Authorized for Publication

All changes requested by the Doctoral Examination Committee have been incorporated into the thesis at hand.

Bremen,

Chair of the Doctoral Examination Committee

Max Planck Institute
for Marine Microbiology

International
Max PlanckResearch School
of Marine Microbiology

JACOBS
UNIVERSITY

School of Engineering
and Science

Scalable bioinformatic methods and resources for ribosomal RNA gene based studies

by

Dipl.-Inf. Elmar Alexander Prüße

A thesis submitted in partial fulfillment of requirements for the degree of

DOCTOR OF PHILOSOPHY

in Bioinformatics

Approved Thesis Committee:

Prof. Dr. Frank Oliver Glöckner (chair)
Max Planck Institute for Marine Microbiology
Jacobs University

Prof. Dr. Marc-Thorsten Hütt
Jacobs University

Dr. Wolfgang Ludwig
Technische Universität München

Date of Defense: December, 9th 2011

Für meine Großeltern, Eltern und Geschwister

Thesis Abstract

The identification and classification of microorganisms relies heavily on the interpretation and manipulation of genetic material. In constrast to for example plants or animals, microbes have few easily observed morphological or phenetic traits by which they can be distinguished. Yet, microorganisms are ubiquitous, having adapted to essentially every environment on earth. The extreme diversity that can therefore be expected is observable on a genetic level. In order to structure microbial life into taxonomic hierarchies and assess both diversity and relative abundances, molecular and computational methods make use of marker genes. In microbiology, the most frequently used marker genes are the small and large subunit (SSU and LSU) ribosomal RNA (rRNA) genes (16S/18S and 23S/28S, respectively). Their popularity in combination with technological progress, especially relating to sequencing methods, has created a vast pool of characterized SSU and LSU gene sequences. The breadth of available and described sequences is of great benefit to diversity studies, as it enhances the precision at which organisms can be identified. The wealth of information inherent in this pool of data can also be harnessed in phylogenetic studies. However, the work-flows employed were developed at a time when sequence data was scarce and expensive, thus made no consideration of scalability in their design. Yet today, sequence data has become both cheap and abundant.

With the SILVA database project we have created a central resource that provides a comprehensive collection of preprocessed, high quality sequence data. The databases include both the small and the large subunit rRNA genes (SSU and LSU) and cover all three domains. The sequences are quality controlled, enriched with contextual data from diverse sources and mutually aligned. A taxonomically labeled phylogenetical guide tree is included with the databases. Standardized subsets of the databases are offered to address the competing demands for comprehensiveness (Parc dataset), optimal quality (Ref dataset) and manageable database size (RefNR dataset). The alignment tool SINA was developed for use in the SILVA pipeline and made generally available. SINA pursues an add-to-alignment approach using partial order alignment (POA) techniques and a modified dynamic programming recursion that guarantees fixed alignment width. SINA is sufficiently reliable and robust to allow unsupervised multiple sequence alignment (MSA) computation. As the

sequences are aligned individually, it also scales very well to large sequence numbers. Scalability limitations in the ARB software for sequence analysis were resolved. This included porting ARB to 64 bit architecture, fixing database schema limitations and improving performance and usability. Several tools have been implemented as part of the SILVA web interface. These allow extracting arbitrarily defined subsets through search and filtering mechanisms, aligning user submitted sequence data and evaluating probes using entire respective SILVA database. Three related studies aiming at improving the primary data situation have been completed. A standardization effort was undertaken to increase the availability of complete and consistent contextual data. A comparison between SSU and LSU resolution based on the Global Ocean Survey (GOS) meta-genomes showed the potential of relying on LSU data instead of or in addition to SSU data. Lastly, the large amount of high quality sequence data in the SILVA database and the mechanisms developed to build these databases were employed in an evaluation of commonly used primers.

Acknowledgements

The road has been arduous and challenging, but also rewarding, in ways different from what I had expected. I am glad I did not have to walk it alone. To all of you who have accompanied me on my way, who offered guidance at crossroads, who helped me along the sometimes rough and winding paths, who enriched my journey with precious morsels of wisdom, insightful discussions and heated debates – thank you! Thank you for bearing with me when I was on a roll, or thought I was. Thank you for listening, even when, as I know has often been the case, I was going on endlessly about ideas of only marginal interest to you. I am glad I did not walk this road alone, for had it been that way, my journey would have ended prematurely. Instead, due to you all, it became a rich, enlightening and, small exceptions notwithstanding, joyful experience.

With me along the entire way was my supervisor Prof. Dr. Frank Oliver Glöckner. His vision, perseverance and relentless attention to minute details were essential to the success of the entire SILVA project, and by extension to this thesis. Yet, the high standards he demands – and sets by example – would have been merely frustrating, were it not for his almost uncanny ability to recognize the potential and bring out the best in the minds he gathers around him. I feel honored to have had the chance to be part of this team. Thank you, Frank Oliver, for your patience and support.

My gratitude also goes to Prof. Dr. Marc-Thorsten Hütt and Dr. Wolfgang Ludwig, who completed my thesis and dissertation committee. Your perspective has helped me greatly in giving focus and structure to this work.

I would like to thank all the people with whom I have worked together in the past years and with whom I have collaborated in accomplishing the work presented in this thesis: the ARB team in Munich, especially Ralf Westram; the SILVA team, especially Christian Quast and Jörg Peplies; Carsten John and the entire Molecular Genomics Group, especially all of you!

One of the greater challenges in this thesis was parting with the belief that computer science alone is enough to solve bioinformatical problems. Verena Salman, Melissa DuDe, Ivo Kostadinov, Pelin Yılmaz, Anna Klindworth – thank you for imparting me not only with domain specific knowledge, but also with your enthusiasm and the fascination for marine microbiology.

Lastly, I would like to thank all my friends and my parents, my grandparents, my sisters Silvia and Natalie and my brother Felix for giving me unconditional, unwavering support and being there for me, whenever I needed them. Thank you!

Elmar Prüße
November 2011

Contents

List of Figures

List of Tables

Acronyms

A Adenine. 18

BLAST Basic Linear Alignment Search Tool. 18

BLOSUM Block Substituion Matrix. 17

CARD-FISH catalyzed reporter deposition FISH. 34, 131, 142

COI cytochrome oxidase I. 4

DAG directed acyclical graph. 20

DDJC DNA Database of Japan. 5, 24

DNA deoxyribonucleic acid. 3, 13, 15

DP dynamic programming. 16, 17, 32

EBI European Bioinformatics Institute. 150

ELIXIR European Life sciences Infrastructure for Biological Information. 150

EMBL European Molecular Biology Laboratory. 5, 150

ENA European Nucleotide Archive. 24

ETWG Eukaryotic Taxonomy Working Group. 153, 158

FISH fluorescence *in situ* hybridization. 13, 30, 33, 127, 142

G Guanine. 18

GOS Global Ocean Survey. iv, 29, 33

GPS Global Positioning System. 23

GSC Genomic Standards Consortium. 23, 29, 34

GTR Generalized Time Reversible. 18

GUI graphical user interface. 159

HMM hidden Markov model. 19

HSP highest scoring segment pair. 19

HV hyper variable. 128

INSDC International Nucleotide Sequence Database Collaboration. 5, 24

LCA lowest common ancestor. 147

LSU large subunit rRNA gene. 6, 148

LTP Living Tree Project. 31

MIGS Minimum Information about a Genome Sequence. 23, 34

MIMARKS Minimum Information about a Marker gene Sequence. 24, 31, 34

MIMS Minimum Information about a Metagenome Sequence. 23, 34

MIxS Minimum Information about any Sequence. 24, 34

ML Maximum Likelihood. 22, 151

MP Maximum Parsimony. 22, 33, 151

MSA multiple sequence alignment. iii, 19, 20, 27, 31, 153

mtDNA mitochondrial DNA. 4

NAST nearest alignment space termination. 21

NCBI National Center for Biotechnology Information. 5

NGS next generation sequencing. 159

NJ neighbor joining. 21, 151

PAM Point Accepted Mutation. 17

PCR polymerase chain reaction. 13, 33, 34, 127, 128, 141, 148

POA partial order alignment. iii

qPCR quantitative PCR. 13

RDBMS relational database management system. 147

rDNA ribosomal RNA gene. 127, 128, 141

RDP Ribosomal Database Project. 5

rRNA ribosomal RNA. 4, 6, 25, 142

SAGA Sequence Alignment by Genetic Algorithm. 20

SMRT single-molecule real-time. 128, 130, 144

SP-score sum-of-pairs score. 19

SSU small subunit rRNA gene. 6, 10, 148

T Thymine. 18

U Uracil. 18

UPGMA Unweighted Pair Group Method with Arithmetic Mean. 21

Part I

Preamble

CHAPTER 1

Introduction

Life in its multitude of forms, in the complexity of detail and the beautiful simplicity of recurring patterns it exhibits, has always been a source of fascination. For millennia, mankind has sought, through observation and deduction, to understand its behavior and the source of its existence. The level of understanding of its inner functioning we take for granted today, however, has been a relatively recent development. Only with the technologies developed in the course of the 20th century has it become possible to observe life at a subcellular level, to analyze these observations and to draw conclusions about the mechanisms behind life. An example of such technology is X-ray crystallography, developed in the wake of the discovery of X-rays by Roentgen shortly before the turn of the century [135, 238, 260]. This method is used to determine the atomic substructure of crystals, and, among many other discoveries, allowed Watson and Crick to discover the structure of deoxyribonucleic acid (DNA) [302]. A host of other technologies, such as electron microscopy, allowed slowly unraveling the sub-cellular mechanisms constituting what we summarily call life [1, 133, 232, 240, 246]. The perhaps most consequential insight gained was the recognition of mechanisms for information processing as the core of the self maintenance and reproduction capabilities of life [46]. Interacting, information carrying macromolecules composed of chains of smaller building blocks, amino acids and nucleic acids, define a cell's behavior. The genomic DNA contained within each cell can be considered to be its blueprint. The genome contains, in surprisingly modular structure [106], instructions for building each of the components comprising the cell. A segment of a genome coding for a particular component is termed a gene. These genes often relate directly to traits observable in the entire organism. Mapping genes to traits and modeling the expression of genes and their recombination during sexual reproduction has served well to explain hitherto mysterious behavior, such as the laws of inheritance observed by Mendel [189, 242]. The concept of gradual evolution from a common ancestor hypothesized by Darwin and Wallace

[47, 296] can also be neatly explained on a genomic level [279], although the complex mechanisms accelerating evolution beyond what a mere combination of random point mutations and selection could achieve have yet to be fully understood [147, 178]. The dichotomy of observable evolution in whole organisms and individual genes has led to the proposition of "the selfish gene" by Dawkins [48], who suggested that the former is but an artifact of the latter.

Whether or not that is the case, evolution is observable in individual types of macromolecules, which can therefore be used as a "molecular clock" to trace the evolution of organisms [319]. The value of analyzing the evolutionary relationships among organisms also goes beyond more philosophical interests, such as the quest for the origins of life. Evolution is by definition a historical process, and just like it is necessary to understand the history of a society to understand its culture, the history of life must be understood to explain its current shape. Furthermore, lineage is a natural metric for a hierarchical classification structure. Biological taxonomy had already pursued such a structuring of organisms prior to the development of molecular methods, relying solely on phenotypic characters. While these methods remain in use where phenotypic data is sufficiently informative or where no genetic data can be obtained, as for example in the classification of organisms based on fossilized traces (ichnotaxa) [187], microbial taxonomy is today usually informed by phylogenetic analysis of marker genes. In addition to improved objectivity, molecular phylogeny has the benefit of being able to discriminate phenotypically similar organisms. This is especially important in microbiology, where morphological simplicity is the rule, but also has zoological or botanical applications. The latter, however, focus more on identification, especially of seeds and infants, in which the identifying morphological traits are not yet expressed, than on phylogenetic classification [108, 144, 196]. Depending on scale, scope and targeted organisms, many different genes have been used as molecular markers. Zuckerkandl and Pauling [319] analyzed homologous hemoglobin and myoglobin chains; in "DNA barcoding", popular with eukaryotic organisms, the mitochondrial DNA (mtDNA) gene cytochrome oxidase I (COI) has been extensively sequenced [13, 141, 254]; in microbiology ribosomal RNA (rRNA), in particular the 16S, has become the "gold standard" for identification and classification of organisms [6, 215].

Raw sequence data, however, is hard to interpret directly. While it is possible to detect patterns visually, unaided interpretation requires extensive training and is necessarily subjective. Various methods originating from fields such as applied mathematics, statistics, signal theory, theoretical computer science or artificial intelligence research have therefore been applied in the analysis of sequence data. The basic problems of sequence analysis can be grouped into pairwise sequence alignment, homology search, multiple alignment and

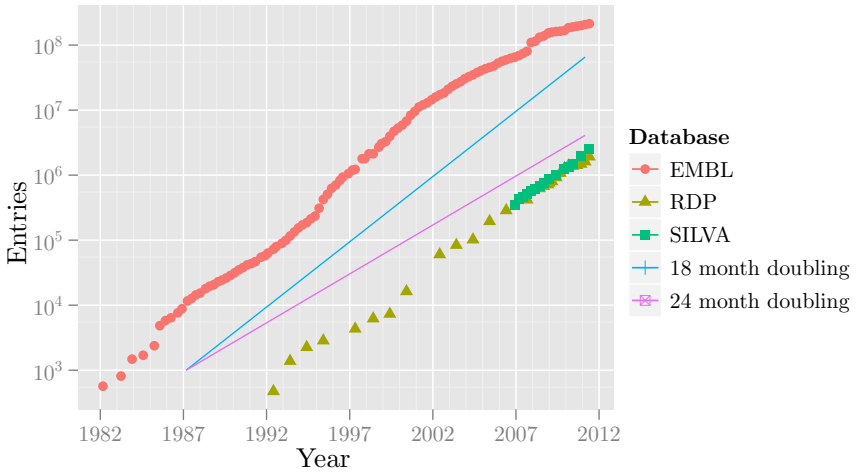


Figure 1.1 *The number of available sequences grow exponentially with a doubling rate of approximately once every 18 month. The sizes of three databases are shown in a logarithmic plot over time. The two lines indicate doubling rates of once every 18 and 24 month, respectively. The EMBL Nucleotide Sequence Database hosts sequences independent of gene type and is synchronized with the other two INSDC databases NCBI and DDJC. The RDP database hosts only SSU sequences; for SILVA, only the size of the SSU database is shown. A roughly proportional growth in the total number of available sequences and SSU sequences can be observed. The data for this plot was drawn from the database websites.*

phylogenetic tree reconstruction [60]. However, many other topics have been addressed, modeling interactions at spacial scopes ranging from the atomic to the ecological level. As both sequencing and microchip technology have progressed in exponential fashion (see Figs. 1.1 and 1.2), data oriented approaches to research questions that may be addressed by sequence characterization and analysis have become very cost effective and popular. This development towards data-driven science is not unique to biology, but extends to every science where measurements can be acquired digitally in large amounts, such as for example particle physics or astronomy. Within biology, it is also not limited to sequence data, but extents to all manners of observations that can be made in an automated, high-throughput fashion, ranging from satellite imagery to GPS motion profiles. The term “data deluge” refers to the phenomenon of overwhelming amounts of available data [26, 115]. The data-intensive approach to science that aims at harnessing these data volumes has been called the fourth paradigm, the third being computer simulations and the first two the time honored classification of science into theoretical and experimental research [21]. In a sense, the third and fourth paradigm are merely

the application of information technology to the original two paradigms. As computer simulations enhance the power of modeling in theoretical sciences, computational data analysis enhance observational and deductive capability of experimental sciences. In biology, the two disciplines building on information technology may be distinguished along the same lines. Bioinformatics focuses on making the information inherent in the observed data accessible, whereas computational biology focuses on modeling the biological systems underlying the observed data [198]. A precise delineation of the two disciplines is impossible, as they share a large zone of interest. Considering that science commonly progresses through the interaction and integration of theoretical and experimental approaches, this situation is unsurprising and in fact necessary. In Muerta et al. [198], bioinformatics is defined as follows:

> **Bioinformatics:** Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

In its emphasis on tools for data management, analysis and visualization, this definition clearly places bioinformatics at the interface between computer science and the theoretical and experimental life sciences. It provides to the experimental life sciences the means to interpret data, drawing on computer science for algorithmically automating data analysis and visualization and drawing on the theoretical life sciences to provide the mathematical foundations. Algorithmic, bioinformatic methods for sequence analysis are frequently published accompanied by implementations facilitating their application. Depending on whether they target casual use, an expert audience or high-throughput scenarios, the implementations take the form of web, desktop or command line oriented applications. The consistent organization of data is facilitated through bioinformatical databases, where data collection and dissemination are centralized and standardized.

