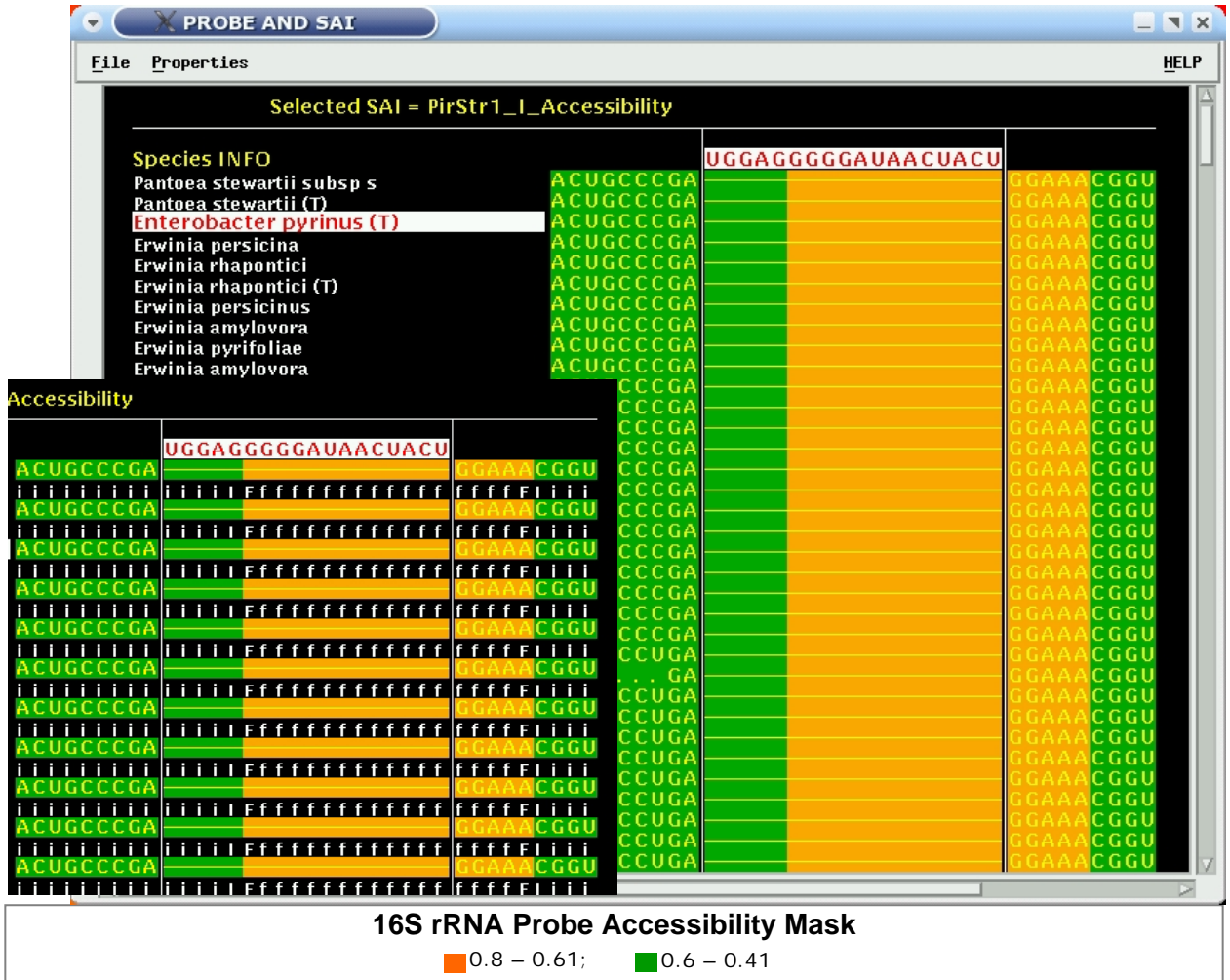


Figure 7.1. Probe Accessibility Mask. Screenshot of SAIprobe tool showing oligonucleotide probe targets overlaid with 16S rRNA accessibility map. Oligonucleotide probe target (5'-UGGAGGGGGGAUAACUACU-3') targeting enterobacteria group is retrieved along with the neighboring region, from the underlying probe server. A local alignment of the extracted rRNA sequence with respect to probe target is displayed along with the full name of the target species. In this screenshot, the oligonucleotide probe target region is overlaid with 16S rRNA accessibility map. Probe accessibility map contains color codes that are assigned to six intensity classes of *in situ* hybridization signals as observed by Behrens et al. (2003b) and are visualized in different background colors (red (0.81 - >1); orange (0.61 – 0.8); yellow (0.41 – 0.6); green (0.21 – 0.4); blue (0.06 – 0.2) and black (0 – 0.05)). Here, the target region falls in between classes IV (shown as green) and II (shown as orange) with the observed fluorescence intensities 0.21-0.4 and 0.61-0.8, respectively. Optionally, each color code is translated into the characters (e, f, i, o, p and x attributed to classes I, II, III, IV, V and VI, respectively) and is visualized below the local alignment of probe target region.

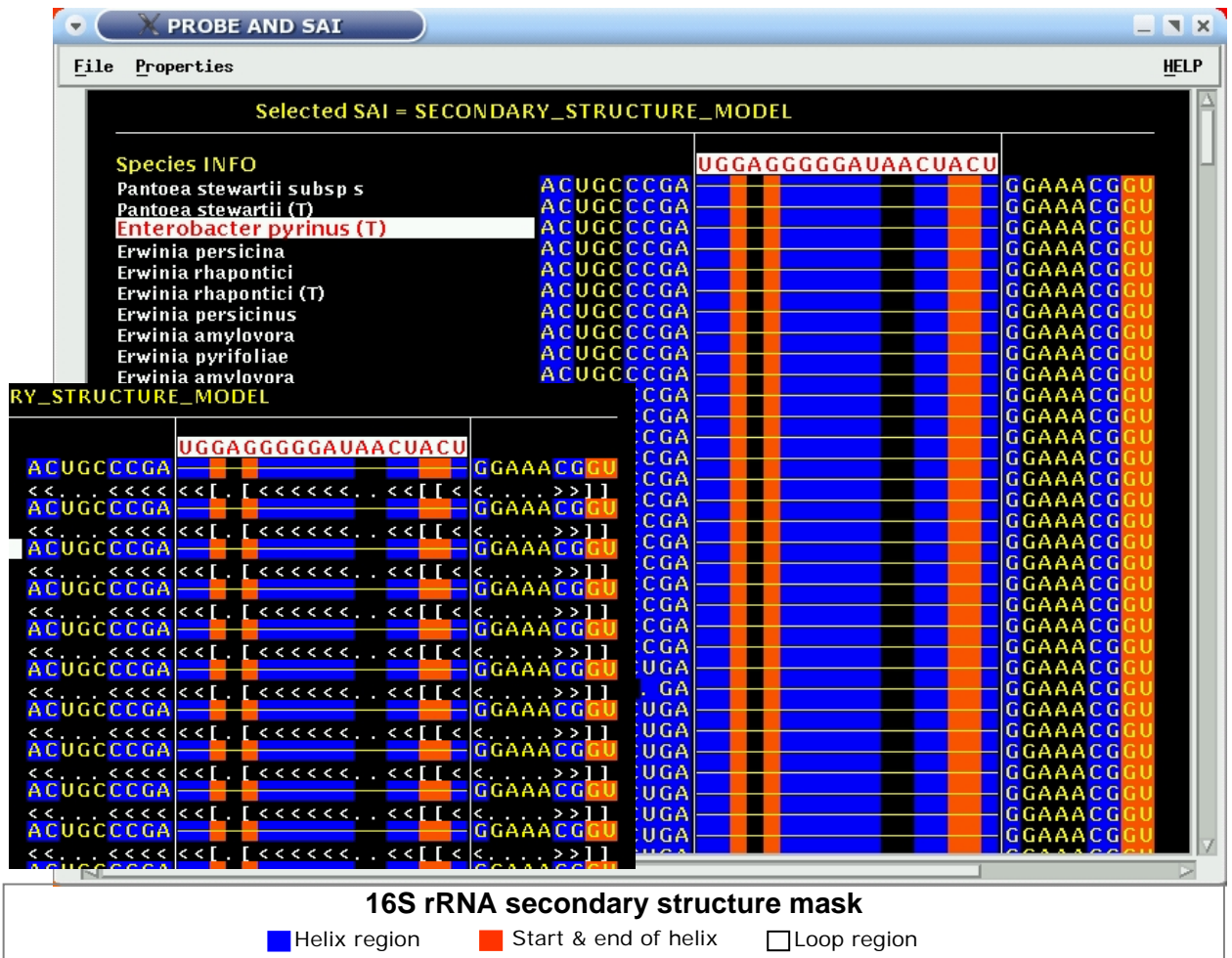


Figure 7.2. Ribosomal RNA Secondary Structure Mask. Screenshot of SAIprobe tool showing oligonucleotide probe targets overlaid with secondary structure information of 16S ribosomal RNA. Oligonucleotide probe target (5'-UGGAGGGGGGAUAACUACU-3') targeting enterobacteria group is retrieved along with the neighboring region, from the underlying probe server. A local alignment of the extracted rRNA sequence with respect to probe target is displayed along with the full name of the target species. In this screenshot, 16S rRNA secondary structure consensus model is superimposed on the oligonucleotide probe target region. Bases in the target region participating in helix formation (in the consensus model) are colored blue, starting and ending positions of helix halves are colored red and the bases without background represent commonly non-base-paired positions. Optionally, symbolic representation of the consensus model is displayed below the local alignment of the target region.

regions of *E. coli* ribosome are virtually inaccessible for oligonucleotide probes when FISH is performed (Fuchs et al. 1998; Fuchs et al. 2000; Fuchs et al. 2001). 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 and optionally visualized as background colors of the sequences in primary structure (see chapter 6), secondary structure (see Chapter 4), and probe visualization windows (Figures 7.1, 7.2 and 7.3) of ARB. Furthermore, any type of information such as secondary structure masks (Figure 7.2) or any statistical calculations performed on the sequence level like sequence consensus, positional variability using parsimony method (Figure 7.3) or any other user defined models, filters or statistics can be used to evaluate the recommended probe targets. Since all the displays produced by the ARB software are interconnected, any changes in one window are automatically updated in other windows as well providing a very dynamic platform for sequence analysis.

Visualization and evaluation of potential probe candidates along with the associated information can be performed at different levels: the local alignment (SAIprobe tool), global alignment (ARB Primary Structure Editor – see chapter 6), secondary structure (SECEDIT tool – see chapter 4) and tertiary structure levels (RNA3D tool – see chapter 5). 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* studies.

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 (Behrens et al. 2003a; Behrens et al. 2003b; Fuchs et al. 1998). 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.

Finally, 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.

This work is published in BMC Bioinformatics Journal under software section with the title "Graphical representation of ribosomal RNA probe accessibility data using ARB software package" (Kumar et al. 2005).

Chapter 8

CONCAT: A tool to concatenate sequence data

8.1. Background

Owing to the rapid advances in DNA sequencing, a considerable amount of sequence data is now available to molecular systematists for inferring the evolutionary history of species. Consequently, multiple gene and genome sequence datasets can now be used to reconstruct more robust evolutionary relationships (Baldauf et al. 2000; Murphy et al. 2001; Wolf et al. 2004; Hedges et al. 2004). Albeit there are many ways of inferring phylogenetic trees from multiple genes for the same set of species (Nei et al. 2001; Suchard et al. 2003), two fundamentally different ways are considered most often. In the first one, phylogenetic reconstruction is done after the gene sequences are concatenated head-to-tail to form a super-gene alignment, which is more often called "the concatenation approach". In the other, phylogenies are inferred separately for each gene and the resulting gene trees are used to generate a consensus phylogeny, which is called "the consensus approach".

The concatenation approach has been used for its presumed statistical advantages which mean a greater phylogenetic accuracy can be conferred by the increased sample size (number of sites) for the given set of taxa. This increased sample size is attributed to improve phylogenetic accuracy in different ways. For example, several researchers found that combining data from two or more different molecular datasets produced a more resolved phylogeny than when a single gene was used (Baldauf et al. 2000; Rokas et al.

2003). On the other hand, the consensus approach summarizes congruence among individual gene trees and produces high resolution in the branching pattern only when there is at least a majority consensus among the different data sets. Thus, it gives a "conservative" or "safe" estimate of the phylogeny (Hillis 1987). Unlike the concatenation approach, the consensus approach takes into account of extensive differences in evolutionary rates and substitution patterns among genes in a gene-specific manner. But the availability of substitution models (Yang 1996) has overcome such limitations of the concatenation approach while most investigations are carried out by simply concatenating sequences and applying a single substitution model to the entire alignment (Murphy et al. 2001; Delsuc et al. 2003; Rokas et al. 2003; Wolf et al. 2004; Hedges et al. 2004).

Nowadays, most of the multigene studies routinely use the concatenation approach to infer phylogenies. Such an approach is more inevitable with the rapid growing of genome sequence databases. So, CONCAT, a concatenation tool, was developed and integrated to ARB software package providing a platform to investigate such approaches with respect to phylogeny reconstruction.

8.2. Glimpse of the tool

The CONCAT tool presents a simple graphical interface for performing concatenation of molecular sequence data with customizable options. It can handle both nucleic acid (RNA/DNA) and protein data and can be invoked from "Sequence/Concatenate" menu of main ARB window.

8.3. Sequence Data

Sequence (nucleic acid and/or protein) data retrieved from public databases such as GenBank (Benson et al. 2004; Kulikova et al. 2004) and from laboratory sequencing projects are imported into ARB using import tools. Additionally, genes extracted from the whole genomes either from public databases or from ARB Genome Package can be used for concatenation. For example, the orthologous genes extracted from the genomes of the desired species can be used as phylogenetic markers for inferring phylogenies based on concatenation approach.

8.4. Concatenation

The CONCAT tool establishes connection with the underlying database and retrieves all the alignments found in the curated database. Retrieved alignments are displayed in the scrollable tabular form in the CONCAT window. Users can select desired sequence alignments to be concatenated from the displayed alignments list. Selection and removal of alignments from the concatenate list are done using respective buttons of the CONCAT window. Additionally, users can rearrange the alignments in the concatenate list using corresponding arrow buttons, to define the order of concatenation. Sequence alignments are concatenated with or without the alignment separator, which can be defined by the user in "Alignment Separator" field. Alignment separator can be used to swiftly identify the concatenation region in the final sequence alignment. Both nucleic acid (RNA/DNA) and protein sequence alignments can be concatenated using the CONCAT tool. Users can define the type of sequence for concatenation using "sequence type" selection box. And also can define name for the final alignment in the "Alignment Name" field. Before performing concatenation operation, the genes or sequence data of the respective species/organism has to be merged. Merging operation is done using "Merge sequence data" feature.

Once the necessary options for merging and concatenation of sequence data has been set, the algorithm of CONCAT tool performs the merging and concatenation operations on sequence data. First, the sequence or gene data associated with common species/organism is merged to the respective species/organism. Then, the sequence alignments (of merged species/organism) in the concatenate list are concatenated in the order of listing with alignment separator, if defined. Final alignment (concatenated) is stored as a separate alignment in the database along with the original alignments.

Finally, the concatenated alignment is further aligned, evaluated and subjected to phylogenetic treeing algorithms using respective tools of ARB. Phylogenetic trees obtained by concatenation approach can be compared with the individual or consensus trees for the final inference and accuracy of phylogenies.

8.5. Related work

Concatenation of multiple genes for multigene phylogenetic studies is usually done either by manual concatenation of genes (Gadagkar et al. 2005) or by designing tailor-made programs (Teeling et al. 2004). But, when large number of datasets is included, the former method becomes more tedious while the latter has to be redesigned extensively. Furthermore, such methods are not easy to use and restricted to command line execution. The CONCAT tool, described here, is an easy-to-use tool and offers the systematists with an interactive platform to perform concatenation of sequence and/or alignments in a more intuitive way.

8.6. Example

As an example, ribosomal operons constituting 5S, 16S and 23S rRNA genes were used to demonstrate the application of the CONCAT tool. 5S, 16S and 23S rRNA genes were extracted from the 118 prokaryote genomes and 15 archaeal genomes. Extracted rRNA genes were then imported to ARB software package and performed alignment using `arb_aligner` (Ludwig et al. 2004). Multiple alignments were checked manually using respective rRNA secondary structures as a reference. Then, the sequence alignments were subjected to treeing algorithms. ARB parsimony tool (Ludwig et al. 2004) was used for phylogenetic reconstruction of rRNA genes and the final tree was generated for individual rRNA genes. In the case of concatenation approach, 5S, 16S and 23S rRNA genes were concatenated from head-to-tail along with the alignment information and subjected to phylogenetic treeing. Finally, individual rRNA gene trees were compared visually with the concatenated (rRNA operon) tree and the inferences were drawn.

Figures 8.1, 8.2 and 8.3 represents the trees reconstructed using 5S, 16S and 23S rRNA genes, respectively. The concatenated rRNA tree is shown in the figure 8.4. From the trees it is more evident that there are clear differences in the resolution and topologies conferred from different rRNA genes. In the case of 5S rRNA gene due to the lesser nucleotide content (120 bp), there is not enough phylogenetic information to deduce meaningful evolutionary relationships among the organisms (Figure 8.1). Tree reconstructed by 16S rRNA genes (Figure 8.2) depict a