Central to this thesis is the development of a bioinformatical sequence database for the marker genes small subunit rRNA gene (SSU) and large subunit rRNA gene (LSU). This database is complemented by newly developed web applications and an improved desktop application, enabling analyses based on the data contained within the database. An algorithm for sequence alignment geared towards the specific requirements in pre-existing rRNA analysis workflows is developed and implemented for use in database generation as well as via the web and desktop applications. Furthermore, the methods developed are applied to assess the taxonomic resolution offered by the LSU as compared to the SSU and to asses the taxonomic bias introduced by primers commonly used in SSU sequencing. The experiences made during the development of

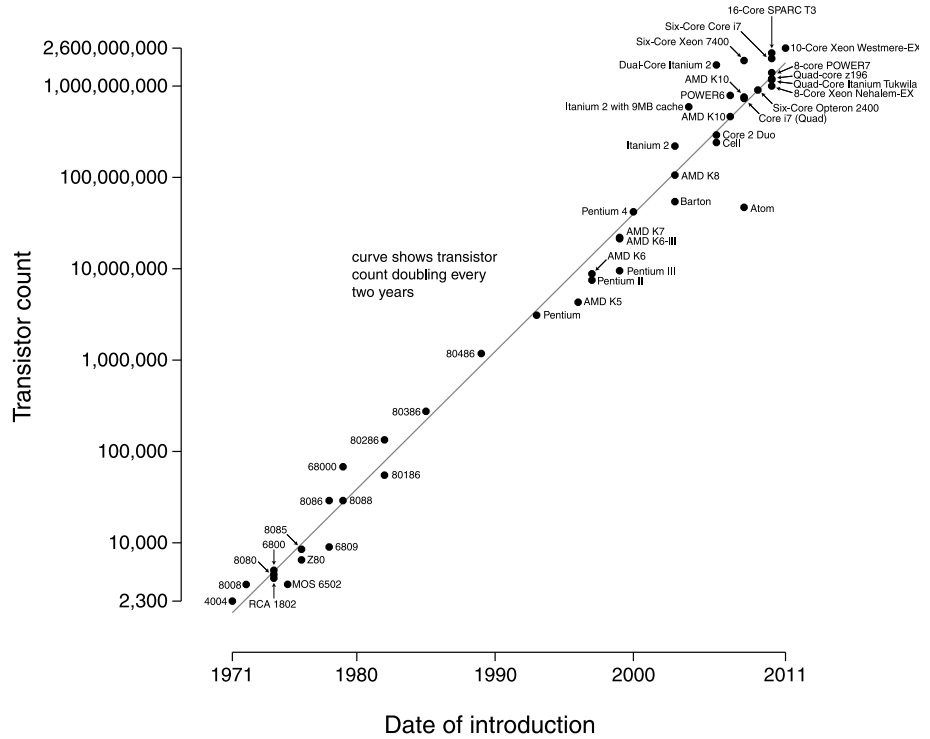


Figure 1.2 *Development of microprocessor transistor counts between 1971 and 2011. The diagonal line indicates a doubling rate of once every two years. This behavior was already predicted in 1965 by Gordon Moore[194], one of the founders of Intel, and is known as Moore's law. The number of transistors per processor is shown, because it is more easily quantifiable than computer performance or the total cost of solving a given computational problem. However, all of these are governed by the miniaturization of microprocessor structures. Increased transistor density leads to decreased production costs and to increased compute performance [245, 284]. (Figure prepared by wikipedia author "wgsimon"[1])*

database and applications are contributed to a standard aiming to improve both consistency and abundance of contextual, environmental data associated with sequence data obtained in empirical studies.

1.1 Taxonomy

Science depends at its very core on precise and accurate names. The scientific process of forming and testing hypotheses until nature can be reliably

[1] `http://en.wikipedia.org/wiki/File:Transistor_Count_and_Moore's_Law_-_2011.svg`

predicted requires names to identify the objects under investigation. While a field is still new, the terms used are ambiguous and in flux. Only once an established system of names has been arrived at, can science truly move on. Often, completing the terminology coincides with completing understanding at the level of abstraction described by the new nomenclature. The periodic table of elements may serve as an example for this. The table taught in chemistry classes today is nearly identical to the one created by Mendeleev in 1871. From this table, or more precisely from a gap within it, he was able to predict the existence of "eka-silicon" which is called germanium, today. Although he himself believed atoms to be without constituents, his table of elements was instrumental to the discoveries made by nuclear physics [137]. Eventually, the classification structure he created became common knowledge and is now one of the foundations of chemistry.

Completing the classification of the constituents of its objects of study has been a major milestone for chemistry. In the domain of biology, however, the situation is much more complex. Although a basic unit of life exists in the cell and all living organisms are composed from cells, cells and atoms cannot be further likened. Molecules are composed from atoms classified into a small number of elements (118, Barber et al. [18]), whereas (multicellular) organisms are composed from genetically identical cells, capable of forming a multitude of different tissue types. The classification of life has therefore focused on grouping individuals into species. Although an individual organism can host a multitude of smaller organisms from other species in parasitic or symbiotic relationships (see for example the human microbiome project: Turnbaugh et al. [289]), although there are exceptions to the assumption that an individual is composed from genetically identical cells (for example, a case of a phenotypically normal albeit chimeric woman is described in Tanaka et al. [274]), and although a host of other issues surround the species concept, it remains without alternatives [50, 114, 186, 228].

Historically, the concept of a "species" as the basic systematic unit of organizing life was defined as a group of individuals that through interbreeding are capable of producing fertile offspring. In pre-darwinian times, these species were thought of as completely distinct, their organization into larger groups such as plants and animals based upon physical appearance and functional capability. Darwin's theory of evolution extended the concept of descent from individuals to species and life as a whole ("gradualism"). Eventually, speciation was recognized as a gradual process. This, the observation of cases in which the above definition of species does not satisfy the criteria for mathematical equivalence, and the inapplicability of interbreeding to asexually reproducing organisms have perforce made "species" a term of only vague definition. Quoting Darwin himself, "No one definition has satisfied all naturalists; yet every

naturalist knows vaguely what he means when he speaks of a species." Both the difficulty of defining the term and the demand for doing so are outlined by the multitude of publications discussing the topic until today (see e.g. Lee [156] or Sites Jr and Marshall [257]).

While the recognition of evolution blurred the definition of a species it was of immense value to the classification of life. Taxonomies outside of biology are always troubled by the fact that there is no single, correct classification hierarchy. Vehicles, for example, could be classified according to the engine type, the number of wheels, the passenger capacity or even their color. All taxonomies based on any combination of these criteria are equally valid. If, however, every individual to be classified originates from a parent individual, the topology of the correct taxonomy can be defined by demanding that all individuals within a group share a common ancestor. Thus, current biological taxonomies are always at least informed by phylogenetics, which is the study of the evolutionary relationships between organisms. Named groups of organisms are called taxa (singular: taxon). Taxa comprising only organisms which are descendants from a shared ancestor are called monophyletic taxa. However, groups that are not monophyletic remain in common usage for pragmatic as well as historic reasons.

Modern biological taxonomy still reflects the work of Carl Linnæus, whose Systema Naturæ [161] marks the founding of the discipline. His scheme for naming species, the binomial nomenclature, remains in use until today. His system of fixed ranks, kingdoms, classes, orders, genera and species, is still recognizable in currently accepted taxonomies. Yet, his work predates Darwin, and thus makes no consideration of evolutionary relationships, but focuses on morphological and behavioral traits. After Darwin published his "On the origin of species" [47], Ernst Haeckel published one of the first revised taxonomies in "Generelle Morphologie der Organismen" [101], introducing the kingdom Protista (Fig. 1.3). He was also the first to formally describe the amniotes, the subgroups of which are a frequently used example for non-monophyletic groups. Traditionally, the amniotes were separated into the classes Reptilia, Aves and Mammalia (Fig. 1.3). Birds, however, have descended from reptiles, making the traditional class Reptillia a paraphyletic group as it excludes the sub-group Aves. Aves and Mammalia together form the polyphyletic group of warm blooded animals (Fig. 1.4). The designation of Reptilia as the Amniota that are not Aves or Mammalia can be argued for as these form a cohesive group sharing a common mode of life [30]. Equally, the term "warm blooded" is useful as it describes a group of organisms with a significant shared trait – even though the similarity is a homoplasy, the trait having developed independently in a case of convergent evolution.

Constructing a purely monophyletic taxonomy is very difficult to do, as

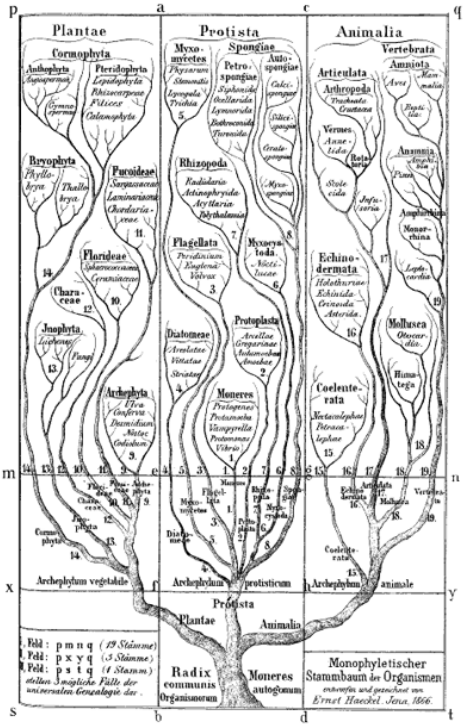


Figure 1.3 *Haeckel's Monophyletic Tree of Life*

it depends on reconstructing a correct phylogeny. Ultimately, the difficulty is evident in the ongoing process of revising and discussing the structure of the tree of life [36, 38, 153]. Archaeological evidence is limited to the fossilized traces that have survived time – bones and imprints – and are costly to acquire. Molecular evidence, on the other hand, can today be acquired cheaply in large amounts, but it is, for the most part, limited to extant organisms. A further challenge is finding characters that are comparable across the investigated taxa. A character deriving from skull shape can only be used to compare taxa actually having a head. Fox, Pechman and Woese therefore used the prokaryotic small subunit ribosomal rRNA, the 16S, in their research [79]. Including the eukaryotic variant 18S, the SSU is present in all known organisms, allowing the the construction of a phylogeny encompassing any living organism. This phylogeny led them to propose three urkingdoms [304], today known as the three domains of life, the *Archaea*, *Bacteria* and *Eukarya* [305] (Fig. 1.5). In addition to its ubiquity, the SSU is sufficiently conserved to allow reliable separation into characters for phylogenetic analysis using multiple sequence alignment techniques [79, 213]. However, even though lateral transfer of the SSU is thought to be unlikely [79], a purely SSU based phylogeny can ulti-

mately only show the history of that particular gene. Protein phylogenies also contradict the results of rRNA phylogenies – as well as each other – especially at the root of the tree of life [78]. Approaches incorporating multiple genes have therefore been pursued [38]. In order to model gene transfer, Dagan et al. [45] replaced the binary tree commonly assumed in phylogeny with a split network computed from 562,321 proteins extracted from 191 completely sequenced genomes. While other studies had come to the conclusion that the root of prokaryotic life lies within the domain *Bacteria* [35, 149], their analysis concluded that the greatest divide can indeed be found between the domains *Bacteria* and *Archaea* [45].

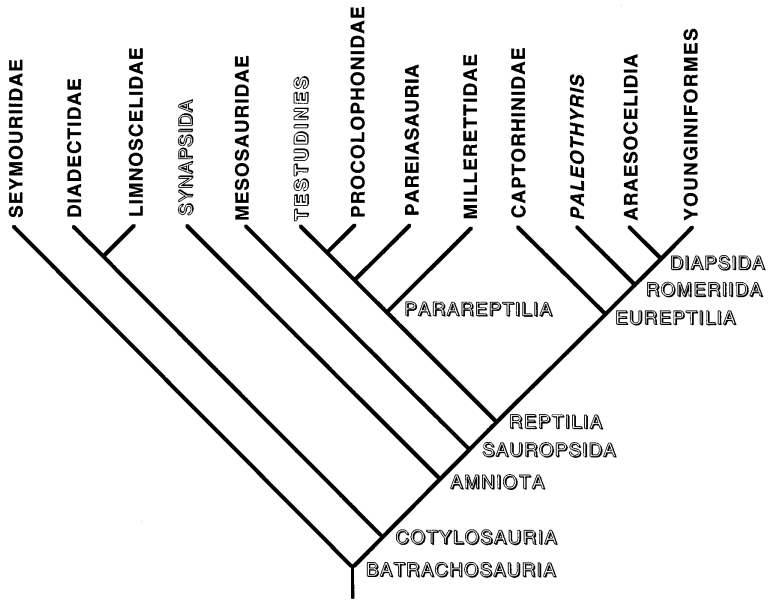


Figure 1.4 *Amniote phylogeny proposed by Laurin and Reisz. The tree was reconstructed from 13 taxa and 124 morphological characters using PAUP 3.1 [271]. Synapsidia includes the Mammalia and Diapsida the Aves[153]*

As illustrated by the example of the three domains proposed by Woese, taxonomy can be informed and refined by phylogenetic research. Yet, at its core, taxonomy remains a hierarchy of labels used as a basis for communication. The imposition of semantic meaning, such as monophyly, merely serves to moderate what would otherwise be an arbitrary structure. As altering a taxonomy is prone to cause confusion and inconsistencies with published literature, phylogenetic evidence is only incorporated slowly and after careful consideration. Ultimately, the purpose of taxonomy is to serve as a framework for classification and identification of organisms in a stable and reliable manner.

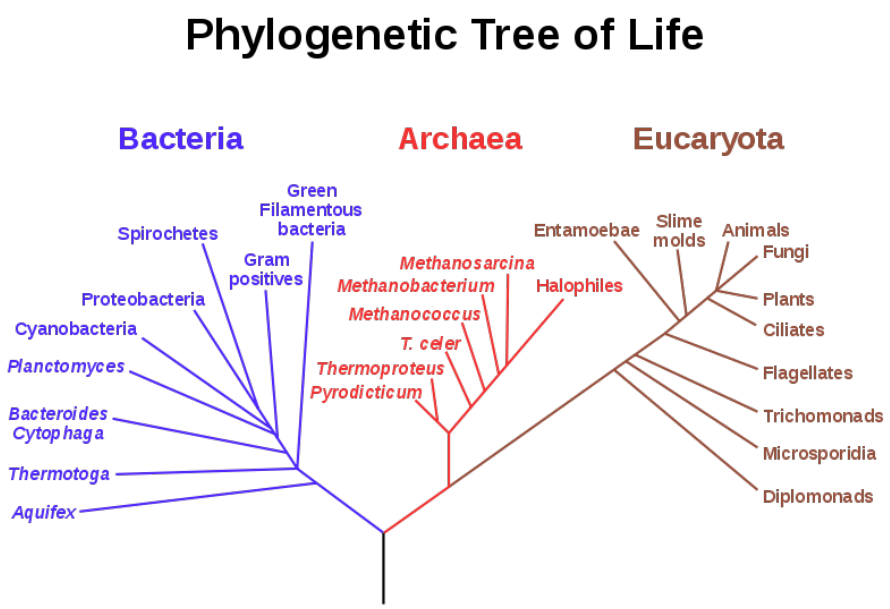


Figure 1.5 *The Phylogenetic Tree of Life as proposed in Woese et al. [305]. (Figure courtesy of NASA)*

1.2 Identification and Classification of Organisms in Microbiology

The study of microorganisms, microbiology, has from its very beginning fought with the difficulty of observing its miniscule subjects. They are invisible to the naked eye and even the use of a microscope only reveals a number of different shapes, colors or formations that is insufficient to identify members of a species. This issue is compounded by the fact that such directly visible features are not stable throughout an individual's life cycle (endospore formation for example, as first described by Koch [140]). For many decades, the only method available for identification was the preparation of cultures. By adding the precise conditions, temperature and composition of the growth medium to the list of features and by observing the growth patterns, enough information can be gathered to identify a species or strain of microorganisms [285]. This process is, however, not only tedious and time consuming, but also limited to species for which a suitable environment can be created in the laboratory. Microorganisms depending on complex environments, such as for example a host organism or on extreme environments, such as high pressure or temperature, are difficult or nearly impossible to culture [6, 287]. Yet, these organisms are both abundant and diverse in nature and are therefore critical to understand-

ing microbial ecology as a whole [230]. Amann et al. [6] estimated that over 90%, possibly up to 99% of the microbial diversity yet resists cultivation.

While both visual observation through microscopes and the preparation of cultures remain important methods used in microbiology, they are insufficient for studies in which the exact structure or composition of the microbial community present in a sample is sought. Cultivation-independent methods for identification and quantification of microorganisms have been developed in answer to this challenge. In the full-cycle rRNA approach [6] (see Fig. 1.6), the fluorescence *in situ* hybridization (FISH) method is used to highlight cells belonging to a specific group of organisms. In this approach, short signature sequences specific to a target group of organisms are determined from sequenced rDNA genes. Oligonucleotides complementary to these signatures are then synthesized and labeled with fluorescent dye. These oligonucleotides, termed probes, are introduced into the cells where they are hybridized to matching rRNA molecules. Not hybridized probes are removed in a washing step. As many rRNA instances of the rDNA gene exist in a cell [213], sufficient optical signal can be obtained to visually identify the labeled cells under a microscope. The FISH method has the advantage of allowing whole, fixed cells to be enumerated. It is therefore both fully quantitative and location preserving. These properties are especially important when investigating symbiotic or parasitic relationships as co-occurrence or inhabited tissues can be visually inspected. An alternative method is dot blotting, where probes are hybridized to extracted DNA or RNA. In this method, locality is lost but quantitative results may be obtained [5]. Polymerase chain reaction (PCR) based methods, such as quantitative PCR (qPCR), have also been used, but can only be considered semi-quantitative due to PCR biases [5, 37]. Recently, purely sequencing based approaches have become popular in diversity studies. Typically, primers targeting conserved regions of marker genes are used to generate large volumes of short sequences (tag sequences). The identity of the organisms is then determined by comparing the sequences to reference databases. Sequence data obtained in meta-genome projects can also be analyzed and interpreted in this manner. However, as PCR remains a necessary step in most sequencing techniques, sequencing based approaches are subject to PCR biases as well. They are employed in spite of this caveat mainly because the “next-generation” sequencing technology has made volume sequence acquisition a cost-effective opportunity for insight into biodiversity.

For a long time, chain-termination sequencing, also known as Sanger sequencing after its developer Frederick Sanger [243], has been the most prevalent method for characterizing nucleotide sequences. In brief, a single stranded DNA template is sequenced as follows: A mixture of DNA polymerase, normal and dye-labeled chain-terminating nucleotides is used to produce all prefixes

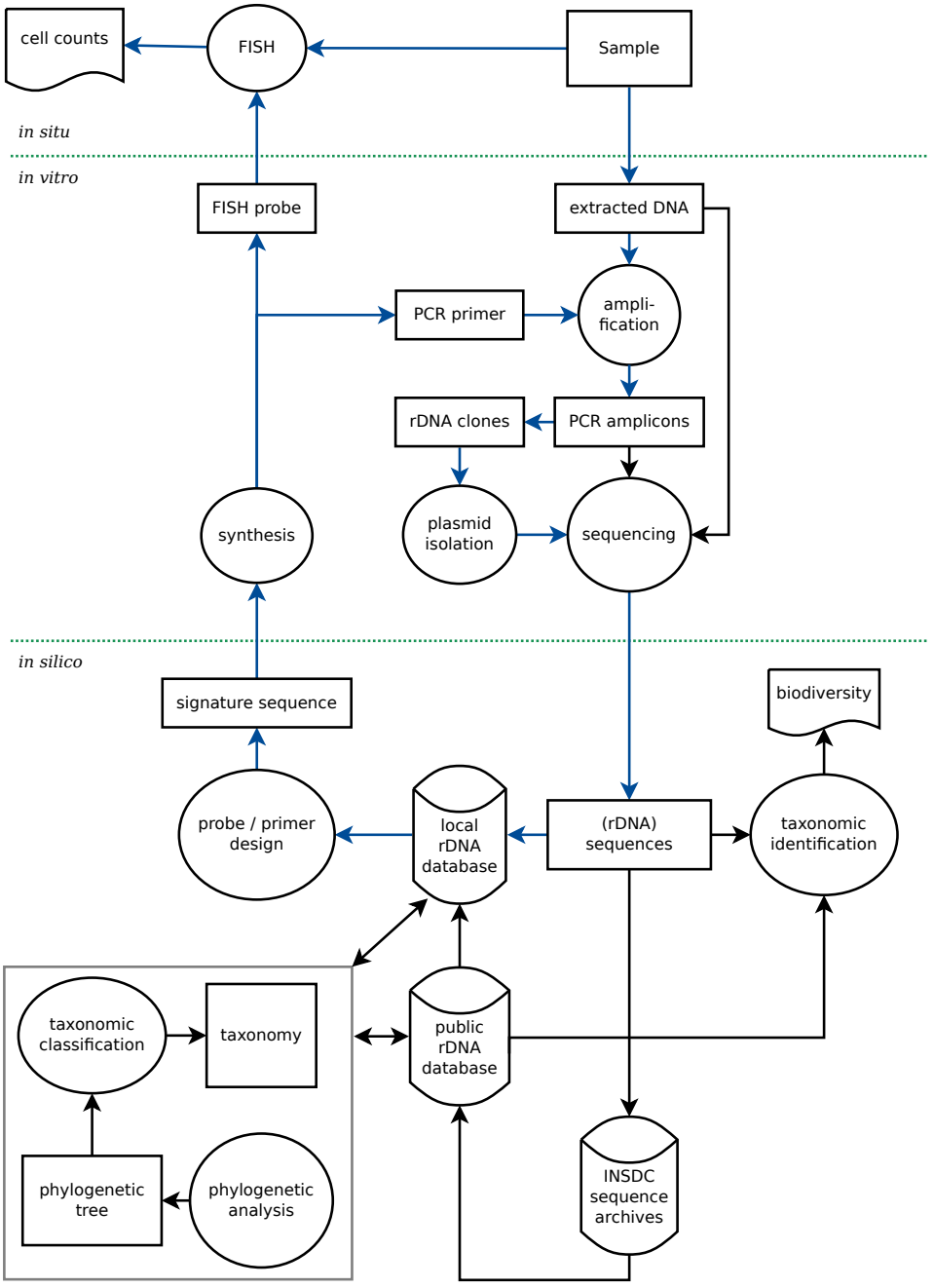


Figure 1.6 *Diagram of common work-flows in ribosomal RNA based classification and identification of microbial organisms. The traditional full-cycle rRNA approach is highlighted in blue.*

of the DNA template. The prefixes are then sorted according to their length through capillary electrophoresis. As the fragments exit the capillary, the type of dye-label affixed to the chain-terminator is determined. The resulting color histogram shows the nucleotide at the end of each prefix and thereby the composition of the sequence. Read length is limited by the sorting mechanism. As the prefix length increases, the relative length difference decreases until peaks in the histogram become indistinguishable. Today, read lengths of up 1000 bases are possible with Sanger sequencing [255]. While the second generation sequencing technologies cannot yet match such read lengths, these methods allow sequencing much larger amounts of nucleotides at lower cost. This is achieved through parallelism using cyclic array sequencing. In essence, second generation sequencing methods characterize a large number of DNA templates at once by iteratively applying enzymatic manipulations to generate an optical signal indicating the type of appended nucleotide [255]. Yet, some form of biochemical amplification is used by all methods to generate a signal of sufficient strength. Third generation sequencing technologies expand on this by directly observing the duplication of a single molecule in real-time [67, 98]. As single molecule sequencing does not require amplification, these new technologies hold the promise of quantitative sequence based identification. Collectively, next-generation sequencing technologies progress at an extremely rapid pace. A host of new methods has been devised and implemented in commercially available instruments [190]. Projections of the technological developments show that even basic sequence data analysis, rather than sequence data acquisition, will very soon become the most limiting factor [268].

1.3 Sequence Analysis

1.3.1 Alignment

Sequence alignment is a frequently used method for comparative sequence analysis. In this process, the residues comprising two sequences are arranged such that equivalent parts of the two sequences are placed beneath each other in shared columns. Empty columns are typically filled with gap characters ("-"). Thus, the fractions of the respective sequences that are identical and those that differ are made explicit and become simple to visually inspect. The ideal alignment of two sequences depends on the type of equivalence sought in its construction and interpretation. Residues may be considered equivalent if they derive from a single evolutionary event. This relation is termed homology and sought for if the alignment is to be used for phylogenetic tree reconstruction. Alternatively, residues can be considered equivalent if they serve the same function. The third general option focuses on the folded structure formed by the

sequence due to hydrogen bonds between residues. In this case, equivalence is assigned according to the role of each residue in structure formation. It is clear, however, that it is impossible to exactly determine the ideal alignment of two sequences as defined by any of these relations. Both the function and the structural position of a residue are not intrinsically defined by the sequence, but depend on external factors. Homology is also not assertable, as even identical sequence can be of distinct descent. On a smaller scale, an insertion of a specific base may have happened in distinct lineages at exactly the same position. This constitutes two distinct evolutionary events, yet they are indiscernible in the observed sequence. Computational sequence alignment can therefore only approximate the correct alignment as defined by the purposes in downstream analysis. The intent is modeled by objective functions and the alignment is computed by optimizing the result of this function. Alignment methods striving for primary structure homology typically minimize the number of evolutionary events required to cross from one sequence to the other. Models of evolution are applied to weight the events according to their observed frequency.

The most simple objective function used in sequence alignment computes the number of substitutions, insertions and deletions (together termed "indels") indicated by the alignment by the number of gap characters and the number of different residues placed in shared columns. The distance expressed by an optimal alignment according to this objective function is called the Levenshtein distance [159]. As insertions, deletions and the different substitution possibilities are not equally likely or consequential, especially in the case of aligning amino acid sequences (peptides), generalized edit distances are commonly employed which assign different weights to each edit operation. The Needleman-Wunsch algorithm [206] was the first to allow computing an optimal alignment with respect to the edit distance. This algorithm is an application of the optimization strategy dynamic programming (DP) formally described by Richard Bellman [22].

DP can be used to efficiently find the optimal solution to problems with optimal substructure and a high degree of overlap among the sub-problems. A problem is said to have optimal substructure if it can be decomposed into subproblems such that the optimal solution of the entire problem is composed of optimal solutions for the sub-problems. Finding the shortest path between two locations is an example for a problem with an optimal substructure: each section of a shortest path is itself a shortest path between its end-points. Considering that the optimal alignment of two sequences based on the edit distance can be interpreted as a shortest path of edit operations translating one sequence into the other, it becomes immediately apparent that this type of alignment has an optimal substructure. In DP-type alignment such as Needleman-Wunsch, the optimal alignment of two sequences of lengths M and N is decomposed into

the $M * N$ optimal alignments of all prefixes[2] of the two sequences. The optimal alignment of the prefixes of length m and n, the one with the lowest edit distance, can be computed from the alignment of the prefixes of lengths $(m-1, n-1)$, $(m-1, n)$ and $(m, n-1)$ by extending them with a mutation, insertion or deletion event, respectively, and choosing the alignment with the lowest resulting total distance. By iteration, the alignment can be computed in a total of $M * N$ steps. The distances and the edit operation for each prefix combination are stored in a matrix as they are computed and subsequently extracted by walking backwards from the prefix combination of length (M, N) to the combination of length $(0, 0)$. Thus, the algorithm has a complexity of $O(MN)$ in both time and space.

The Needleman-Wunsch algorithm performs a "global alignment". It assumes that the two sequences as a whole are homologous and thus comparable from the first to the last base of each sequence. The Smith-Waterman algorithm [258] instead searches for the best scoring "local alignment", comprising only a pair of sub-sequences. Like Sellers [252], they use an inverted objective function, maximizing sequence similarity rather than minimizing sequence distance. Smith et al. [259] showed that Needleman-Wunsch and Sellers are, given suitable parameters, equivalent. By default, Smith-Waterman assigns a score of 1 to matches and a score of $-\frac{1}{3}$ to mismatches. This results in an average total score of 0 for long, random sequences composed from four base types occurring with equal frequency. If the alignment of any pair of prefixes would result in a score below zero, it is instead set to zero. The local alignment is then extracted from the DP matrix by determining the cell containing the highest alignment score and following the path of edits until a cell with a score of 0 is encountered.

By using scoring functions that reflect evolutionary processes more accurately, the results of the alignment process can be improved. Affine gap scoring is used to reflect that consecutive insertions or deletions can be the result of a single evolutionary event. A linear function determines the cost of the entire gap. This was realized without impacting algorithm complexity by Gotoh through the use of additional matrices [92]. Match and mismatch scoring can be generalized by using a substitution matrix. The mechanistic specialization of match/mismatch scores is then reflected by a matrix in which the diagonal is set to the match score and the remaining cells to the mismatch score. For protein sequences, the Point Accepted Mutation (PAM) [49] matrix and the Block Substituion Matrix (BLOSUM) [110] are commonly used. The methods used to create these matrices differ, but both derive from observed data, with BLOSUM intended (and shown) to perform better at detecting remote homologies.