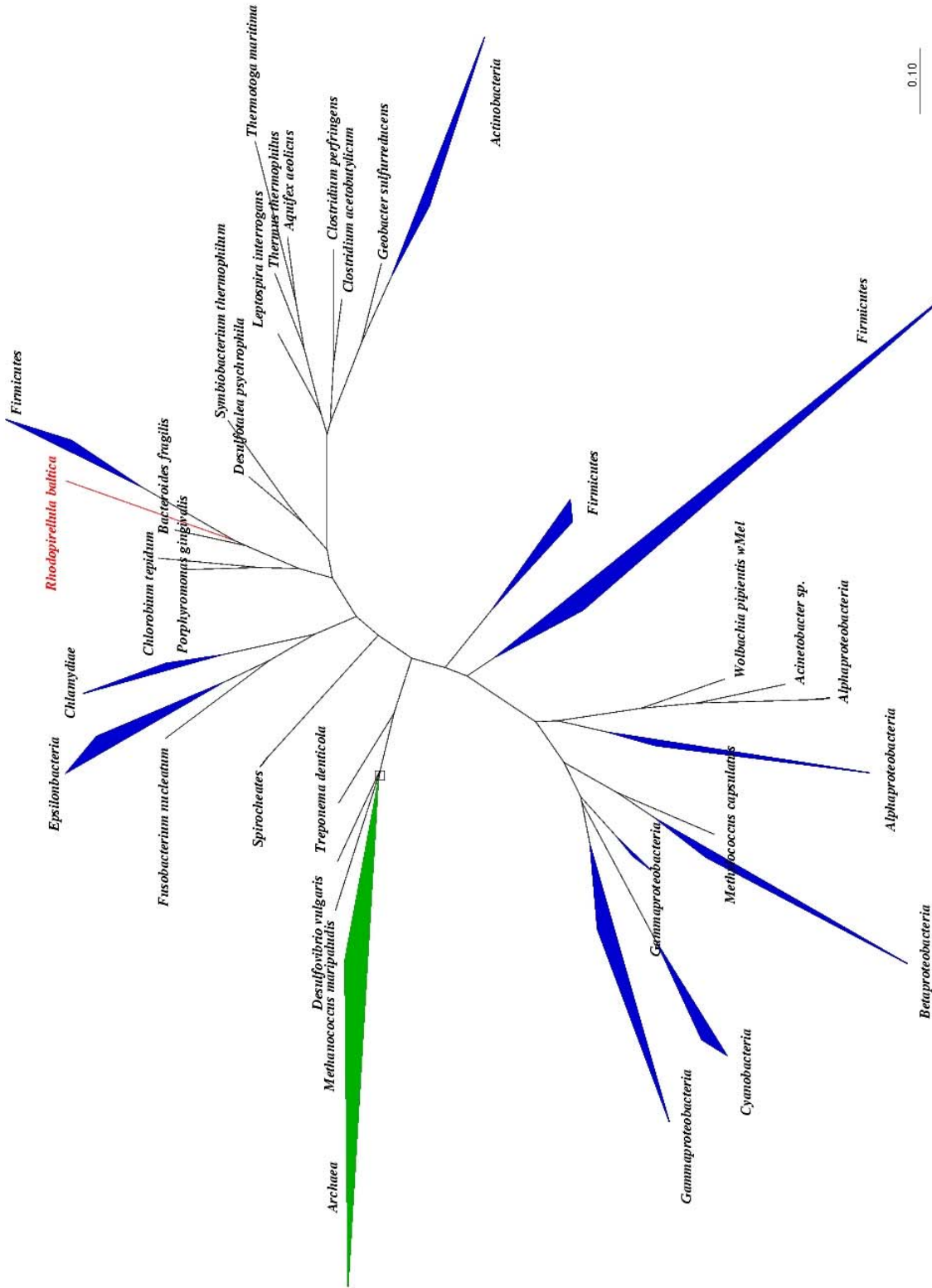


Figure 8.1. 5S rRNA tree. 5S rRNA based tree showing the major bacterial phyla. The triangles indicate groups of related organisms and the edges of triangle represent the shortest and longest branch within the group. The tree was reconstructed and optimized using the ARB parsimony tool. Data set represents 118 bacterial and 15 archaea species. Both bacterial and archaeal genes were extracted from the complete genomes. Archaea was used as outgroup references to root the tree.

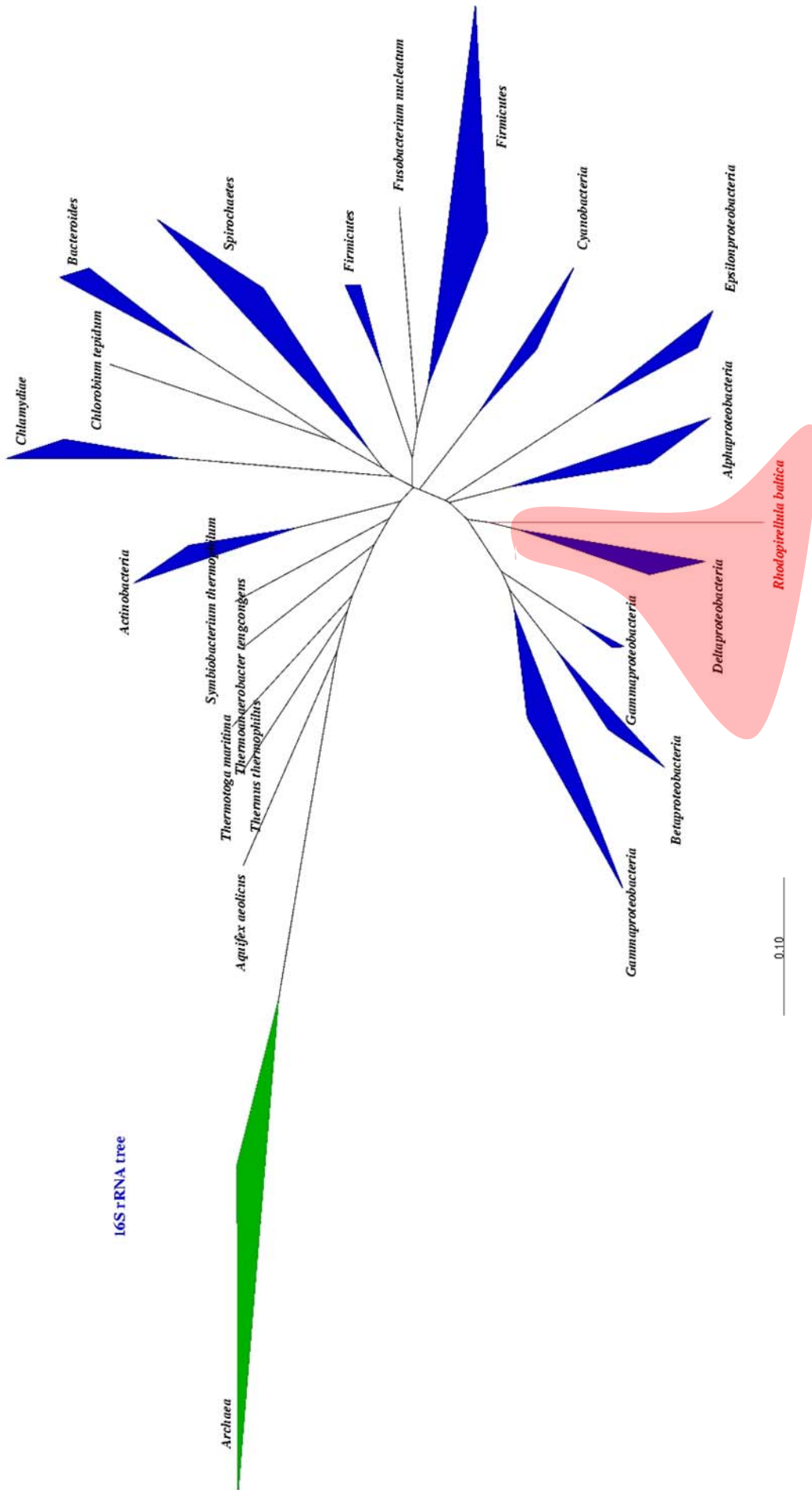


Figure 8.2. 16S rRNA tree. 16S rRNA based tree showing the major bacterial phyla. The triangles indicate groups of related organisms and the edges of triangle represent the shortest and longest branch within the group. The tree was reconstructed and optimized using the ARB parsimony tool. Data set consisted of 118 bacterial 16S rRNA genes and 15 archaeal 16S rRNA genes. Both bacterial and archaeal genes were extracted from the complete genomes. Archaea was used as outgroup references to root the tree.