[2]Needleman and Wunsch actually used suffixes, but as the results are identical irrespective of sequence direction the distinction is not relevant to sequence alignment.

For nucleotide sequences, the mechanistic scoring performs well enough, but substition models are used to reflect for example the different probabilities for transisitions and transversions. A transition is a change among the purines Adenine (A) and Guanine (G) or a change among the pyrimidines G and Thymine (T) or Uracil (U), whereas a transition is a change between pyrimidines and purines. Similarly, models of DNA evolution corresponding to these matrices are used to estimate the evolutionary distance from a similarity score by accounting for unobserved mutations. This is important for tree reconstruction (see below). The Jukes-Cantor model [126] accounts only for matches and mismatches, the Kimura model [138] also distinguished between transitions and transversions. The models devised by Felsenstein [71] and Hasegawa [107] extend these by allowing for different base frequencies. The most general usable model is Generalized Time Reversible (GTR) [276].

1.3.2 Homology Search

The most important use case for local alignments is homology search. The interest here is not to make residue-level homology among a pair of sequences explicit, but to find sequences that share a common ancestor with the query sequence in a large database. Searching for the most similar sequences alone does not suffice to satisfy this intent. The database matches must also be statistically assessed to determine whether they constitute significant evidence for the homology hypothesis. Also, scalability becomes a critical issue. Space efficiency is of particular importance as exceeding the amount of available physical memory and having to resort to secondary storage (i.e. hard drives) results in a performance decrease of several orders of magnitude. Based on Hirschberg [116], Myers and Miller [203] showed a DP-alignment algorithm requiring only $O(N)$ space at typically twice the computation time. Their algorithm recursively calculates the mid-point of the conversion path between the two sequences, removing the need to store the complete DP-matrix. A reduction in both time and space complexity was achieved in the FAST-P algorithm [163] by sacrificing optimality. FAST-P and the nucleotide version FAST-N, published together along with the Smith-Waterman based search tool SSEARCH in the FASTA package [217], use word-based matching to confine the Smith-Waterman algorithm to a diagonal band. The Basic Linear Alignment Search Tool (BLAST) software [2] and its successors Gapped and PSI-BLAST [3] and BLAST+ [31] further increased search speed. While FASTA searches only for shared words, BLAST extends the word list derived from the query sequence to all words with sufficient similarity (as determined by a parameter). This allowed increasing word lengths (from two to three for protein sequences and from 6 to 11 for DNA sequences) and resulted in more significant word matches. The matched words

are then extended on each side to form highest scoring segment pairs (HSPs). Close HSPs are joined and the resulting set of HSPs evaluated for statistical significance. In catering to the wide range of applications for which BLAST is used today, it employs are large number of methods which are beyond the scope of this short description. Important to note is only the E-value which is reported for each match and indicates the number of times such a match can be expected given the size of the database. While BLAST allows subjecting the matching HSPs to Smith-Waterman alignment in some configurations, the alignment reported by BLAST is usually heuristic and thus inferior to optimal alignment techniques.

While BLAST remains the most commonly used tool for homology search, an alternative approach exists in using probabilistic models such as hidden Markov models (HMMs) to treat homology search as a problem of statistical inference [60, 129]. Until recently, however, HMM-based tools have not been able to match the speed achieved by BLAST. This has changed with HMMER3, which promises to be as fast while showing improved detection of remote homologies [61, 77].

1.3.3 Multiple Sequence Alignment

Multiple sequence alignment is the natural extension of the approach to sequence comparison followed in pairwise alignment. Visually, the sequences comprising a MSA are represented as the rows of a matrix with the residues spread over the columns such that homologous or structurally equivalent residues align. The dynamic programming algorithm used to compute pairwise alignments can also be generalized to multiple sequences by increasing the number of matrix dimensions. This approach will find an optimal solution of the alignment problem as defined by the sum-of-pairs score (SP-score) [34]. The SP-score is simply the sum of the pairwise scores of all sequence combinations. However, the number of cells of the n-dimensional matrix grows exponentially with the number of sequences. Space and time requirements therefore quickly become intractable. The problem of multiple alignment with SP-score was thoroughly investigated and classified as “NP-hard” [68, 127, 299], thus, no tractable algorithm exists. Multiple sequence alignment methods have therefore focused on heuristics to approximate the optimal alignment. The most commonly used method, termed progressive alignment, was devised by Feng and Doolittle [73]. In this method, sequences are first clustered according to their pairwise distance, creating a binary tree. Progressing from the leaves of this tree to its root, pairs of sequences are aligned to form column vectors which are in turn subjected to pairwise alignment until at the root all sequences have been mutually aligned. Commonly used alignment tools employing

this method include CLUSTALW [280], MaFFT [130–132], MUSCLE [63, 64], POA [93] and ProbCons [58]. One major drawback to the progressive alignment strategy is caused by what Feng and Doolittle termed the "once a gap always a gap" rule. A gap that is introduced at lower levels of the guide tree cannot be reverted later on in the alignment process, leading to an accumulation of errors. The method therefore depends on the quality of the guide tree. MaFFT and MUSCLE both include modes in which two rounds of progressive alignment are executed. In the first round, the distance matrix used to cluster the sequences is built using k-mer counting, that is determining the number of shared words of length k. In the second round, the MSA computed in the first round is used to determine pairwise distances. ProbCons instead minimizes error in one round by incorporating probabilistic assessment of alignment accuracy in guide tree construction and objective function. All three tools also include iterative refinement stages to further improve the total SP-score. POA includes neither refinement but uses directed acyclical graphs (DAGs) to represent MSAs internally instead of the column profiles employed by the other methods. An alternative to heuristic SP-score optimization is followed by Sze et al. [272] in PSalign by reformulating the objective function such that an optimal alignment can be computed in polynomial time. PSalign uses a graph theoretic alignment representation similar in concept to the partial order alignment used by POA but differing in its structure. Methods not relying on progressive alignment include Align-m [291] and diagonal alignment (DIALIGN) [195, 270] which compose the final MSA from many, short local alignments, and Sequence Alignment by Genetic Algorithm (SAGA) [210] which through its use of evolution to optimize the alignment allows choosing arbitrary objective functions. Other multiple sequence alignment tools of note include Kalign [152], T-Coffee [211] and M-Coffee [297].

All algorithms or tools listed above are *de novo* alignment methods that build a complete MSA from unaligned sequences. Certain use cases, however, require that sequences are incorporated into a pre-existing MSA. One example is the construction of curated MSAs. Given an already curated MSA, the ability to incorporate an unaligned sequence into the alignment without modifying the existing alignment facilitates a cycle of sequence addition and alignment review. Expert curated alignments are still considered to be superior in accuracy to computed alignments [66] and are therefore commonly used evaluate the accuracy of alignment methods [14, 229, 282, 292]. In part, the human superiority is due to the possibility of incorporating information from diverse sources in manually curated alignments. For example, experimentally verified 3D-structures can be used corroborate or refute alignment decisions. Another use case would be the addition of a sequence to an existing phylogenetic tree. Here, it is desirable to maintain the structure of the columns used as charac-

ters in previous calculations. The "fast-aligner" tool provided with the ARB software suite [177] is an alignment method supporting an "add-to-alignment" mode of operation. Similar to BLAST, this tool first searches for the longest common sub-sequence between the new sequence (query) and any sequence in the existing MSA (reference MSA). The corresponding part of the query is then placed in the same columns as the matching sub-sequence. The match is extended on both sides until a maximum number of mismatches is encountered. The process is recursively repeated with the unaligned remainders of the query. While this method is very fast, it can only be used to create a rough, preliminary alignment. Especially the ends of the alignment of a query sequence exhibit unnecessary outward scattering. Another method of aligning a sequence with a given reference MSA was introduced with nearest alignment space termination (NAST) [54]. NAST uses the BLAST software to obtain a pairwise alignment between the query sequence and the best match in the reference MSA. This alignment is then used to map the query sequence into the reference MSA via a series of gap character reintroduction and removal operations. Improved implementations using the same basic principle have been published as PyNAST [33] and as part of mothur [247]. PyNAST uses UCLUST [65] instead of BLAST whereas mothur relies on its own implementations of a k-mer search to select the reference sequence and a Needleman-Wunsch type alignment algorithm to perform the pairwise alignment. The three NAST-type alignment tools are mainly aimed at high-throughput scenarios. As each sequence is aligned individually and no mutual comparisons, such as needed to build the guide trees for progressive alignment, are required, the reference MSA based approach to MSA construction can scale linearly with the total number of sequences in the resulting MSA.

1.3.4 Tree Reconstruction

The methods used for tree reconstruction group into those that consider only the distance between taxa and those that consider multiple characters. The simplest method is Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [262], an agglomerative method for hierarchical clustering. That is, UPGMA builds the tree "bottom-up". At first, each taxon is assigned its own cluster. These clusters are iteratively merged in order of the least average distance between clusters. UPGMA assumes a constant rate of evolution and does not compute branch lengths. It is therefore not suited to true phylogenetic tree reconstruction but is only used to build the guide trees for progressive alignment. neighbor joining (NJ) [241] works in essentially the same way, but adjusts the distance to account for differences in divergence before each clustering iteration. NJ also computes branch lengths. It is a popular method for

reconstructing phylogenetic trees due to its very high speed and reasonable accuracy. However, reducing all characters to a dingle distance value looses information. Methods considering characters individually can therefore reconstruct the true phylogeny more precisely.

The two most common optimality criteria used by such methods in phylogenetic tree construction are Maximum Parsimony (MP) and Maximum Likelihood (ML). As the names suggest, MP methods search for the most parsimonious and ML methods search for the tree with the highest likelihood. MP grades a tree by the minimal number of mutations that must have occurred to explain the observed data. The number may be weighted according to a cost model associated with the types of mutations. In ML, a tree is graded by its likelihood given the observed data, which is equal to the probability of observing the data given the tree. The probabilities are calculated using models of evolution. As the search space spanned by all possible trees is too large to be searched exhaustively, the implementations of either method use various techniques to search for a good solution. In the most simple case, hill-climbing is performed starting at a number of randomly selected trees. These are iteratively improved by modifying the branching order and evaluating the resulting tree according to the objective function. As a number of equally good trees may be found, MP and ML can yield multiple resulting trees. The statistical method "bootstrapping" is often used to obtain confidence values for the inferred branches. The characters are not used directly, but the calculations are repeated multiple times using equal numbers of randomly selected characters (characters may repeat). Examples of tools for phylogenetic tree reconstruction are ARB [177], PAUP [271], RAxML [265–267] and PHYLIP [72], but many more tools implementing one or more of the above methods exist.

1.4 Holistic Data Analysis

Although dominant in volume, sequence data is just one type among many that are being accumulated in biology. Many types of image data are also being generated in volume, ranging from optical or electron microscopy to magnetic resonance tomography or even satellite imagery. All of this data is acquired under diverse experimental conditions, creating a descriptive type of data termed metadata or contextual data. The relevance of the latter becomes immediately apparent when considering data of lesser intrinsic informational content than for example biological sequences. A single temperature value contains little intrinsic information. Yet, when associated with spatiotemporal metadata, temperature measurements can be linked with measurements of humidity, wind-speed, precipitation and other factors. The integrated body of data may then be used to design and evaluate climate models and therefore

serve to predict weather conditions. Including historical data, even questions regarding long term antropomorphic influences on global climate development can be addressed. In the same manner, integrative analysis of the large body of data acquired in biological experiments can yield answers beyond the original research questions.

Integrative data analysis in biology is, however, vastly more complex than in other fields. Whereas the accuracy of e.g. Global Positioning System (GPS) data is sufficient for spatiotemporal linking of measurements in the above, meteorological example, a global frame of reference is insufficient in biology. The position of a population of microorganisms relative to the body of the inhabited host may be more relevant, for example, than global position on even millimeter scale (e.g. Costello et al. [43]). Environmental conditions also vary widely within millimeter scales, as for instance above or below the seawater or seafloor surface, or even within sediment or soil. This is not to say that a global frame of spacial reference is entirely meaningless, merely that it is insufficient for many research questions. Similarly, the time relative to local events may be relevant, examples being the beginning of an incubation experiment or stages in an organism's life cycle. Many other properties beyond spatiotemporal description offer the opportunity of data integration in "meta studies". To name but a few, data could be linked according to the medical condition of the subject under study, according to co-occurrence of organisms, according to physical or chemical environmental parameters, or even according to methodological differences. The combination of both large volumes of data and large numbers of complex descriptors make the field of integrative data analysis in biology a challenging topic [112].

Several tasks need to be addressed to simplify integrative data analysis and to enable holistic interpretation of diverse biological data in high volumes. Optimally, comprehensive descriptive data should be recorded at the time of data acquisition, including descriptors that are not directly relevant to the specific study. This data should be represented in syntactically and semantically well-defined form to allow automated data processing and meaningful interpretation. Furthermore, the data should be stored such that it can be uniformly accessed and queried. Efforts to arrive at such a situation encompass the development of standards, tools and databases. The development of generally accepted standards is especially critical. In defining which descriptors constitute a comprehensive description, these standards must walk the fine line between the breadth of descriptors potentially desirable for future analyses and the acceptable effort during data acquisition. They must also ensure that the data can be recorded syntactically and semantically consistent. Examples of recently developed standards are the Genomic Standards Consortium (GSC) standards Minimum Information about a Genome Sequence (MIGS), Minimum

Information about a Metagenome Sequence (MIMS) [74] and Minimum Information about any Sequence (MIxS) as well as Minimum Information about a Marker gene Sequence (MIMARKS) ([314], Chapter 9). Tools such as Handlebar [27] and MetaBar [104] aim at simplifying the process of recording data and ensuring standards compliance. Databases such as probeBase [168], strainInfo.net [294] or Megx.net [143] collect, integrate and curate data and provide interfaces for querying the data they hold. The three INSDC databases European Nucleotide Archive (ENA) [158], DDJC [128] and GenBank [23] maintain a synchronized archive of sequence data.

CHAPTER 2

Research Aims

The overarching aim of this thesis is to relieve scientists relying on small and large subunit (LSU/SSU) rRNA gene data analysis from the need to individually prepare and maintain suitable reference databases. A common base of reference data will also serve to improve the comparability of results from different studies. In order to achieve these primary aims, databases meeting the following criteria are to be created:

1. **up-to-date**
   The database content must be based on the most current available upstream data sources.

2. **comprehensive**
   The databases must contain all relevant data.

3. **high quality**
   The databases must be quality controlled. The data should be screened to improve the signal to noise ratio.

4. **easy to use**
   The databases must be easily accessible and the data prepared such that the effort required in its application is minimized.

From item one, up-to-date, we can immediately derive that a system for automated database preparation must be fashioned. Considering data volume and data growth rates, manual maintenance of up-to-date databases is not feasible. A certain amount of remaining manual effort will, however, be impossible to eliminate. The studies relying on the database cannot be expected to update on a daily basis. This is also not desireable, as it conflicts with the intention of improving comparability. A balance between highly current data and a manageable effort in both preparation and application of the databases

must therefore be found and the database release cycle must be adjusted accordingly.