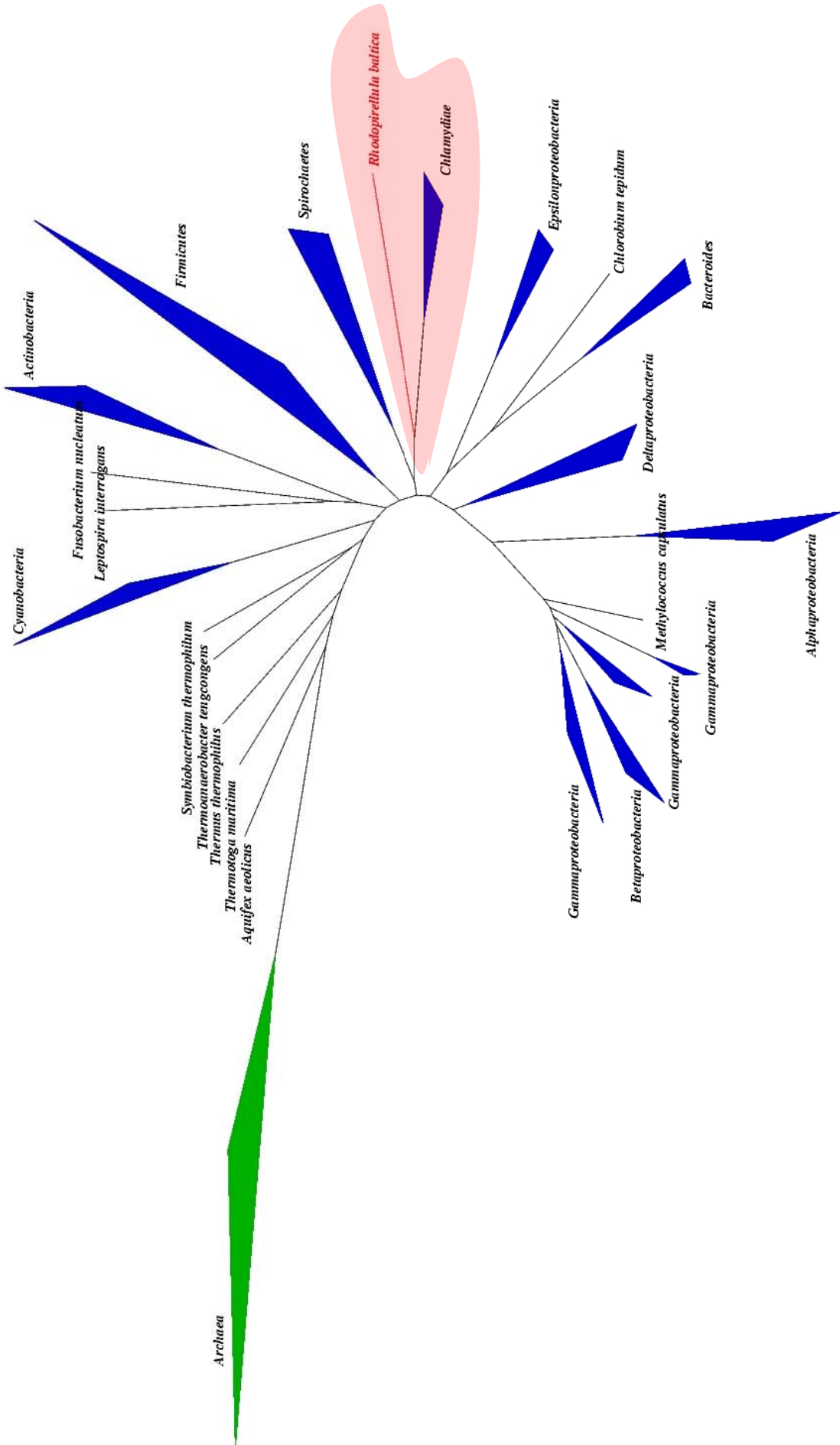


Figure 8.3. 23S rRNA tree. 23S rRNA based tree showing the major bacterial phyla. The triangles indicate groups of related organisms and the edges of triangle represent the shortest and longest branch within the group. The tree was reconstructed and optimized using the ARB parsimony tool. Data set consisted of 118 bacterial 23S rRNA genes and 15 archaeal 23S rRNA genes. Both bacterial and archaeal genes were extracted from the complete genomes. Archaea was used as outgroup references to root the tree.

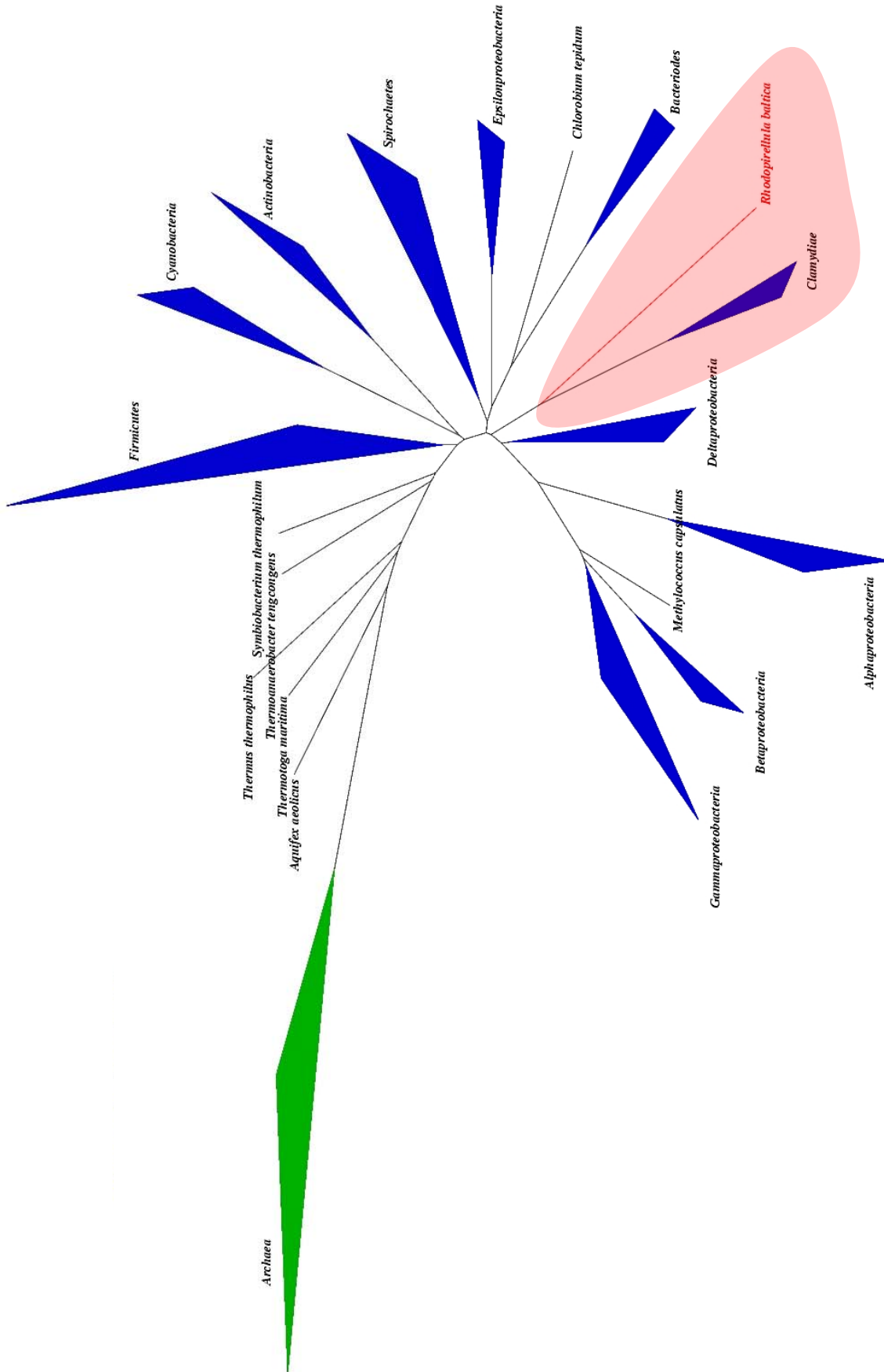


Figure 8.4. rRNA operon tree. Concatenated rRNA based tree showing the major bacterial phyla. The triangles indicate groups of related organisms and the edges of triangle represent the shortest and longest branch within the group. The tree was reconstructed and optimized using the ARB parsimony tool. Genes constituting rRNA operon (5S, 16S and 23S) were concatenated and phylogenetic reconstruction was carried out. Data set represented 118 bacterial and 15 archaeal species. Both bacterial and archaeal rRNA operon genes were extracted from the complete genomes. Archaea was used as outgroup references to root the tree.

considerable degree of resolution in deducing the evolutionary relationships of organisms. Albeit the groupings of major bacterial phyla are more or less in agreement with the accepted phylogeny, there are some incongruencies exists with respect to certain topologies in the tree. For example, Planctomycete, *Rhodopirellula baltica* is placed along with the *Deltaproteobacteria* clade in 16S rRNA tree where as it is placed near to *Chlamydiae* in 23S rRNA-based tree and concatenated rRNA tree. This inconsistency can be attributed to the branch attractions while the trees were reconstructed with limited number of rRNA sequences (118). Furthermore, such inconsistencies are plausible because of the fact that one cannot expect identical tree topologies from different phylogenetic markers (Ludwig and Klenk, 2001) and also there were no additional filters included in the treeing process, as the purpose was to demonstrate the functionality of CONCAT tool and not to produce the best tree.

Higher resolution of phylogenetic trees can be obtained with 23S rRNA because of the fact that large subunit rRNA (23S) carries more phylogenetic information than the small subunit rRNA (16S). There are more than twice as many informative residues in the large subunit rRNA owing to higher nucleotide content (around 2900 bp). Such increased information can be used to achieve greater resolution of different bacterial clades in the phylogenetic trees. For example, alpha-, beta-, gamma-proteobacteria are clearly resolved with the rest of the groups in the tree (Figure 8.3 and 8.4).

With the concatenated approach, further greater resolution of phylogenetic trees can be achieved. This can be explained for the fact that by combining multiple gene data the available phylogenetic information content is greatly increased which results in greater resolution and accuracy of the resulting trees. Any incongruency present in the single-gene trees might be well resolved with the multigene trees as the combined data increases the phylogenetic signal and disperse the noise. For example, association of *Rhodopirellula baltica* with *Chlamydiae* in 23S rRNA tree (Figure 8.3) is maintained in the same clade in the concatenated tree (Figure 8.4) although it was placed wrongly in 5S and 16S rRNA trees (Figure 8.1 and 8.2, respectively). And the affiliation of *Rhodopirellula baltica* with *Chlamydiae* is also supported by the various workers using the concatenation approach (Teeling et al. 2004).

One might expect more significant differences between the tree topologies and the final resolution in single-gene trees and multigene trees (concatenation approach) but such an observations cannot be drawn with the limited data set and the taxon sampling included in the example. Selection of dataset was solely done to demonstrate the application of CONCAT tool and the evaluation of such an approach to find best trees is beyond the scope of the study.

8.6. Discussion

Small sub-unit of ribosomal RNA (SSU rRNA) sequences constitute the single most comprehensive database available for phylum-level systematics (Van De Peer, 1999; Wuyts, 2002; Ludwig, 2004). But these data alone is not sufficient to resolve the phylogenies, completely. For example, higher order systematics of eukaryotes, relying on the analysis of SSU rRNA, depict the eukaryotes as a series of deeply diverging lineages branching successively towards a dense unresolved cluster, the so-called eukaryote crown (Knoll 1992). Alternatively, the availability of genome sequence datasets due to extensive genome sequencing projects, have tempted the systematists to use different genes to resolve such issues and reconstruct more robust evolutionary relationships. Such investigations have led to different gene trees and phylogenies. And one of the most notable difficulties is the widespread occurrence of incongruence between alternative phylogenies generated from single-gene data sets (Rokas et al. 2003). Such incongruencies have been attributed to several factors such as limited data availability, choice of optimality criterion, taxon sampling, and evolution model. In order to overcome an extensive incongruence revealed by single-gene phylogenies, many researchers have attempted to increase the signal-to-noise ratio by analyzing concatenated data sets to address the difficult phylogenetic questions (Baldauf et al. 2000; Murphy et al. 2001; Wolf et al. 2004; Hedges et al. 2004). Combining data increases phylogenetic accuracy both by increasing signal and decreasing noise, which facilitate in resolving conflicting phylogenies. The integrated CONCAT tool can be used for such investigations with in the ARB software environment.

Factors such as sequence length, number of genes, optimality criterion and rate of evolution may influence phylogenetic reconstruction (Hillis 1987). The effect of these factors can be systematically explored with large data sets

(concatenated/combined) derived from biological sequences. And the progress in genome sequencing has presented systematists with an unprecedented opportunity to evaluate these issues in a broader perspective. A simple concatenation of multigene data sets appears to be better than the consensus for phylogenetic reconstruction by resolving conflicting branches of the tree (Rokas et al. 2003).

Phylogenetic studies on multigene/combined data suggests that each gene has its own unique set of strengths and weaknesses as a phylogenetic marker, and it is unlikely that any one gene will ever be able to strongly, or perhaps even accurately, resolve all deep branches of a universal tree (Baldauf et al. 2000). Thus, the concatenation approach of inferring phylogeny is the better way to resolve such incongruencies and produce more accurate and reliable phylogenies. Such an approach may provide unprecedented power not only in testing specific phylogenetic hypothesis but also precise reconstruction of the historical associations of all the taxa analyzed. Furthermore, through analysis of a larger amount of sequence data, greater confidence in the proposed phylogenetic reconstruction can be achieved. Together with the integrated phylogeny tools, CONCAT tool extends the ARB software environment and presents the systematists with a comprehensive platform for multigene phylogenetic analysis.

Chapter 9

Summary

In biology, where data sets are becoming larger and more complex with the structural data, graphical analysis is felt to be ever more pertinent. Although some patterns and trends in data sets may only be determined by sophisticated computational analysis, viewing data by eye can provide the users with an extraordinary amount of information in an instant. Software tools for comparative sequence analysis to visualize the data sets along with the structural data can provide the researchers with a powerful view of the differences and similarities between the gene sequences. This thesis describes the development of various graphical tools aiming to achieve the structure based comparative sequence analysis particularly of ribosomal RNA in a more comprehensible way. SAlviz, SECEDIT and RNA3D are the tools dealing with the secondary and tertiary structures while SAlprobe aids in evaluation of oligonucleotide probes and CONCAT tool facilitates multigene studies.

Ribosomal RNA, a key molecule in protein synthesis with its ubiquitous presence and highly conserved nature has formed the backbone of modern molecular taxonomy. Much of the phylogenetic information is attributed for the fact that rRNA forms distinct secondary and tertiary structures and is highly conserved across all the taxa that have been examined so far. With the advent of secondary structure models, the higher-order structure information of rRNA is routinely used in comparative sequence analysis of rRNA genes. The SECEDIT and RNA3D tools that are developed in the study provide a powerful graphical platform for systematists in order to carry out systematic structure based analysis of rRNA genes. The knowledge of secondary structure is very helpful in order to preserve the alignment

of well-conserved motifs which largely determines the underlying evolutionary relationship among the rRNA genes. When the secondary structure features of rRNA are considered, the alignment ambiguity can be greatly reduced producing more meaningful alignments of rRNA genes. Consideration of additional structural information with respect to rRNA such as interactions with ribosomal proteins, RNA molecules and antibiotic resistance sites are very useful for determining proper weighting during phylogenetic reconstruction of rRNA genes. Such information is also important for the design and evaluation of oligonucleotide probes intended to be used for *in situ* hybridization experiments.

Availability of high-resolution rRNA crystal structures have led to inclusion of the three-dimensional structures of rRNA in structure based phylogeny and molecular probe design studies. The RNA3D tool offers the dynamic 3D environment to visualize and evaluate the tertiary structure of rRNA in a more perspective way. Along with the rRNA comparative structure models (2D), the visualization of three-dimensional structures helps in validating the sequence alignments that have been initiated, refined and expanded over the past two decades. Observations such as presence of intramolecular interactions, antibiotic resistance regions and the contact sites of ribosomal proteins and tRNAs, can be evaluated with respect to 3D conformations of rRNA. With the additional capabilities like mapping other rRNA sequence data, oligonucleotide probes, phylogenetic information, probe accessibility maps, and other sequence associated information, onto the master structure provides the researcher with more possibilities to observe the phylogenetic forces and performance of perspective probe candidates in a real time virtual reality environment. In future, such virtual reality display environments will become more important as tools for bioinformatics, as they provide much higher possibilities to integrate molecular sequence data, structure data and sequence associated data on one platform to gain more knowledge.

Furthermore, as the demand for oligonucleotide probes that can identify and quantify bacteria is permanently increasing, *in silico* evaluation and visualization of such probe targets are necessary, particularly when used for FISH experiments. Target accessibility is among the crucial criteria to be considered for probe based identification and detection system because hybridization is carried out with nearly native conformations of ribosomes. The SAIprobe tool

developed during the study aims to provide a thorough *in silico* evaluation and visualization of such probe targets. In this regard, SAIprobe, SECEDIT and RNA3D are valuable tools for research facilities where *in situ* hybridization experiments are routinely applied.

Since the alignment of primary structures is a critical step in inferring phylogeny, any significant variability between the sequences results in alignment ambiguity. Such ambiguities are very difficult to identify and evaluate with the conventional text output format of multiple alignments. The SAIviz tool offers the possibility to visually inspect the primary alignments against various criteria before subjecting to treeing algorithms. More informative multiple alignments can be generated using SAIviz tool by overlaying any sequence associated features onto the multiple alignments using color coding schemes. This enables the user to grasp the most salient features of a sequence and the alignment at a glance, and inspect the corresponding regions precisely. Such informative displays are easily understandable and are more ideal for presentation and publication purposes.

In future, molecular systematists will include more than one gene to reconstruct even more robust evolutionary relationships. Most of the multigene studies use the concatenation approach to infer phylogenies. The concatenation approach offers a greater phylogenetic accuracy resolving the incongruencies present in single-gene phylogenies due to the increased sample size. Together with the integrated phylogeny tools, CONCAT tool extends the ARB software environment and provides the systematists with a possibility to carry out multigene phylogenetic analysis. Such an approach may not only help in testing specific phylogenetic hypotheses but also help in the precise reconstruction of the historical associations of all the taxa analyzed with greater confidence.

The integration of the tools developed in the study into the widely used ARB software package readily achieves the interoperation between several applications within the ARB package, offering the researcher an all-in-one powerful software platform to carry out thorough sequence analysis. Finally, the software tools achieving visualization of molecular data in an interactive and intuitive graphical user interface ideally may serve as 'third eye' for a molecular biologist.

Zusammenfassung

In der Biologie, in der die Datensätze durch die Einbeziehung räumlich-struktureller Informationen stetig größer und komplizierter werden, wurde eine Möglichkeit zu graphischen Analyse immer notwendiger. Obgleich einige Muster und Tendenzen in den Daten auch weiterhin nur durch anspruchsvolle Berechnungsmethoden festgestellt werden können, kann der Benutzer oft durch Betrachten mit dem bloßen Auge schon eine außerordentliche Menge von Aufschlüssen erhalten. Softwarewerkzeuge zur vergleichenden Sequenzanalyse, die die Sequenzen zusammen mit den Strukturinformationen darstellen, verhelfen den Forschern zu tieferen Einsichten in die Unterschiede und Ähnlichkeiten zwischen Gensequenzen. Die vorliegende Arbeit beschreibt die Entwicklung verschiedenener graphischer Werkzeuge zur strukturunterstützten, vergleichenden Sequenzanalyse - insbesondere von ribosomaler RNS. Dabei ermöglichen die Werkzeuge SAIviz, SECEDIT und RNA3D die Integration der räumlich-strukturellen Informationen in die Primär- Sekundär- und Tertiärstrukturmodelle während SAIprobe diese Informationen zur Sondenevaluierung ausnutzt. Als weiteres Werkzeug wurde CONCAT zur Erweiterung der Alignments auf mehrere Gene entwickelt und integriert.

Die Verwendung der ribosomalen RNS, einem Schlüssel-molekül in der Proteinsynthese, bildet heute das Rückgrat der modernen molekularen Taxonomie. Phylogenetische Information beruht auf der Tatsache, daß die rRNS klare Sekundär- und Tertiärstrukturen ausbilden, welche aufgrund ihres Beitrags zur Gesamtfunktionalität des Ribosoms stark konserviert sind und bisher in allen daraufhin überprüften Organismen gefunden wurden. Mit dem Aufkommen der Kovariationsstrukturmodelle werden nun rRNS-Strukturinformationen höherer Ordnung ebenfalls in der vergleichenden Sequenzanalyse von rRNS-Genen herangezogen. Die Werkzeuge SECEDIT und RNA3D, die in dieser Arbeit entwickelt wurden, geben Systematikern leistungsfähige graphische Hilfsmittel in die Hand, um systematische, strukturunterstützte Analysen der rRNS-Gene durchzuführen. Das Kenntnis von Sekundärstrukturen hilft das Alignment von konservierten Motiven, die zum

größten Teil die zugrunde liegenden evolutionären Beziehungen der rRNS-Gene bestimmen, zu erstellen. Dabei verringert die Einbeziehung der Sekundärstruktureigenschaften der rRNS deutlich die Anzahl an zweideutigen Alignmentpositionen und ergibt damit aussagekräftigere Alignments der rRNS-Gene. Die Nutzung zusätzlicher strukturbezogener Informationen, wie der Interaktion mit ribosomalen Proteinen, anderen RNS-Molekülen oder für Antibiotikaresistenzen wichtige Stellen, ist für die richtige Gewichtung der Positionen während der Stammbaumrekonstruktion, sowie für die Entwicklung und die Evaluation von Oligonukleotidsonden, die zur *in situ* Hybridisierung benutzt werden sollen, sehr nützlich.