Item two, comprehensiveness, encompasses more than merely including all published sequences of the gene in question. Descriptive data, such as taxonomic classifications, strain type, environmental parameters or referencing publications are needed assert the context of the observed sequence data. Data from multiple source databases must therefore be integrated. In combination with item one, it is important that these databases are as up-to-date as possible as well. The first task, however, will be the detection and extraction of all instances of the respective gene. We expect that a combination of keyword search to find annotated sequences and pattern-matching using the RNAmmer hidden Markov models [148] with the tool HMMer [62] will be sufficient to solve this task.

Item three, ensuring that the databases contain only high quality data, competes with comprehensiveness. Firstly, a way must be found to determine from the sequence data alone to which degree each sequence can be regarded a trustworthy observation. Anomalies and chimera resulting from sequence amplification must be excluded from the database or, if no reliable method for the identification of such sequences can be found, they need to be marked appropriately. Secondly, a way to balance quality criteria with comprehensiveness that is acceptable for all potential applications must be found. Furthermore, sequence detection, as outlined in the above paragraph, focuses on sensitivity alone. We expect that specificity to the gene in question can be derived from sequence alignment (see below).

The fourth and last item is expected to be the most challenging as it raises a number of very broad questions. How can we make it possible for researchers to work with comprehensive datasets in the face of rapidly growing data volumes? How can we keep the bioinformatical knowledge neccessary to work with the databases at a minimum? Which common tasks can be moved into centralized database preparation? For which of the balances mentioned above can we make a decision appropriate for all usage scenarios?

In answer to these questions we intend to focus primarily on a single application for which our databases will provide the “fuel”. A new, stable release of this application, the ARB software [177], must be developed. A way to reduce the entry-barrier for novice users of ARB must be found, as well ways to improve its scalability. A web interface allowing easy cross platform access to the complete databases, the means to browse and search them and a facility to generate custom subsets of the database as required by particular research questions should allow further reduction of the hardware requirements during analyses relying on our databases.

The targeted analysis scenario, as well as the quality assessment stage, also require computing a high quality MSA from all sequences comprising the databases. The accuracy of this alignment must be comparable to the accuracy achieved by manual curation. The work-flows to which phylogenetic analyses using ARB currently adhere must remain maintainable. In these work-flows, novel sequences are semi-manually incorporated into an existing MSA. This MSA is manually curated to accurately reflect the conserved secondary structure of the rRNA molecule. In previous work [225], I have shown a way of automating the process of sequence addition to an existing MSA. In this thesis, the research questions relating to sequence alignment will be to determine whether and how the accuracy can be further improved and whether the attained level of accuracy is sufficient to forego manual curation. The existing prototype and the method upon which it is based must therefore be refined and rigorously evaluated.

In summary, the primary research aims this thesis aspires to adress can be expressed as these three questions:

1. **How can we automatically build sequence databases that can serve as high quality reference for rRNA based analyses?**

2. **How can we build very large yet accurate MSAs compatible with curated reference MSAs?**

3. **How can we make working with the new reference databases possible in spite of their expected size?**

CHAPTER 3

Publication Overview

The results presented in the follwing chapters of this thesis are grouped into two parts. The first part contains the core results covering the SILVA database project (Chapters 4 and 5), the SINA sequence aligner (Chapter 6) and the ARB software suite for sequence analysis (Chapter 7). The second part covers flanking efforts in which the author has participated. Chapter 8 investigates the usefullness and taxonomic resolution of the 23S rRNA gene as compared to the 16S rRNA gene based on metagenomic samples from the GOS expedition. The MIMARKS standard and the MIxS checklist presented in Chapter 9 are a result of the efforts made by the GSC to resolve the current situtation of inconsistent and sparse annotation of sequence data with environmental parameters. Chapter 10 evaluates 16S primers for taxonomic bias both *in silico* and experimentally.

3.1 SILVA database project

SILVA is named after the latin word for forest, alluding to the sequence analysis suite "ARB", the name of which derives from *arbor*, the latin word for tree. The name SILVA was not chosen to express that the project will contain many phylogenetic trees – that is already possible within ARB – but resulted from the idea of building an encompassing infrastructure for ARB. SILVA is not limited to ARB but provides databases to all applications requiring rRNA reference data. The databases contain preprocessed sequence data that are enriched with contextual data from multiple sources, screened for quality and complemented with guide trees. Using these databases, rather than relying on unpreprocessed primary data, greatly simplifies the phylogenetic analysis of newly characterized sequences. Such analysis is frequently required in the context of a broad range of microbiological research projects, including global diversity surveys, the investigation of specific habitats or even the study of model organisms. Prior

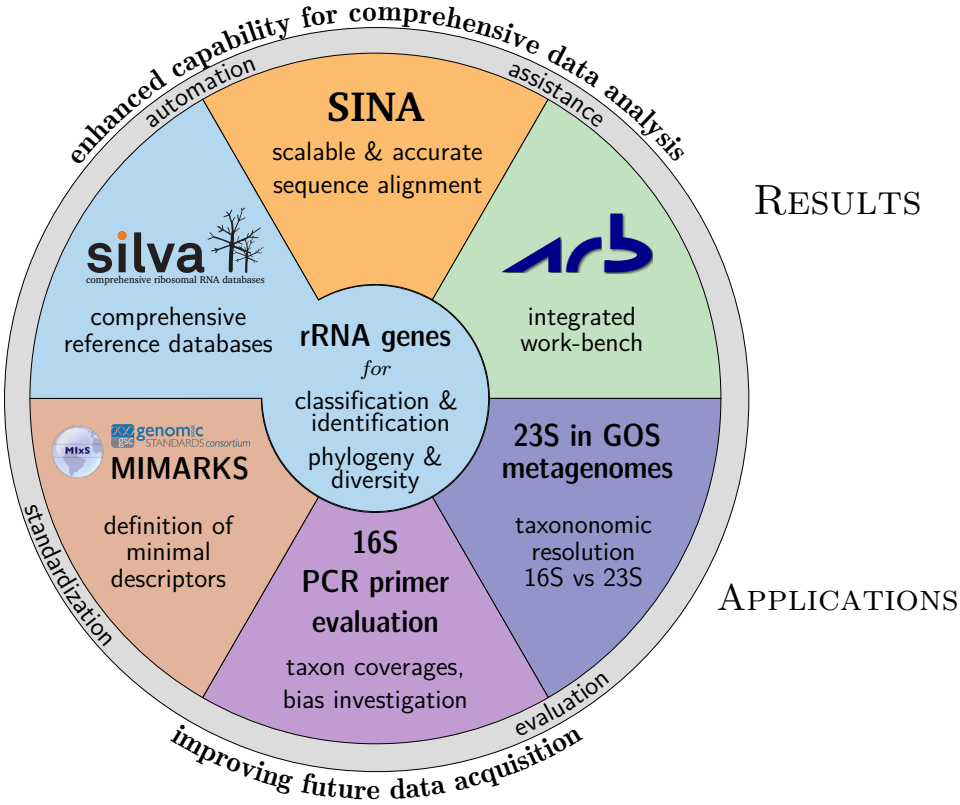


Figure 3.1 *Thesis Structure*

to the availability of the SILVA databases, it was necessary to extend existing, smaller databases manually with rRNA sequences from related organisms. This entailed BLAST searches and time consuming manual sequence alignment. Beyond pure phylogenetic work, rRNA databases are used to design group specific probes for FISH as well as primers for sequencing. Both the design and the evaluation of specificity and sensitivity achieved by such oligonucleotides depends on the use of diverse and dense reference data. SILVA addresses these needs and in providing updates to its databases at regular intervals also improves the comparability of results between studies. Previously, databases were often handed from researcher to researcher and incrementally extended, resulting in a diverse set of databases of unknown consistency.

The software behind the SILVA database project as it exists today can be grouped into two parts: the pipeline that is used to build the databases themselves and the web site that provides services operating on these databases. While originally only meant to disseminate databases in ARB format, it quickly became apparent that there was broad interest in the SILVA databases beyond the ARB user community. Tools for searching and visualizing database con-

tent, extracting subsets of the database in a variety of formats for further processing, sequence alignment compatible with the SILVA alignments and probe evaluation have therefore been developed and added to the SILVA website. Furthermore, the website has been used to host several related projects, drawing mutual benefit from the extended reach accomplished via the website. One such project is the Living Tree Project (LTP), which provides curated databases and phylogenetic trees for type strain sequences [199, 311, 312]. A user survey executed in 2010, mainly to prioritize further development of SILVA, yielded results useful to the development of the MIMARKS standard (Chapter 9).

Contributions

The SILVA pipeline was initially conceived and realised in an intense, collaborative effort by Frank Oliver Glöcker, Jörg Peplies, Christian Quast and the author of this thesis. Assessing individual contributions on the level of ideas, design, implementation and operation is extremely difficult, especially as responsibilties changed according to work load during and after the initial development phase. The author was primarily responsible for the design of the job execution system, the pipeline configuration system, the object relational mapping layer, the ARB interface, the chimera checking, the sequence alignment, the quality calculation and the web integration, whereas Christian Quast was responsible for the design of the SILVA database schema, primary data import, secondary database integration, homopolymer and vector checking. As an ongoing project, however, SILVA has been continuously improved and had to be continuously operated. The author has been responsible for pipeline operations for roughly one and a half years and was until recently in charge of web site operation.

3.2 SINA aligner

SINA is an alignment tool targeted at scenarios in which an existing, high-quality alignment needs to be extended with a large number of sequences in a robust and accurate manner. Computing *de novo* MSAs from sequence volumes as handled by the SILVA database project is not feasible as these alignment would have to be reviewed extensively with each database release. The existing ARB reference databases also included a balanced primary and secondary structure alignment. Aligning sequences *according* to this alignment, rather than *de novo*, preserves the knowledge represented in this alignment. Furthermore, the established method of extending existing phylogenetic trees within ARB requires that alignment columns remain stable.

Contributions

SINA was developed solely by the author, however, parts of SINA derive directly from the galign software developed as part of the authors "Diplomarbeit" [225]. SINA constitutes a complete rewrite of the framing components and has a flexibly configurable and easily extensible pipeline structure. A search and classification module has been added to this pipeline. The alignment module has been extended with a bypass for direct matches to subsequences found in the reference alignment. The alignment algorithm and its implementation have been extended to allow more flexible weighting of reference graph nodes and a more effective method for node weighting has been found. The issue of large insertions that was largely ignored by galign is now treated by two alternative approaches, either during or after DP alignment. Furthermore, a thorough evaluation was executed to empirically establish validity and to determine the degree of achieved accuracy as well as optimal parameters.

3.3 ARB software project

ARB is a software suite for sequence analysis, mainly targeting phylogeny and related areas, that has been under continued development for over 16 years. It is designed as a work bench type application that allows handling large amounts of homologous sequences, inspecting and curating MSAs, visualizing secondary structure and phylogenetic trees and serves as a graphical shell for command line based bioinformatics tools. With the sudden increase in data volume caused by the SILVA project, the lack of scalability of this application became an urgent problem. These were addressed in order to allow full exploitation of the databases provided by SILVA with the ARB release of 2007 and the release of version 5 of the ARB Software in 2009.

Contributions

The author has actively participated in ARB development and testing. This included the contribution of bug fixes and new features to the ARB project. It must be noted, however, that these contributions are only a fraction of the work invested into ARB in the recent years, most importantly by its primary developer Ralf Westram, but also by Yadhu Kumar and Kai Bader. A set of patches initiating the port of ARB to AMD64 architecture, remedying the most crucial scalability issue caused by limited address space was developed by the author. The full port was completed by Kai Bader. A solution to the issue of sequences not uniquely identifiable by their accession numbers, which is the case in sequences from genomes, was provided in a patch to the ARB name

server. A more generic solution has since been implemented by Ralf Westram. The high entry barrier to ARB caused by the overwhelming amount of features and options available within ARB was lowered by the introduction of an “expert mode” to which the legacy, non-standard, potentially dangerous or rarely used functions were delegated. The classification of interface elements was provided by Frank Oliver Glöckner and the implementation harnessed the remainders of a preexisting, although unusable, “novice mode”.

Minor patches include the preservation of color information during tree export for printing, fixes for rendering issues in the sequence editor, disabling the continuous mouse-over focus of interface elements, a standard compliant web browser integration, a generalization of socket configuration to unix sockets, typed import statements in input data format definitions and an interface allowing user initiated type conversion database fields, increased search speed and decreased start-up times in the PT server, improved estimation of available system memory and accessing multiple servers from a single client via the ARB RPC system.

3.4 GOS 23S Evaluation

Due to technical limitations of sequence acquisition the SSU has been preferred over the longer LSU in phylogenetic and diversity studies. This chapter investigates the benefits and drawbacks of using either gene in a case study based on the metagenomes obtained in the course of the GOS expedition. The total number of identifiable reads, the taxonomic rank up to which they can be identified and the resulting abundances of major marine bacterial and archaeal taxa are determined for both genes. These results are subsequently compared. Based on the identifiable reads, 16 PCR and 2 FISH probes are also evaluated *in silico* for sensitivity.

Contributions

The publication derived from a student project conceived and supervised by Pelin Yilmaz and myself. This project aimed at comparing the performance of best BLAST and incremental MP identification. One student used the LSU gene and another student the SSU gene to identify the metagenome reads with both methods.

3.5 MIMARKS standard

The MIMARKS standard complements the previous GSC standards MIGS and MIMS with a minimal set of descriptors for marker gene sequences. The also presented MIxS unifies the three standards and allows their extension with minimal descriptors depending on the sampling environment via the MIxS environmental packages.

Contributions

Experiences made during the development of the SILVA database pipeline were contributed to the standardization process. These include the challenges in extracting contextual data at the current state of sequence reporting, data structures preferable to providers of derived databases, descriptors which are typically missing, incomplete or inconsistent in current databases, and observations regarding the interest in specific descriptors available or missing in the SILVA databases.

3.6 Primer Evaluation

The investigation of biodiversity using high-throughput sequencing of PCR amplicons can be biased by a suboptimal choice of primer sequences. In Chapter 10, a large number of commonly used primers is evaluated with regard to their taxonomic coverage. An *in silico* analysis assesses taxon coverage for all investigated primers and investigates a selected set of forward/reverse primer combinations. This analysis relies on the SILVA SSU database and taxonomy. The primers chosen are experimentally evaluated in a comparison of results obtained through PCR amplicon sequencing with results obtained via metagenome sequencing and catalyzed reporter deposition FISH (CARD-FISH).

Contributions

A micro-pipeline was developed by the author for *in silico* assessment of probe sensitivity and specificity for all taxa in any of the taxonomies included in SILVA. The ARB PT server and a tool based on the SINA source code were used for match detection. The nested-set taxonomy representation from the SILVA website was used to calculate results for each taxon from the list of matches. Special consideration was paid to the issues caused by missing sequence data at the primer match position and the use of forward and reverse primers or pairs of forward and reverse primers.

Part II

Results

CHAPTER 4

SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB

Authors: Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig,Jörg Peplies and Frank Oliver Glöckner

Status: Published in Nucleic Acids Research, 2007, Vol. 35, No. 21, pages 7188–7196.