Mit den Ergebnissen von hochauflösende Röntgenstrukturanalysen von rRNS-Kristallen wurden auch dreidimensionale Strukturen der rRNS in die strukturbasierte Phylogenie und das Sondendesign miteinbezogen. Das RNA3D-Werkzeug, welches in dieser Arbeit entwickelt wurde, stellt eine dreidimensionale Ansicht der Tertiärstruktur zur Verfügung und hilft die Sequenzalignments zu überprüfen. Entdeckungen von molekularen Interaktionen, Antibiotikaresistenz-assoziierte Stellen sowie Kontaktstellen mit ribosomalen Proteinen und tRNSs können im Hinblick auf die tatsächliche Konformation der rRNS untersucht werden. Zusätzliche Funktionen erlauben es, Oligonukleotidsonden, phylogenetische Informationen, die Zugänglichkeit für Sonden und weitere sequenzassoziierte Informationen (Mutationen, etc.) in der räumlichen Struktur einzublenden und geben dem Forscher bessere Möglichkeiten phylogenetische Gesetzmäßigkeiten und die Eignung von Sondenkandidaten zu erkennen. In Zukunft werden solche virtual-reality-Werkzeuge noch mehr Bedeutung erlangen, da sie noch bessere Möglichkeiten zur Integration von Sequenzdaten, Strukturen und weiteren Informationen in einer Darstellung bieten.

Da die Nachfrage nach Oligonukleotidsonden zur Identifizierung und Quantifizierung von Bakterien ständig wächst, ist eine Visualisierung und *in silico* Evaluierung der Sondenzielstellen sehr nützlich, insbesondere, wenn die Sonden zu FISH Experimenten benutzt werden sollen, da die Hybridisierung mit Ribosomen in einer nahezu nativen Konformation ausgeführt wird. Die Zugänglichkeit der Bindungsstelle gehört zu den entscheidenden Kriterien, die sichergestellt werden müssen. Das SAIprobe-Werkzeug, welches in dieser Arbeit entwickelt wurde, bietet dazu eine vollständige *in-silico* Evaluierung und Hervorhebung solcher

Zielstellen an. SAIprobe, SECEDIT und RNA3D sind daher wertvolle Werkzeuge für Forschungseinrichtungen, die FISH routinemäßig durchführen.

Da das Alignment der Primärstrukturen der kritische Schritt beim Erstellen von Phylogenien ist, führen Sequenzunterschiede zu unsicheren Alignmentpositionen. Solche mehrdeutigen Positionen sind mit dem herkömmlichen Textformat multipler Alignments sehr schwer zu identifizieren und zu beheben. Das SAIviz Werkzeug stellt eine Möglichkeit bereit, die Sequenzen hinsichtlich verschiedener, farblich hervorgehobener Kriterien visuell zu überprüfen, bevor daraus Stammbäume rekonstruiert werden. Mit dem SAIviz Werkzeug können durch die farbliche Überlagerung jeder beliebigen Sequenzeigenschaft aussagekräftigere multiple Alignments erstellt werden. Dies ermöglicht dem Benutzer die hervorstechendsten Eigenheiten von Sequenzen und Alignments schnell auf einen Blick zu erfassen und die entsprechenden Regionen zu kontrollieren. Darüberhinaus sind solche aufschlußreichen Darstellungen leichter verständlich und besser zur Präsentation und Publikation geeignet.

In der Zukunft werden molekulare Systematiker oft mehr als ein Gen einbeziehen, um Evolutionsbeziehungen noch genauer rekonstruieren zu können. Bisherige Mehrfachgenuntersuchungen nutzen normalerweise den Konkatenationsansatz um Phylogenien abzuleiten. Dieser bietet eine höhere phylogenetische Genauigkeit, da er aufgrund der breiteren Datenbasis die Unstimmigkeiten von einzelgenbasierten Phylogenien besser auflöst. Das CONCAT-Werkzeug erweitert das ARB-Programmpaket und stellt dem Systematiker die Möglichkeit für mehrfachgenbasierte phylogenetische Untersuchungen zur Verfügung. Dieser Ansatz ermöglicht es spezielle phylogenetische Hypothesen zu überprüfen sowie auch eine präzisere Rekonstruktion von evolutionären Zusammenhängen zwischen den betrachteten Taxa.

Die Integration der entwickelten Werkzeuge in das ARB Softwarepaket ermöglicht das Zusammenspiel der verschiedenen Anwendungen und bietet Forschern damit eine äußerst vielseitige, integrierte graphische Umgebung zur umfassenden Sequenzanalyse.

Appendix A

Acronyms

CONCAT CONCATenation of sequences. Joining of two or more sequences from head-to-tail.

DNA Deoxy-ribonucleic acid. A macromolecule formed of repeating deoxy-ribonucleotide units linked by phosphodiester bonds between the 5'-phosphate group of one nucleotide and the 3'-hydroxy group of the next. DNA appears in nature in both double-stranded (the *Watson-Crick* model) and single-stranded forms and functions as a repository of genetic information. The information is encoded in its base sequence.

EMBL European Molecular Biology Laboratory. The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource and can be accessed via <http://www.ebi.ac.uk/embl>.

FISH Fluorescence *in situ* hybridization. A method for visualization of a genetic marker on a chromosome by use of a fluorescently labelled polynucleotide probe that hybridizes to and indicates the locus of a gene on a chromosome.

GenBank GenBank is a nucleotide sequence database maintained at National Center for Biotechnological Information (NCBI), USA. It can be accessed via <http://www.ncbi.nlm.nih.gov/Entrez>.

- GUI** Graphical User Interface. A program interface that takes advantage of the computer's graphics capabilities to make the program easier to use. Well-designed graphical user interfaces can free the user from learning complex command languages.
- HTTP** Hyper Text Transfer Protocol. The underlying protocol used by the World Wide Web. HTTP defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands.
- mRNA** Messenger ribonucleic acid. The RNA that contains the coded information, as sequences of codons, for protein synthesis.
- PCR** Polymerase chain reaction. A technique to amplify a specific region of double-stranded DNA. An excess of two amplimers, oligonucleotide primers complementary to two sequences that flank the region to be amplified, are annealed to denatured DNA and subsequently elongated, usually by a heat-stable DNA polymerase from *Thermus aquaticus* (*Taq* polymerase). Each cycle involves heating to denature double-stranded DNA and cooling to allow annealing of excess primer to template and elongation of the primers by the *Taq* polymerase; the number of amplicons, i.e. the target sequence fragments between flanking primers, doubles with each cycle.
- PDB** Protein Data Bank. PDB is a repository for 3-D structural data of proteins and nucleic acids. PDB can be accessed via <http://www.rcsb.org/pdb/index.html>.
- RDP** Ribosomal Database Project. RDP provides

ribosome related data and services to the scientific community, including online data analysis and aligned and annotated 16S rRNA sequences. It can be accessed via <http://rdp.cme.msu.edu/index.jsp>.

RNA Ribonucleic acid. A macromolecule formed of repeating ribonucleotide units linked by phosphodiester bonds between the 5'-phosphate group of one nucleotide and the 3'-hydroxy group of the next. RNA has several biological functions, most of which depend upon its ability to form sequence-specific interactions with DNA. RNA comprises the genome of some viruses.

tRNA Transfer RNA; the RNA that serves in protein synthesis as an interface between mRNA and amino acids. It carries an anticodon sequence that pairs bases with a codon of mRNA, and it binds an amino acid at its 3'-end through an ester bond.

SAI probe Evaluation of oligonucleotide probes using Sequence Associated Information (SAIs).

SAI viz Visualization of Sequence Associated Information (SAIs) on multiple alignments.

SECEDIT SECondary structure EDITor for visualizing secondary structure models of small and large subunit ribosomal RNA.

Some of these definitions are taken from the David M. Glick *Glossary of Biochemistry and Molecular Biology* (<http://www.portlandpress.com/pp/books/online/glick/search.htm>) and Wikipedia: The Free Encyclopedia (<http://en.wikipedia.org/>).

Appendix B

Definitions

Algorithm A finite set of well-defined instructions for accomplishing some task which, given an initial state, will terminate in a corresponding recognizable end-state. Generally, algorithms are implemented by a computer program for solving a particular problem.

Alignment The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Archaea One of the three primary domains of life, along with the Bacteria and Eukarya. Like bacteria, Archaea are prokaryotes, i.e. simple, unicellular organisms with small cells containing no distinct organelles. Archaeal systems of replication, transcription and translation are phylogenetically and mechanistically affiliated with those of eukaryotes, whereas the metabolic and signaling systems largely resemble those of bacteria.

Bacteria One of the three primary domains of life, along with the Archaea and

Eukarya. Bacteria are microscopic, unicellular organisms with a relatively simple cell structure lacking cell nucleus and other organelles such as mitochondria, cytoskeleton, etc. that are present in higher life forms, Eukarya.