ABSTRACT

Sequencing ribosomal RNA (rRNA) genes is currently the method of choice for phylogenetic reconstruction, nucleic acid based detection and quantification of microbial diversity. The ARB software suite with its corresponding rRNA datasets has been accepted by researchers worldwide as a standard tool for large scale rRNA analysis. However, the rapid increase of publicly available rRNA sequence data has recently hampered the maintenance of comprehensive and curated rRNA knowledge databases. A new system, SILVA (from Latin silva, forest), was implemented to provide a central comprehensive web resource for up to date, quality controlled databases of aligned rRNA sequences from the *B*acteria, Archaea and *E*ukarya domains. All sequences are checked for anomalies, carry a rich set of sequence associated contextual information, have multiple taxonomic classifications, and the latest validly described nomenclature. Furthermore, two precompiled sequence datasets compatible with ARB are offered for download on the SILVA website: (i) the reference (Ref) datasets, comprising only high quality, nearly full length sequences suitable for in-depth phylogenetic analysis and probe design and (ii) the comprehensive Parc datasets with all publicly available rRNA sequences longer than 300 nucleotides suitable for biodiversity analyses. The latest publicly available database release 91 (August 2007) hosts 547 521 sequences split into 461 823 small subunit and 85 689 large subunit rRNAs.

4.1 Introduction

Initiated by the pioneering studies of Fox and Woese [79] 30 years ago and later on pursued by Pace, Olsen, Giovannoni, and Ward [88, 213, 216, 301], the ribosomal RNA (rRNA) molecule has been established as the 'gold-standard' for the investigation of the phylogeny and ecology of microorganisms [6, 215]. Today the more than 500 000 publicly available small and large subunit (SSU and LSU) rRNA sequences ask for specialized quality controlled databases and appropriate software tools. In anticipation of this impending deluge of rRNA data, the development of the ARB software suite and the curation of its associated databases began more than 12 years ago [177]. The software suite offers a graphical user interface and a wide variety of interacting software tools built around a common database. Furthermore, the ARB project provides structured, integrative knowledge databases for small and large subunit rRNAs. Based on regularly offered international workshops and the ARB mailing list it is currently estimated that the ARB software suite and its databases are employed worldwide by several thousand users from academia and industry. In addition to the ARB approach, there are currently three projects offering access to a set of curated ribosomal RNA sequence and alignment databases: the European rRNA databank at the University of Gent (`http://www.psb.ugent.be/rRNA/`) [310] the Ribosomal Database Project II (`http://rdp.cme.msu.edu/`) at Michigan State University in East Lansing, MI [40, 41], and the greengenes project (`http://greengenes.lbl.gov/`) maintained by the Lawrence Berkeley National Laboratory in Berkeley, CA [55]. All four projects offer at least one 16S rRNA dataset, but vary in the amount of sequences, quality checks, alignments, and update procedures. However, the ARB project is the only platform that actively incorporates homologous small (SSU) as well as large (LSU) subunit sequences from all three domains of life, the *B*acteria, Archaea (16S/23S) and *E*ukarya (18S/28S). All projects provide web-based software tools for the alignment and classification of sequences as well as probe match functionalities. Downloading of sequences is provided in various formats including the commonly used FASTA and GenBank file formats. Additionally, greengenes provides ARB compatible datasets, but only for nearly full length sequences (>1250 bases) of *B*acteria and *A*rchaea.

An increasing awareness and motivation to catalogue and protect the biodiversity on Earth using molecular techniques demands comprehensive knowledge databases spanning all three domains of life. Furthermore, a majority of the sequences available is derived from cultivation independent biodiversity surveys, which rely on rapid pattern- or clone-based approaches that often generate partial rRNA sequences. To conserve this suboptimal information especially for diversity studies, state of the art databases need to retain partial

sequences.

To compensate for the limited phylogenetic resolution of the SSU rRNA [175, 221] the two fold larger LSU rRNA should now also be included in the rRNA approach [6]. Especially for Eukaryotes the highly variable regions in the LSU rRNA are already commonly used for species discrimination [309]. Triggered by a new capacity for cheap and rapid sequencing, there is a steady flow of approximately 10 000 rRNA sequences per month into the public databases of the International Nucleotide Sequence Database Collaboration (`http://www.insdc.org`). To make full use of these data for reliable phylogenetic reconstructions and biodiversity analysis careful inspection of each sequence and alignment is necessary. To support the users with this task, standardized procedures to assign a defined set of contextual information to each sequence must be established. Unified quality control mechanisms are urgently needed to intuitively flag potential problems with each sequence. The recent introduction of accelerated and less expensive sequencing technologies, such as pyrosequencing [183], and their successful application for a census of marine microbial diversity [261] further substantiates the need for comprehensive quality controlled databases for comparisons. Such databases provide a stable framework enabling biologists to transfer the copious data into reliable biological knowledge. The SILVA database project is designed to satisfy the request for comprehensive quality controlled and aligned rRNA datasets. It is intended to provide a central knowledge resource to alleviate users of the time consuming manual curation process.

4.2 Materials and Methods

4.2.1 Sequence data

The SILVA release cycle and numbering corresponds to that of the EMBL database, a member of the International Nucleotide Sequence Database Collaboration (`http://www.insdc.org`). Thus, the ribosomal RNA sequences used to build version 91 of the SILVA databases, which is referred to in this paper, were retrieved from release 91 (June 2007) of EMBL. A complex combination of keywords including all permutations of 16S/18S, 23S/28S, SSU, LSU, ribosomal and RNA was used to retrieve a comprehensive subset of all available small and large subunit ribosomal RNA sequences. All candidate rRNA sequences extracted from the EMBL database were stored locally in a relational database system (MySQL). The specificity of the SILVA databases for rRNA is assured by the subsequent processing of the primary sequence information.

The source database providing the seed alignment, required for the incremental alignment process, included a representative set of 51 601 aligned

rRNA sequences from *B*acteria, Archaea and *E*ukarya with 46 000 alignment positions. The SSU alignment positions are currently kept identical with the ssu_jan04.arb database which has officially been released by the ARB project (`http://www.arb-home.de`) in 2004. For the large subunit RNA databases, an in-house, aligned database was used as the seed. It encompasses a representative set of 2868 sequences from all three domains (150 000 alignment positions). Since the quality of the final datasets critically depends on the quality of the seed alignments both datasets were iteratively cross-checked by expert curators during database build-up. Within this process, all sequences that could not be unambiguously aligned were removed from the seed.

4.2.2 Quality checks

Every imported SSU and LSU sequence had to pass a multi-stage quality inspection. Sequences were rejected if they were shorter than 300 unaligned nucleotides, if they were composed of more than 2% of ambiguities or more than 2% homopolymeric stretches longer than four bases, which means only bases exceeding homotetramers are counted, or if they had more than 5% identity to vector sequences. The identity was checked by querying a database of commonly used vector sequences, based on the EMVEC (`http://www.ebi.ac.uk/blastall/vectors.html`) and UniVec (`http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html`) databases using the blastn tool. All thresholds to reject sequences were defined based on statistical analysis of the retrieved SSU and LSU sequences. Each sequence in the SILVA databases carries the percentages of ambiguities, homopolymers, and vector contamination. Additionally, a summary 'sequence quality' score is calculated according to the following formula, where Sq = sequence quality, A = % ambiguities, H = % homopolymers and V = % vector identity:

$$Sq = 1 - \frac{\frac{A}{A_{max}} + \frac{H}{H_{max}} + \frac{V}{V_{max}}}{3} * 100 \tag{4.1}$$

This score represents the mean of the three individual parameters, such that 100 is the best possible value. All sequences that passed the quality thresholds were automatically aligned against the seed alignment using the new SILVA INcremental Aligner (SINA).

4.2.3 Aligner

To cope with the huge amount of sequence information and to minimize the workload for manual curation, a new dynamic incremental profile sequence aligner (SINA) was developed. In the first step the aligner uses the suffix tree

concept of ARB [177] to search for up to 40 closely related sequences in the seed alignment. The reference sequences from the seed are transferred into a partial order graph as used in [155], while preserving the positional identity from the reference alignment. The sequence under investigation is then aligned to this graph using a variant of the Needleman-Wunsch algorithm [206] with affine gap penalties and cost free overhang. The graph concept allows 'jumping' between the different references to find an optimal alignment for the different sequence regions. This technique enables the algorithm to correctly place bases that were e.g. deleted from the closest relative, but are present in the candidate sequence and other relatives. It also eliminates the need for synthetic full length sequences in the reference alignment as introduced for the NAST aligner [54] To further improve the alignment quality a variability statistic is used to give more weight to conserved positions. Results of each step of the aligner are reported to the database and shown in the corresponding fields of the exported ARB file (Tables 4.1 – 4.3). The 'alignment quality' score is a measure of the similarity with the seed sequences that are taken into account for the alignment process. The score is derived from the dynamic programming score by removing the effects of sequence length and positional weighting. High values (>90) mean that nearly identical sequences have been found within the seed alignment, resulting in a high likeliness for the alignment to be accurate. Low values indicate a high distance as perceived by the aligner, making the alignment task more difficult and lowering the average accuracy. Due to the size of the seed alignment, low values are rather rare and ask for manual inspection of the alignment. The 'basepair' score is calculated from the

Table 4.1 *Description of database fields in ARB files exported from SILVA for ARB specific fields and entries.*

ARB Field Name	Owned By	Description
aligned	User	User-defined entry, e.g. name and date of the person who aligned the sequence
ambig	ARB	Ambiguities calculated in ARB using 'count ambiguities'
ARB_color	ARB	Stores the information about sequence colors
name	ARB	Internal ARB database ID, do not change!
nuc	ARB	Number of nucleotides; calculated by ARB using 'count nucleotides'
nuc_term	ARB	Number of nucleotides coding for the respective rRNA gene; calculated by 'count nucleotides gene'
remark	User	Field for remarks
tmp	ARB	Used by several ARB modules

number of bases involved in helix binding according to the secondary structure model of Gutell et al. [99] as already implemented in the ARB package. Canonical and non-canonical base pairings are evaluated, weighted according to the cost model implemented in the Probe_Match ('weighted mismatches') tool in ARB [177]. To fit our unified scoring scheme, the alignment quality and the base pair score were normalized to values between 0 and 100, such that 100 represents the maximum score. After aligning, the number of successfully aligned bases was again counted and sequences with less than 300 bases within the boundaries of the respective SSU or LSU rRNA genes were discarded.

4.2.4 Anomaly check

To check for sequence anomalies, a custom version of the Pintail software [12] was used. The software was specifically adapted for batch processing by the RDP II team. By design, Pintail can only detect anomalies between two sequences. To circumvent this limitation, a pairwise comparison of all sequences in the seed against a group of 20 sequences was performed. If a majority of the comparisons was deemed anomalous, the sequences were iteratively eliminated from the seed alignment until no such sequences remained. Subsequently, all aligned sequences of the SSU database were tested against their five closest relatives within this pruned seed. The number of 'yes', 'likely' and 'no' reported by Pintail was counted for each sequence and transferred into the 'Pintail quality' value. This score was normalized between 0 and 100, such that 100 indicates the best quality and a low probability that the sequence is anomalous or chimeric. Only SSU sequences were checked for anomalies because the Pintail software is currently not designed to handle LSU sequences.

4.2.5 Taxonomy

Every sequence in the SILVA databases carries the EMBL taxonomy assignment. Where available, the greengenes and RDP taxonomies where added for comparison. The EMBL taxonomy was retrieved simultaneously with the sequence, whereas the other taxonomies have been assigned to the sequences based on accession numbers. The greengenes taxonomic outline was acquired in June 2007 from the greengenes website (`http://greengenes.lbl.gov/`) and the RDP Nomenclatural Taxonomy was acquired from RDP II release 9.51. At the moment, no other up to date databases containing aligned LSU sequences are available. Therefore, the only taxonomy provided with the LSU database is the taxonomy used by EMBL. Type strain information has been added to the field 'strain' and is indicated by [T]. Mapping was done based on the RDP II dataset and is therefore only available for *B*acteria.

ARB Field	EMBL Field	Description
acc	AC	Accession number
ali_xx/data	sequence	Sequence information
author	RA	Reference author(s)
clone	FT/clone	Clone from which the sequence was obtained
collected by	FT/collected_by	Name of the person who collected the specimen
collection_date	FT/collection_date	Date that the specimen was collected
country	FT/country	Geographical origin of sequenced sample
date	DT	Entry creation and update date separated by;
description	DE	Description
full_name	OS	Organism species
gene	FT/gene	Symbol of the gene corresponding to a sequence region
insdc	PR	The International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry
isolate	FT/isolate	Individual isolate from which the sequence was obtained
isolation_source	FT/isolation_source	Describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived
journal	RL	Reference location
lat_lon	FT/lat_lon	Geographical coordinates of the location where the specimen was collected
nuc_region	FT source	Identifies the biological source of the specified span of the sequence
nuc_rp	RP	Reference positions
product	FT/product	Name of the product associated with the feature
publication_doi	RX	Cross-reference DOI number
pubmed_id	RX	Cross-reference Pubmed ID
specific_host	FT/specific_host	Natural host from which the sequence was obtained
specimen_voucher	FT/specimen_voucher	An identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution

Table continued on next page

ARB Field	EMBL Field	Description
start	FT rRNA	Start of the ribosomal RNA gene
stop	FT rRNA	Stop of the ribosomal RNA gene
strain	FT/strain	Strain from which the sequence was obtained
submit_author	RL	Submission authors from reference location
submit_date	RL	Submission date from reference location
tax_embl	OC	Organism classification according to EMBL
tax_embl_name	OC	Organism name taken from the classification field
tax_xref_embl	FT/db_xref	Database cross-reference: pointer to related information in another database
title	RT	Reference title
version	ID SV	Subversion from identification line

4.2.6 Nomenclature

All organism names have been synchronized with the 'Nomenclature up to date' website of the "Deutsche Sammlung für Mikroorganismen und Zellkulturen" DSMZ (released June 2007[1]) in order to stay consistent with the constant renaming of validly described species according to the recommendations published in the 'International Journal of Systematic and Evolutionary Microbiology'. All former names are stored in the database and are visible on the web page, as well as in the corresponding field of the ARB databases (Tables 4.1 – 4.3).

4.2.7 SSU and LSU rRNA databases for ARB

Two types of precompiled databases for both small and large subunit ribosomal RNA sequences are available in the ARB format: the high-quality Ref databases and the comprehensive Parc databases. The Ref databases are subsets of Parc, which are exclusively comprised of nearly full length 16S/18S and 23S/28S rRNA sequences. A sequence is accepted if it is at least 1200 bases long. Additionally, sequences as short as 900 bases are included if they belong to the domain *Archaea*. Applying a strict cut-off at 1200 bases would result in the loss of the majority of sequences of this domain. Sequences in the LSU Ref database have to be at least 1900 bases long. For quality control, all sequences that

[1] `http://www.dsmz.de/download/bactnom/names.txt`

could not be unambiguously aligned (alignment quality score <50 and <30 for SSU and LSU, respectively) were removed from the Ref databases. Both Ref databases are supplemented with a guide tree based on the full length sequence tree of the ARB Jan 04 SSU and the Ludwig LSU databases, respectively. The trees were built using the ARB parsimony tool with filters to remove highly variable positions. Common filters like the positional variability filters were recalculated for the Ref databases. Sequences with long branches in combination with low alignment qualities (<80) were removed from the Ref databases.

The rRNA Parc databases are a collection of all quality checked and automatically aligned rRNA sequences longer than 300 bases of the aligned rRNA gene (field 'nuc_gene_slv', Tables 4.1 – 4.3). The name Parc has been chosen according to the UniProt concept [10] where Parc stands for the comprehensive protein sequence archive. All sequences in the SILVA databases are associated with a rich set of sequence and process parameters. Included is information from the initial quality checks to the alignment process, as well as information taken directly from the EMBL entry (Tables 4.1 – 4.3). Together with the search and query functionalities on the web site and in ARB, one can quickly search for problematic sequences or generate individual high or low quality sequence subsets for further processing or curation. The ARB package can be used to export sequences in various formats like EMBL, GenBank, or aligned and unaligned FASTA.

4.2.8 Availability / Webpage

The SILVA databases are available via a web-based interface at `http://www.arb-silva.de`. The web interface is divided into six sections: the browser, search, list, download, background, and FAQs pages. Download of the complete Parc and Ref datasets in ARB format is available in the download section. It is also possible to download files that gain additional sequences from new releases. Subsets of aligned sequences from the Parc dataset can be retrieved from the website. This is currently possible via two entry points: a taxonomic browser and advanced search functions. After selecting a database and the desired taxonomy in the browser, the user can navigate through the taxonomy by clicking on the respective nodes. A cart system is used to easily select subsets of single sequences, complete groups or even phyla. The cart system keeps the selections for the SSU and LSU databases distinct. This allows the user to select sequences from both databases simultaneously without mixing the two sequence types. However, it must be noted that any misclassification or erroneous information provided by INSDC is currently propagated on the SILVA webpage.

Additionally, the advanced search functions of the SILVA website can be

Table 4.3 *Description of database fields in ARB files exported from SILVA for SILVA specific fields and entries.*

ARB Field Name	Description
align_bp_score_slv	Calculates the number of bases in helices in the aligned sequence taken into account canonical and non canonical basepairing. The cost matrix is taken from ARB Probe_Match [177].
align_cutoff_head_slv	Unaligned bases at the beginning of the sequence
align_cutoff_tail_slv	Unaligned bases at the end of the sequence
align_family_slv	Names and scores of reference sequences in the alignment process
align_log_slv	Detailed aligner comments
align_quality_slv	Maximal similarity to reference sequence in the seed
aligned_slv	Data and time of alignment by Silva
ambig_slv	Calculated percent ambiguities in the sequences, a maximum of 2% is allowed
homop_slv	Calculated percentages repetitive bases with more than four bases, a maximum of 2% is allowed
homop_events_slv	Absolute number of repetitive elements with more than four bases
nuc_gene_slv	Aligned bases within gene boundaries
pintail_slv	Information about potential sequence anomalies detected by Pintail [12]; 100 means no anomalies found.
alternative_name_slv	Synonyms or basonyms of the species according to the DSMZ 'nomenclature up to date' catalogue
seq_quality_slv	Summary sequence quality value calculated based on values from vector, ambiguities and homopolymers, 100 means very good
tax_gg	Taxonomy mapped from greengenes
tax_gg_name	Organism name in greengenes
tax_rdp	Nomenclatural taxonomy mapped from RDP II
tax_rdp_name	Organism name in RDP II
vector_slv	Percent vector contamination, a maximum of 5% is allowed

used to build custom subsets of sequences. In addition to simple searches e.g. for accession numbers, organism names, taxonomic entities, or publication DOI/PubMed IDs, complex queries over several database fields using constraints such as sequence length or quality values are possible. The results can be sorted according to accession numbers, organism names, sequence length, sequence and alignment quality and Pintail values. Before download, the search results must be added to the 'List'. This can be done by either manually selecting the sequences by mouse click or by clicking on 'Add complete result to List' to mark and transfer all results.

The coloured bars on the search page and in the short and detailed sequence views of the browser given a fast overview of the different quality aspects assigned to every sequence. The length of the bars is a graphical representation of the respective quality value. The colours classify the information into four categories: A green bar represents a value equal to or greater than 75. Yellow bars stand for values equal to or greater than 50 but less than 75. Values less than 50 are expressed by an orange bar. Red bars are only used for scores of 0. Since 'problematic' sequences, sequences of inadequate quality, as well as insufficiently aligned sequences were discarded from the databases only the Pintail scores can have 0.

In the 'List' section of the website, the entries can be inspected, items can be deleted, and the download files can be created. By clicking on the 'generate download' button the user will be asked whether he would like to download the sequences as a multi-FASTA or ARB file from the download section of the web page. All generated files will be available for download on the download page for up to 24 h. The background section of the website provides additional information about the current status of the databases, and the FAQ section describes the main steps necessary to download subsets of sequences and how to merge the retrieved ARB databases with the user's personal ARB database.

4.2.9 Operating systems and programming languages

The SILVA core system was written in C++ and runs on an Ubuntu GNU/Linux 6.06 LTS based 64bit Dual Dual-Core Opteron cluster with at least 16 GB of main memory on each node. The database server runs MySQL 5.0 and features 32 GB of main memory. The Sun-grid engine (N1GE 6.0) is used to distribute jobs, such as importing, quality check, and aligning on the cluster. The web server is a LAMP system running Ubuntu GNU/Linux 6.06 LTS, Apache 2, MySQL 5.0, and PHP 5. It is connected to the internet via a synchronous 34 Mb connection. The website was written in PHP and Ajax and it is managed using the typo3 content management system in version 4.1. Due to the complexity of the system and the high hardware requirements the system is currently not

Table 4.4 *Sequence retrieval and processing for SILVA 91*

	SSU Parc	LSU Parc
Candidates	900 573	417 217
<300 Bases	320 327	297 218
>2% Ambiguities	8018	2193
>2% Homopolymers	19 240	4772
>5% Vector contamination	14 973	2573
Insufficient relatives	49 063	13 081
<300 Gene bases	25 961	7510
<30 Alignment quality or base pair score	6583	3390
Total sequences in Parcs	461 823	85 689

intended for local installation.

4.3 Results and Discussion

4.3.1 Data retrieval and processing

The total numbers of retrieved sequences and the number of and reasons for rejected sequences are listed in Table 4.4. Cross checks with RDP II and greengenes indicated a sensitivity of our search procedure of >99%. For ambiguities, homopolymers and vector contamination the numbers are non-additive, since some of the sequences may be affected by two or three parameters. Cut-off values have been determined based on a statistical evaluation with relaxed parameters (data not shown), and are intended to balance the quality of the databases with any loss of information. Manual inspection of the sequences rejected by the aligner showed that most of these sequences were not ribosomal RNA sequences.

A comparison of the length distribution immediately after importing the SSU sequences with the length distribution of aligned sequences confirmed that no unexpected loss of sequences in certain length classes occurred (Figure 4.1). Partial sequences between 300 and 800 bases were more frequently rejected than longer ones. A closer comparison of sequence quality versus sequence length confirmed that sequences below 700 bases tend to be of low quality. These 'problematic' sequences may be generated in diversity studies based on single strand sequencing. The high number of rejected sequences with less than 300 bases is evidence for the increase in short length tag sequencing using e.g. pyrosequencing machines. The LSU database shows a similar distribution for rejected sequences as the SSU database (Figure 4.2).

As expected, the SSU sequence length distribution follows the prominent

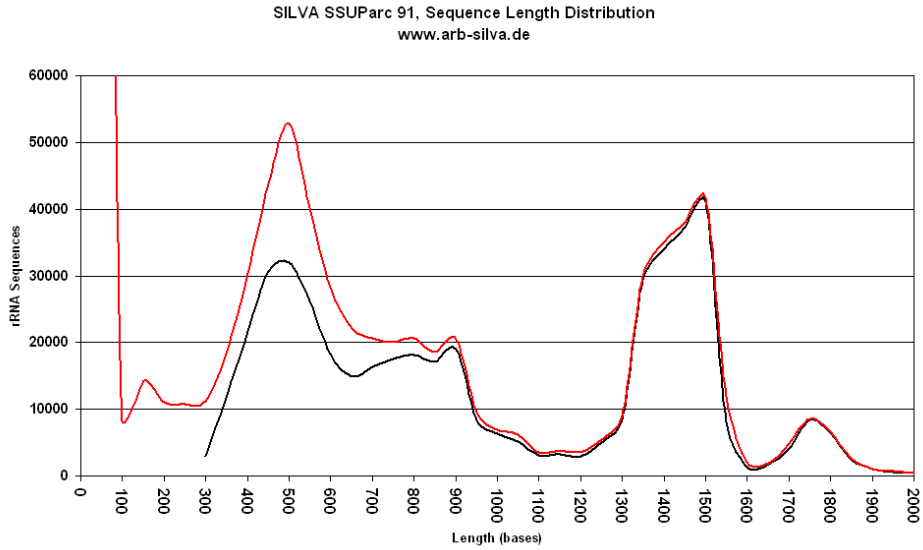


Figure 4.1 *Sequence length distribution of rRNA genes in the SILVA 91 SSU database. The dotted line represents the sequence distribution directly after importing, the solid line after quality checks and alignment. The huge amount of sequences around 100 bases reflect the first impact of tag sequencing approaches.*

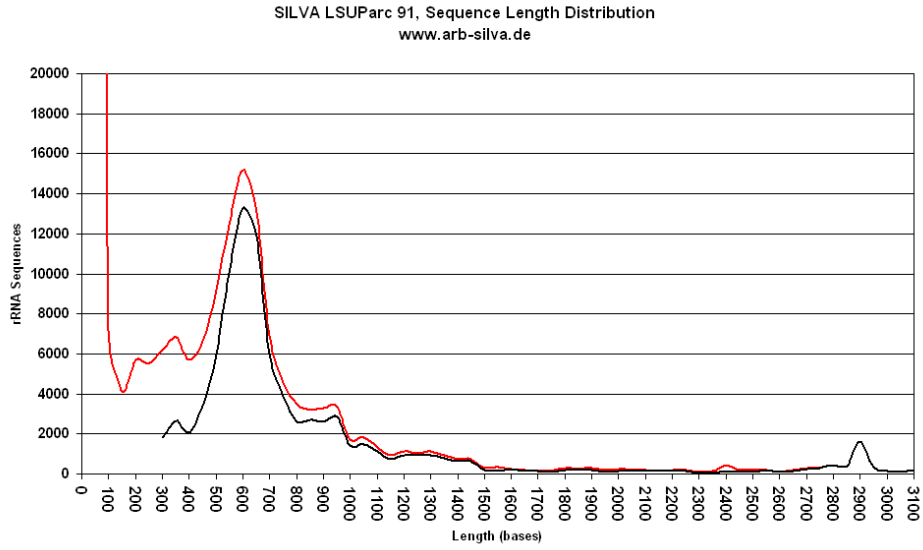


Figure 4.2 *Sequence length distribution in the SILVA 91 LSU database. The dotted line represents the sequence distribution directly after importing, the solid line after quality checks and alignment. The huge amount of sequences around 100 bases reflect the first impact of tag sequencing approaches.*

primer sets used for sequencing specific conserved regions on the 16S/18S rRNA gene [181]. A distinct peak exists around 500 bases, a small one at 900 bases, and a peak between 1300 and 1500 bases. The large number of sequences with 300 and 600 bases is typical for diversity studies that use single reads or fingerprint techniques like DGGE [201]. A text search for 'DGGE' across all fields of the SSU Parc database using ARB showed that 8241 (93%) out of 8889 'DGGE' sequences found belong to the 300 – 600 nucleotide length class. A taxonomic breakdown for the 300 to 600, 600 to 1000, and 1300 to 1600 bases length classes revealed that 80 to 90% of all sequences per class were of bacterial origin. Nevertheless, from the shortest to the longest length class, the relative numbers for *E*ukarya decreases, whereas *A*rchaea and *B*acteria peaked in the 600 – 1000 and 1300 – 1600 length classes, respectively. This again reflects the application of the typical universal primers for *B*acteria [181] and *A*rchaea [52].

A comparison of the number of sequences hosted by the SILVA, greengenes, and RDP II projects revealed that the SILVA SSU Ref database contains roughly the same amount of bacterial and archaeal sequences as greengenes [55] [SILVA: 165 928, greengenes: 165 759 (July 2007)]. Furthermore, SILVA contains 2423 more nearly full length sequences for *B*acteria than RDP II (163 505, release 9.52) [41]. This is surprising considering SILVA's less frequent release cycle (currently synchronized with major EMBL releases); one would thus anticipate SILVA to contain fewer sequences. This may have been due to a higher sensitivity in SILVA's search and alignment protocol. Different quality control mechanisms should not have a significant influence, since only nearly full length sequences have been taken into account for this comparison.

With this respect it has to be emphasised that the primary intention of the SILVA project is not to offer the biggest database by size but more importantly to provide reliable rRNA datasets with a robust set of processing and quality values assigned to each sequence. Such quality values enable users to easily evaluate sequences in order to create subsets of sequences for specific applications, or to extract the sequences that need further attention with respect to sequence and/or alignment quality or anomalies. The alternative taxonomies and type strain information, as well as the latest nomenclature, will facilitate the daily work flow of diversity analysis using classical clone based and high throughput sequencing approaches. Additionally, SILVA provides two LSU databases to support the increasing use of molecular markers with a higher resolution than the SSU rRNA [175]. A taxonomic breakdown of the LSU Parc database contents showed that already 91% of the sequences are of eukaryotic origin.

4.3.2 Alignment and aligner

The current SILVA alignment is based on 46 000 and 150 000 alignment positions for the small and large subunit rRNA, respectively. The reasons for the large amount of alignment positions are: (i) large insertions often present in *E*ukarya and (ii) sequencing errors, such as additional artificial bases often found in homopolymeric sequence stretches. Such errors are common and require placement to be filtered before phylogenetic tree reconstruction, without corrupting the rest of the alignment.

In the 'align-to-seed' approach implemented in the SILVA system, well aligned sequences from seed datasets are used as references for new, unaligned sequences. Therefore, the quality of the final alignment strongly depends on the accuracy of the seed alignment. To further improve the quality of the SSU and LSU seed databases a manual curation process was performed on the latest officially released ARB dataset from January 2004.

The SSU seed hosts currently over 1000 unpublished sequences that primarily cover the domain *A*rchaea. These sequences further improve the alignment in regions of the original SSU January 2004 dataset with sparse sequence coverage. In summary, the quality and consistency of all of the seed alignments is excellent. Only minor inconsistencies could not be resolved in the *E*ukarya. Nevertheless, the Parc datasets exceed the corresponding SSU and LSU seeds by a factor of 8 to 25. This probably indicates that not every phylum is equally represented in the seed. Hence, before using the alignments for in-depth phylogenetic analysis, the alignment of the selected sequence should be carefully checked. Problematic sequences can be easily filtered out by the quality values followed by manual curation. The SILVA team highly appreciates the return of manually inspected and corrected alignments of sequence subsets for inclusion in the SILVA seed. This will allow us to further increase the quality of future alignments.

To manage the deluge of data currently available in the public databases, a new aligner (SINA) has been developed. Similar to existing aligners, such as the Fast Aligner implemented in ARB [177] or the NAST aligner [53], the tool uses related sequences from the reference alignment as a template. For benchmarking the performance of SINA, standard tools, such as BAliBASE [281], could not be used since they are restricted to protein sequences. Benchmark results were obtained by removing and realigning each sequence from the seed. The results were internally compared to the original alignment by counting the number of matching and non-matching columns. Overall, SINA correctly placed 99.8% of all bases in the alignment. Furthermore, 33% and 80% of all sequences tested had no, or less than 1%, alignment errors, respectively. The high accuracy was gained in extensive test runs by changing parameter sets for gap insertions/extension parameters and family sizes combined with sub-

sequent manual inspection of the results by expert curators. The design and implementation of SINA as individually running processes allows distributed aligning on cluster nodes. More than one sequence per second can be aligned per CPU.