Bioinformatics Narrowly and most appropriately defined, methods of computer science and informatics as applied to storage and retrieval of biological (particularly sequence and structural) information. More often used to designate merging of any biotechnology and information technology with the goal of revealing new insights and principles in biology.

Consensus sequence One of the forms of representation of the sequence conservation in an alignment of DNA or protein sequences. A consensus includes residues that are most frequent in each position of the alignment.

Conservation Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Deletion Absence of residue or a nucleotide base corresponding to the reference sequence in an alignment reflecting evolutionary history of two proteins or genes is often referred as deletion.

Domains Groupings of life forms according to the primary lines of descent. Three major domains of such groupings based on ribosomal RNA studies are

Bacteria, Archea and Eukarya. This form of classification is widely accepted by molecular systematists.

Eukarya One of the three primary domains of life, along with the Bacteria and Archaea. Eukaryotes are macroscopic, mostly multicellular organisms with much complex cellular structure. Mammals, birds, fish, invertebrates, mushrooms, plants are the examples of eukaryotes.

Filtering Also known as Masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently leads to spurious high scores.

Gap A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

Gene A piece of genomic DNA that encodes the synthesis of a (at least one) mRNA or structural RNA molecule.

Genome The complete DNA sequence of a life form; consists of genes and intergenic regions.

Homology	Similarity attributed to descent from a common ancestor.
Horizontal (lateral) gene transfer	Represented as acronyms HGT or LGT. Transfer of genes from one phylogenetic lineage to another, as opposed to vertical descent along a lineage.
Identity	The extent to which two (nucleotide or amino acid) sequences are invariant.
Insertion	Presence of residue or a nucleotide base corresponding to the reference sequence in an alignment reflecting evolutionary history of two proteins or genes is often referred as insertion.
Masking	Also known as Filtering. The removal of repeated or low complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence.
Motif	A short conserved region in a nucleotide or protein sequence. Motifs are frequently highly conserved parts of domains.
Multiple Sequence Alignment	An alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. Clustal W is one of the most widely used multiple sequence alignment programs.
Oligonucleotide Probe	A small stretch of nucleotide sequence (usually containing 18 – 24 nucleotides) used in detection

and identification systems in various applications of molecular biology.

Orthologs Homologous genes in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Paralogs Homologous genes within a single species that arose by gene duplication.

Parsimony A general principle of evolutionary reconstruction aimed at constructing scenarios with the minimal number of event required to account for the available data. Specifically, embedded in maximum parsimony methods for phylogenetic tree construction.

Phylogenetics A taxonomical classification of organisms based on how closely they are related in terms of evolutionary differences.

Primer An RNA sequence hybridized to a DNA template whose elongation by a DNA polymerase constitutes DNA synthesis. A random primer is a mixture of polynucleotides with all four bases at each sequence position; an arbitrary primer is a single species with a single base at each sequence position.

Prokaryotes Prokaryotes are mostly unicellular organisms lacking nucleus. Bacteria and Archaea are commonly referred as prokaryotes.

Similarity The extent to which nucleotide or

protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation.

Substitution In the ideal case that a sequence alignment genuinely reflects the evolutionary history of two genes or proteins, residues that have been aligned, yet are not identical, would represent substitutions.

Some of these definitions are taken from the following WWW sites:

1. <http://www.ncbi.nih.gov/Education/BLASTinfo/glossary2.html>
2. Wikipedia: The Free Encyclopedia (<http://en.wikipedia.org/>)

Acknowledgements

The entire work described in this thesis was done over the span of three years while working in ARB Bioinformatics Unit, Department of Microbiology, Technical University of Munich under the leadership of Dr. Wolfgang Ludwig.

With the immense gratitude, I earnestly acknowledge the expedient and able guidance, constant encouragement, invaluable suggestions, pertinent comments and keen interest rendered by my supervisor and group leader, **Dr. Wolfgang Ludwig** throughout the course of my study.

I express my heartfelt gratitude to my advisor, **Prof. Dr. Karl-Heinz Schleifer**, Chair of the Department of Microbiology, Technical University of Munich, for providing me an opportunity to carry out doctoral study under his pragmatic guidance and kind counseling. Suggestions and pertinent comments in improving the manuscript were highly valuable and gratefully acknowledged.

I am grateful to my advisory committee member, **Prof. Dr. Stefan Kramer**, Department of Bioinformatics, Technical University of Munich for his valuable suggestions and guidance during the preparation of this manuscript.

Being a biologist, initially I was stumbling over to understand the technical related problems, during which my colleague, **Ralf Westram**, was of immense help. Lengthy technical discussions including coffee breaks were very fruitful and helped me in realizing the ideas presented in this work. His timely support in this regard is thoughtfully appreciated.

General to specific discussions with my colleagues, **Dr. Harald Meier, Dr. Lothar Richter, Arno Buchner** and **Gangolf Jobb**, has certainly contributed to my overall knowledge and are worth acknowledging. Also I am thankful for their time especially for the "brain-twitching" caffeine moments.

Special thanks to Arno Buchner for having provided the accommodation for initial months of my stay and also to Dr. Lothar Richter for helping in German translation of summary section.

I warmly acknowledge the help rendered by the members of Microbial Ecology group, MPI Bremen, **Prof. Dr. Rudolf Amann, Prof. Dr. Frank Oliver Glöckner, Dr. Bernhard Fuchs** and **Dr. Sebastian Behrens** during the development of probe visualization tools.

Technical discussion with the members of Scientific Visualization Department, **Dr. Peter Kipfer** and **Jens Krüger** with respect to the OpenGL problems was very helpful and highly appreciated.

Party times spent with the members of Microbiology Lab is worth mentioning and I am thankful for the hospitality rendered by the in-mates especially **Barbara Wunner-Füßl**.

I owe a lot to my mother, brothers and friends for their encouragement, love and inspiration in my endeavors.

The financial support rendered by Federal Ministry of Education and Research (BMBF) during the course of the study is greatly acknowledged.

Many minds and hearts have contributed a lot in making my stay in Munich a very pleasant and a memorable one. Not to forget the Bavarian Alpine mountains and Weis beer. Big thanks to all of them. **"Danke vielmals!"**

References

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol.Rev.*, **59**, 143-169.

Ashelford, K. E., Weightman, A. J., & Fry, J. C. (2002) 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.*, **30**, 3481-3489.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., & Doolittle, W. F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, **290**, 972-977.

Ban, N., Nissen, P., Hansen, J., Moore, P. B., & Steitz, T. A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905-920.

Behrens, S., Fuchs, B. M., Mueller, F., & Amann, R. (2003a) 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.*, **69**, 4935-4941.

Behrens, S., Ruhland, C., Inacio, J., Huber, H., Fonseca, A., Spencer-Martins, I., Fuchs, B. M., & Amann, R. (2003b) In situ accessibility of small-subunit rRNA of members of the domains Bacteria, Archaea, and Eucarya to Cy3-labeled oligonucleotide probes. *Appl.Environ.Microbiol.*, **69**, 1748-1758.

- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2004) GenBank: update. *Nucleic Acids Research*, **32**, D23-D26.
- Brimacombe, R. (1981) Secondary structure and evolution of ribosomal RNA. *Nature*, **294**, 209-210.
- Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M., Pande, N., Shang, Z., Yu, N., & 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. *BMC.Bioinformatics.*, **3**, 2.
- Cedergren, R., Gautheret, D., Lapalme, G., & Major, F. (1988) A secondary and tertiary structure editor for nucleic acids. *Comput.Appl.Biosci.*, **4**, 143-146.
- Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., & Tiedje, J. M. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, **33** Database Issue, D294-D296.
- De Rijk, P., Van De Peer, Y., Chapelle, S., & De Wachter, R. (1994) Database on the structure of large ribosomal subunit RNA. *Nucleic Acids Res.*, **22**, 3495-3501.
- De Rijk, P., Wuyts, J., Van De Peer, Y., Winkelmanns, T., & De Wachter, R. (2000) The European large subunit ribosomal RNA database. *Nucleic Acids Res.*, **28**, 177-178.
- Delsuc, F., Stanhope, M. J., & Douzery, E. J. (2003) Molecular systematics of armadillos (Xenarthra, Dasypodidae): contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes. *Mol.Phylogenet.Evol.*, **28**, 261-275.
- Fuchs, B. M., Glockner, F. O., Wulf, J., & Amann, R. (2000) Unlabeled helper oligonucleotides increase the in situ accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes. *Appl.Environ.Microbiol.*, **66**, 3603-3607.

- Fuchs, B. M., Syutsubo, K., Ludwig, W., & Amann, R. (2001) In situ accessibility of Escherichia coli 23S rRNA to fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, **67**, 961-968.
- Fuchs, B. M., Wallner, G., Beisker, W., Schwippl, I., Ludwig, W., & Amann, R. (1998) Flow cytometric analysis of the in situ accessibility of Escherichia coli 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, **64**, 4973-4982.
- Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Expt. Zool.*, **304B**, 64 - 74.
- Gatesy, J., DeSalle, R., & Wheeler, W. (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol.*, **2**, 152-157.
- Gautheret, D., Major, F., & Cedergren, R. (1990) Computer modeling and display of RNA secondary and tertiary structures. *Methods Enzymol.*, **183**, 318-330.
- Gautheret, D., Damberger, S. H., & Gutell, R. R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.*, **248**, 27-43.
- Gutell, R. R. (1994) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucleic Acids Res.*, **22**, 3502-3507.
- Gutell, R. R., Larsen, N., & Woese, C. R. (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.*, **58**, 10-26.
- Gutell, R. R., Lee, J. C., & Cannone, J. J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, **12**, 301-310.
- Gutell, R. R., Noller, H. F., & Woese, C. R. (1986) Higher order structure in ribosomal RNA. *EMBO J.*, **5**, 1111-1113.
- Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J., & Stormo, G. D. (1992) Identifying constraints on the higher-order

structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785-5795.

Gutell, R. R., Weiser, B., Woese, C. R., & Noller, H. F. (1985) Comparative anatomy of 16-S-like ribosomal RNA. *Prog.Nucleic Acid Res.Mol.Biol.*, **32**, 155-216.

Hedges, S., Blair, J., Venturi, M., & Shoe, J. (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC.Evolutionary.Biology*, **4**, 2.

Hickson, R. E., Simon, C., Cooper, A., Spicer, G. S., Sullivan, J., & Penny, D. (1996) Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Mol.Biol.Evol.*, **13**, 150-169.

Hickson, R. E., Simon, C., & Perrey, S. W. (2000) The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Mol.Biol.Evol.*, **17**, 530-539.

Hillis, D. M. (1987) Molecular versus morphological approaches to systematics. *Annu.Rev.Ecol.Syst.*, **18**, 23-42.

Hoerter, J. A., Lambert, M. N., Pereira, M. J., & Walter, N. G. (2004) Dynamics inherent in helix 27 from Escherichia coli 16S ribosomal RNA. *Biochemistry*, **43**, 14624-14636.

Inacio, J., Behrens, S., Fuchs, B. M., Fonseca, A., Spencer-Martins, I., & Amann, R. (2003) In situ accessibility of Saccharomyces cerevisiae 26S rRNA to Cy3-labeled oligonucleotide probes comprising the D1 and D2 domains. *Appl.Environ.Microbiol.*, **69**, 2899-2905.

Kjer, K. M. (1995) Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Mol.Phylogenet.Evol.*, **4**, 314-330.

Knoll, A. H. (1992) The early evolution of eukaryotes: a geological perspective. *Science*, **256**, 622-627.

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, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., & Apweiler, R. (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, **32**, D27-D30.

Kumar, Y., Westram, R., Behrens, S., Fuchs, B., Gloeckner, F. O., Amann, R., Meier, H., & Ludwig, W. (2005) Graphical representation of ribosomal RNA probe accessibility data using ARB software package. *BMC.Bioinformatics*, **6**, 61.

Lee, J. C., Cannone, J. J., & Gutell, R. R. (2003) The lonepair triloop: a new motif in RNA structure. *J.Mol.Biol.*, **325**, 65-83.

Lee, J. C. & Gutell, R. R. (2004) Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. *J.Mol.Biol.*, **344**, 1225-1249.

Ludwig, W., Amann, R., Martinez-Romero, E., Schoenhuber, W., Bauer, S., Neef, A., & Schleifer, K. H. (1998) rRNA based identification systems for Rhizobia and other bacteria. *Plant Soil*, **204**, 1-9.

Ludwig, W. & klenk, H. P. (2001) Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. *Bergey's Manual of Systematic Bacteriology*. (ed. by G. M. Garrity), Vol. 1, pp. 49-66.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., & Schleifer, K. H. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363-1371.

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.

Maidak, B. L., Larsen, N., McCaughey, M. J., Overbeek, R., Olsen, G. J., Fogel, K., Blandy, J., & Woese, C. R. (1994) The Ribosomal Database Project. *Nucleic Acids Res.*, **22**, 3485-3487.

Martinez, H. M. (1988) An RNA secondary structure workbench. *Nucleic Acids Res.*, **16**, 1789-1798.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., & Dubchak, I. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046-1047.

McClure, M. A., Vasi, T. K., & Fitch, W. M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol.Biol.Evol.*, **11**, 571-592.

Morrison, D. A. & Ellis, J. T. (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol.Biol.Evol.*, **14**, 428-441.

Mueller, F. & Brimacombe, R. (1997) A new model for the three-dimensional folding of Escherichia coli 16 S ribosomal RNA. II. The RNA-protein interaction data. *J.Mol.Biol.*, **271**, 545-565.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., & O'Brien, S. J. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614-618.

Nei, M., Xu, P., & Glazko, G. (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences*, **98**, 2497-2502.

Noller, H. F. (1984) Structure of ribosomal RNA. *Annu.Rev.Biochem.*, **53**, 119-162.

Noller, H. F. (1991) Ribosomal RNA and translation. *Annu.Rev.Biochem.*, **60**, 191-227.

Noller, H. F., Kop, J., Wheaton, V., Brosius, J., Gutell, R. R., Kopylov, A. M., Dohme, F., Herr, W., Stahl, D. A., Gupta,

- R., & Waese, C. R. (1981) Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.*, **9**, 6167-6189.
- Pozhitkov, A. & Tautz, D. (2002) An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification. *BMC.Bioinformatics*, **3**, 9.
- Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798-804.
- Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., & Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences. *Genome Research*, **10**, 577-586.
- Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E. D., Hardison, R. C., & Miller, W. (2003) MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.*, **31**, 3518-3524.
- Shapiro, B. A., Maizel, J., Lipkin, L. E., Currey, K., & Whitney, C. (1984) Generating non-overlapping displays of nucleic acid secondary structure. *Nucleic Acids Res.*, **12**, 75-88.
- Stern, S., Changchien, L. M., Craven, G. R., & Noller, H. F. (1988) Interaction of proteins S16, S17 and S20 with 16 S ribosomal RNA. *J.Mol.Biol.*, **200**, 291-299.
- Suchard, M. A., Kitchen, C. M., Sinsheimer, J. S., & Weiss, R. E. (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst.Biol.*, **52**, 649-664.
- Suggs, S. V., Hirose, T., Miyake, T., Kawashima, E. H., Johnson, M. J., Itakura, K., & Wallace, R. B. (1981) Use of synthetic oligodeoxyribonucleotides for the isolation of specific cloned DNA sequences. *Developmental biology using purified genes*. (ed. by D. Brown and C. F. Fox), pp. 683-693. New York: Academic Press Inc..
- Teeling, H., Lombardot, T., Bauer, M., Ludwig, W., & Glockner, F. O. (2004) Evaluation of the phylogenetic position of the planctomycete 'Rhodopirellula baltica' SH 1

by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. *Int.J.Syst.Evol.Microbiol.*, **54**, 791-801.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.

Tung, C. S., Joseph, S., & Sanbonmatsu, K. Y. (2002) All-atom homology model of the Escherichia coli 30S ribosomal subunit. *Nat.Struct.Biol.*, **9**, 750-755.

Van De Peer, Y. & De Wachter, R. (1997) Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *J.Mol.Evol.*, **45**, 619-630.

Van De Peer, Y., Robbrecht, E., de, H. S., Caers, A., De, R. P., & De Wachter, R. (1999) Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Res.*, **27**, 179-183.

Van De Peer, Y., De Rijk, P., Wuyts, J., Winkelmanns, T., & De Wachter, R. (2000) The European small subunit ribosomal RNA database. *Nucleic Acids Res.*, **28**, 175-176.

Watson, J. D. & Crick, F. H. (1953) The structure of DNA. *Cold Spring Harb. Symp. Quant. Biol.*, **18**, 123-131.

Weiser, B & Noller, H. F. (1995) XRNA: Auto-interactive program for modeling RNA. The Center for Molecular Biology of RNA, Santa Cruz, California: University of California; Internet: <ftp://fangio.ucsc.edu/pub/XRNA>.

Wheeler, W. C. (1994) Sources of ambiguity in nucleic acid sequence alignment. *EXS*, **69**, 323-352.

Wheeler, W. C., Gatesy, J., & DeSalle, R. (1995) Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol.Phylogenet.Evol.*, **4**, 1-9.

- Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr., Morgan-Warren, R. J., Carter, A. P., Vonnheim, C., Hartsch, T., & Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327-339.
- Winnepenninckx, B., Van De, P. Y., Backeljau, T., & De, W. R. (1995) CARD: a drawing tool for RNA secondary structure models. *Biotechniques*, **18**, 1060-1063.
- Woese, C. R. (1987) Bacterial evolution. *Microbiol.Rev.*, **51**, 221-271.
- Woese, C. R., Gutell, R., Gupta, R., & Noller, H. F. (1983) Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol.Rev.*, **47**, 621-669.
- Wolf, Y. I., Rogozin, I. B., & Koonin, E. V. (2004) Coelomata and Not Ecdysozoa: Evidence From Genome-Wide Phylogenetic Analysis. *Genome Research*, **14**, 29-36.
- Wuyts, J., de Rijk, P., Van De Peer, Y., Winkelmans, T., & De Wachter, R. (2001) The European Large Subunit Ribosomal RNA Database. *Nucleic Acids Res.*, **29**, 175-177.
- Wuyts, J., Van De Peer, Y., Winkelmans, T., & De Wachter, R. (2002) The European database on small subunit ribosomal RNA. *Nucleic Acids Res.*, **30**, 183-185.
- Yamamoto, K., Sakurai, N., & Yoshikura, H. (1987) Graphics of RNA secondary structure; towards an object-oriented algorithm. *Comput.Appl.Biosci.*, **3**, 99-103.
- Yang, Z. (1996) Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J.Mol.Evol.*, **42**, 587-596.
- Zuckermandl, E. & Pauling, L. (1965) Molecules as documents of evolutionary history. *J.Theor.Biol.*, **8**, 357-366.