4.3.3 Future developments

To account for the growing awareness in ecology that sequence information must be treated in the proper environmental context [75], emphasis was put on the retrieval of contextual (meta)information from public databases. For easy visualisation, the 'Environment' subsection is available in the detailed view of the browser. Additionally, basic environmental parameters, such as exact location and time of sampling as well as physical, chemical, and biological information about the sampling site, will be added in collaboration with the International Census of Marine Microbes (ICoMM), where similar efforts are ongoing (`http://icomm.mbl.edu/`). In upcoming releases of the SILVA databases a crosslink to the genomes mapserver at `http://www.megx.net` [166] will allow the geographic visualization of the sequence information as long as the location is provided. The direct addition of tag sequences below 300 nucleotides as typically produced by pyrosequencing, is not currently planned for SILVA, since it is already a main objective of the ICoMM agenda [261]. Sequence based search options and alignment of user provided sequences are under development for the SILVA webpage. Finally, it must be stressed that an appropriate and stable phylogenetic classification of all rRNA sequences is urgently needed. Efforts in collaboration with Bergey's trust are ongoing and the information will be incorporated as soon as it becomes electronically available.

4.4 Conclusions

The new SILVA system provides comprehensive, quality controlled, richly annotated and aligned, reference rRNA databases to support the molecular assessment of biodiversity, as well as investigations of the evolution of organisms. Applications of the databases range from basic research in microbiology and molecular ecology to the detection of contaminants and pathogens in biotechnology and medicine. Molecular taxonomy and diagnostics have already revolutionized our view on microbial diversity on Earth [117, 218, 261], and the added value of molecular techniques for the determination of eukaryotic diversity has recently been documented by Tautz et al. [275]. The SILVA databases combined with the ARB software suite provide a stable and easy to use workbench for researchers worldwide to perform in depth sequence analysis and phylogenetic reconstructions. It is designed as a knowledge database to assist

in the daily effort to keep pace with the increasing amount of data flooding our general-purpose primary databases.

4.5 Acknowledgments

We would like to thank Ralf Westram for expert assistance with the ARB software suite, the company Pixelmotor for designing and implementing the webpage and all colleagues and students who helped with the manual curation of the databases. We would also thank James Cole, George Garrity and the RDP II team for help with Pintail and fruitful discussions. We are grateful for funding from the Max Planck Society. Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.

CHAPTER 5

SILVA: updates

This chapter outlines the progress of the SILVA project since the original publication presented in the previous chapter.

Section 5.2.5 and the last paragraph of Section 5.3 were published as part of an updated version of the original SILVA publication in "Pruesse, E., Quast, C., Yilmaz, P., Ludwig, W., Peplies, J., and Glöckner, F. O. (2011). SILVA: comprehensive databases for quality checked and aligned ribosomal RNA sequence data compatible with ARB. In de Bruijn, F. J., editor, *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, pages 393–398. John Wiley & Sons." [227]

5.1 Introduction

Since its inception in 2007, the SILVA project has continuously provided regular releases of comprehensive, high quality rRNA databases covering all three domains of life and both small and large subunit genes (SSU and LSU). The demand for such reference databases has not lessened in the past four years. On the contrary, the developments in next generation sequencing technologies, such as pyrosequencing [183], have led to a focus on purely sequencing oriented approaches by studies into microbial diversity [231, 286]. The methods used in these studies for high throughput sequence classification depend on large and diverse sets of accurate training data [303]. Studies such as Grice et al. [95] and Grice et al. [96] have also caused large amounts of high quality, full-length SSU sequences to become available. The sudden increase in data volume and the unbalanced distribution of sequences caused by in-depth sequencing of very specific habitats mandate the development of representative, non-redundant datasets.

In this chapter, we outline the improvements made to the SILVA databases and to the services offered as part of the SILVA website.

5.2 Materials and Methods

5.2.1 Release Schedule

The release schedule of the SILVA databases has been lowered from quarterly to biannual beginning with release 98. The release numbering continues to match that of the EMBL Nucleotide Sequence Database. However, releases of SILVA are only produced for even numbered EMBL releases.

5.2.2 Sequence Data Retrieval and rRNA Extraction

Table 5.1 *Description of database fields in ARB files exported from SILVA for Fields and entries imported from EMBL. Only fields not already described in Table 4.2 are listed.*

ARB Field	EMBL Field	Description
bio_material	RA	Identifier for the biological material from which the nucleic acid sequenced was obtained
clone_lib	FT /clone_lib	Clone library from which the sequence was obtained
culture_collection	FT /culture_collection	Institution code and identifier for the culture from which the nucleic acid sequenced was obtained, with optional collection code
embl_class	EMBL files, relnotes.txt	Describes the data class in EMBL, e.g. CON: Constructed, WGS: Whole Genome Shotgun
embl_division	EMBL files, relnotes.txt	Describes the taxonomic division in EMBL, e.g. ENV: Environmental Samples, PRO: Prokaryotes
env_sample	FT /environmental_sample	Identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE, or other anonymous methods) with no reliable identification of the source organism. Indicated by 'yes' in the ARB files

Table continued on next page

Table 5.1 *Description of database fields in ARB files exported from SILVA for Fields and entries imported from EMBL. Only fields not already described in Table 4.2 are listed.*

ARB Field	EMBL Field	Description
haplotype	FT /haplotype	Name for a specific set of alleles that are linked together on the same physical chromosome.
identified_by	FT /identified_by	Name of the taxonomist who identified the specimen
lab_host	FT /lab_host	Scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained
pcr_primers	FT /PCR_primers	PCR primers that were used to amplify the sequence.
plasmid	FT /plasmid	Name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by /chromosome or /segment.
host	FT /host	Natural host from which the sequence was obtained.
specific_host	FT /specific_host	*removed; now called host*
sub_species	FT /sub_species	Name of sub-species of organism from which sequence was obtained

Since SILVA release 93, the hidden Markov models supplied with the RNAmmer [148] package are used in addition to keyword based matching to find unannotated rRNA gene candidates in the EMBL database. RNAmmer contains twelve models, half of which target the LSU and half the SSU gene. For each gene, a spotter model and a final model is provided for each domain. The spotter model is used to quickly detect regions of interest, the final model is then applied to pinpoint the position of the LSU or SSU gene. Rather than using the Perl script provided with RNAmmer, a custom binary built for speed and a high degree of parallelism is employed in the SILVA pipeline. This allows scanning the whole EMBL sequence archive for gene candidates. An E-value threshold of 10^{-5} is used for all models. Redundant or overlapping regions detected by different models are merged. The sequence is later confined to the actual gene during the alignment stage. Furthermore, all sequences contained in the RDP II database [41] are white-listed as SSU candidate sequences. The annotation sources are documented for each sequence in the field “ann_src_slv”.

Table 5.1 shows the fields imported from EMBL as of SILVA release 108 in addition to the fields that were imported as of release 91 (see Table 4.2

5.2.3 Sequence Alignment

Beginning with release 108, candidate sequences are aligned before passing the quality assurance module. The candidate sequences are aligned in accordance with the SILVA SSU and LSU seed alignments using SINA (Chapter 6). Beginning with release 110, SILVA uses SINA in version 1.2.9. SSU sequences are aligned with the default parameters, LSU sequences are aligned with a full-length value of 2900. The SSU seed alignment has been extended and revised and now contains 57,689 sequences at an alignment width of 50,000 columns. The LSU seed remains unchanged, containing 2,868 sequences at an alignment width of 150,000 columns.

Candidate sequences for which no reference sequence could be found in the seed alignments are rejected by SINA. Reference sequences are searched using a *k*mer search heuristic, a sequence must therefore share at least one 10-mere with at least one of the sequences contained in the seed alignment. The alignment score and the bp-score produced by SINA are used during sequence quality assessment for further filtering. The alignment score reflects the similarity between the candidate sequence and the seed alignment; a score of 1 indicates that the candidate sequence is equal to a reference sequence, part of a reference sequence or can be composed from multiple reference sequences. The bp-score reflects the degree to which the aligned sequence matches the secondary structure of the respective molecule. Candidate sequences detected solely by the RNAmmer models are shortened to contain only the part found to be homologous with the seed sequences. Candidate sequences annotated as SSU or LSU during submission to the INSDC databases are not modified. The non homologous parts at the beginning and end of these sequences are placed in consecutive columns outwards from the outermost aligned bases. The length of the unaligned part is documented in the fields “cutoff_head_slv” and “cutoff_tail_slv” (see Table 4.3).

5.2.4 Quality Checks

The SILVA vector checking was rewritten to account for cyclic vectors and now only considers the part of the candidate sequences that was not found to be part of the respective gene. The vector score is the number of bases, in the unaligned remainder of the sequence annotated as SSU or LSU, found to be part of a vector relative to the number of bases found to be homologous with the reference sequences during alignment. This value is reported in the field

Table 5.2 *Description of database fields in ARB files exported from SILVA for SILVA specific fields and entries. Only fields not already described or described differently in Table 4.3 are listed.*

ARB Field Name	Description
align_family_slv	Shows the accession numbers of the sequences used for alignment
alternative_name_slv	Synonyms or basonyms of the species according to the DSMZ 'nomenclature up to date' catalogue
ann_src_slv	Additional sources of sequence information is indicated in this field. Current identifiers: RNAmmer and RDP
clustered_slv	Members of an OTU (not yet available)
depth_slv	Depth
habitat_slv	Habitat description according to EnvO-Lite
replicates_slv	Replicates in on OTU (not yet available)
tax_slv	SILVA taxonomy path
vector_slv	Percent vector contamination, a maximum of 2% is allowed

"vector_slv".

The calculation of ambiguous base and homopolymer content is confined to the aligned part of the candidate sequences. Both values are computed as a fraction of the length of the aligned part of the candidate sequence. Only homopolymers of at least five bases length are considered and the first four bases of each such homopolymer are ignored. The percentages are reported in the fields "ambig_slv" and "homop_slv", respectively. The field "homop_events_slv" contains the number of homopolymeric stretches (see Tables 4.3 and 5.2).

The SILVA overall sequence quality score Sq is calculated according to the following formula (with $A =$ ambig_slv, $H =$ homop_slv and $A_{\max} = H_{\max} = 2$).

$$Sq = 1 - \frac{\frac{A}{A_{max}} + \frac{H}{H_{max}}}{2} * 100 \tag{5.1}$$

5.2.5 Taxonomy and Type Strain Information

The EMBL taxonomy is retrieved simultaneously with the sequences, whereas the taxonomies from RDP and greengenes are assigned to the sequences based on accession numbers. For LSU rRNA sequences no additional up to date datasets are available. A substantial revision of the classification of all sequences in the Ref datasets was first published with SILVA release 100. Based on the guide trees, all phylogenetic assignments are manually curated, taking into account taxonomic information provided by Bergey's Taxonomic Outline of the Prokaryotes [85], the taxonomic outlines for Volumes 3, 4 and 5 of Bergey's

Manual and the List of Prokaryotic names with Standing in Nomenclature Euzéby [69]. Furthermore, extensive effort is spent to represent prominent uncultured, and not-validly published environmental clades, groups, and taxa, respectively. The majority of these clades and groups are annotated in the guide tree for the SSU Ref dataset based on literature surveys and personal communications. Taxonomic groups consisting only of sequences from uncultured organisms are named after the clone sequence submitted earliest. Due to this exhaustive manual approach SILVA currently contains the most up to date and detailed bacterial and archaeal taxonomic classification. Sequences not classified in a taxonomy are assigned to the pseudo-domain "Unclassified".

5.2.6 Nomenclature and rDNAs from genome projects

Manually curated information about the isolation environment (habitat) of the rRNAs of genome sequences is added based on the EnvO-Lite annotations in the megx.net database [143] (see Table 5.2).

Sequences are marked as originating from type strains, cultured organisms or resulting from genome projects by EMBL using tags added to the "strain" field. This information is retrieved from EMBL, Straininfo.net [294], RDP II [42] and the "All-Species Living Tree" project [312]. Table 5.3 shows which tags are currently in use.

Table 5.3 *Strain Identifiers*

Source	Information	Tag	SSU total	LSU total
EMBL	Typestrains	(t)	135	19
EMBL	Genomes	e[G]	6274	6350
Straininfo.net	Cultured	s[C]	16350	9863
Straininfo.net	Typestrains	s[T]	17492	7719
Living Tree Project	Typestrains (curated)	l[T]	8781	1889
RDP II	Typestrains	r[T]	7839	-

Sequence totals reported as of release 108.

5.2.7 Parc, Ref and RefNR Datasets

The candidate sequences are filtered in three stages to form three baseline datasets meeting the needs of different usage scenarios. The Parc dataset aims at completeness; the Ref dataset includes only high quality full-length sequences from the Parc suitable for tree reconstruction; the RefNR is a pared down subset of the Ref for resource constrained environments. The following paragraphs

describe the criteria used to elect sequences for inclusion with the respective datasets.

Candidate sequences are only excluded from the Parc dataset if a) they contain more than 2% ambiguous bases or more than 2% bases in long homopolymers; if b) the aligned part of the sequence is shorter than 300 bases; or if c) any of the alignment score, bp-score or sequence quality score are below 30%. A blacklist of known defective sequences (i.e. verified chimera) is maintained and its members excluded as well.

In addition to the above criteria, sequences must have an alignment score and a bp-score of at least 50 to be considered for inclusion with the SSU Ref dataset. Furthermore, archaeal SSU sequences must have at least 900 aligned bases, bacterial and eukariotic SSU sequences must have at least 1200 aligned bases and LSU sequences must have at least 1900 aligned bases. The large amounts of high quality full length SSU sequences produced in Grice et al. [95] and Grice et al. [96] were separated from the Ref database and are provided separately as the HSM/MWM (human skin microbiome / mouse wound microbiota) dataset.

Beginning with release 104, a dataset of reduced size is produced from the Ref and HSM/MWM sequences, labeled RefNR. Sequences are selected for inclusion in RefNR based on sequence clustering with UCLUST [65] at an identity threshold of 99%. Sequences are also selected such that all sequences from cultivated species are included in the RefNR.

ARB databases are prepared from the SILVA database contents as described in Chapter 4. The Ref and RefNR ARB databases include a phylogenetic tree, labeled according to the SILVA taxonomy.

5.2.8 Web Tools

The SILVA databases are accessible online at `http://www.arb-silva.de`). Beyond the databases themselves and extensive documentation, the website offers a taxonomy browser, a search tool, an alignment service, a probe evaluation tool and a facility for generating and downloading subsets of the databases. A wealth of additional information about the current status of the databases, as well as FAQs and tutorials are available in the background section of the website. Furthermore, the SILVA website hosts a set of projects including “The All-Species Living Tree” project [312] the “Standard Operating Procedure for Phylogenetic Inference (SOPPI)” [222] and is part of the international Genomic Standards Consortium [74] currently developing the Minimum Information about an MARKer gene Sequence checklist and standard (see Chapter 9).

The metaphor of a “sequence cart” is used to describe a selection of sequences of interest. Cart entries are identified by their accession number. The cart

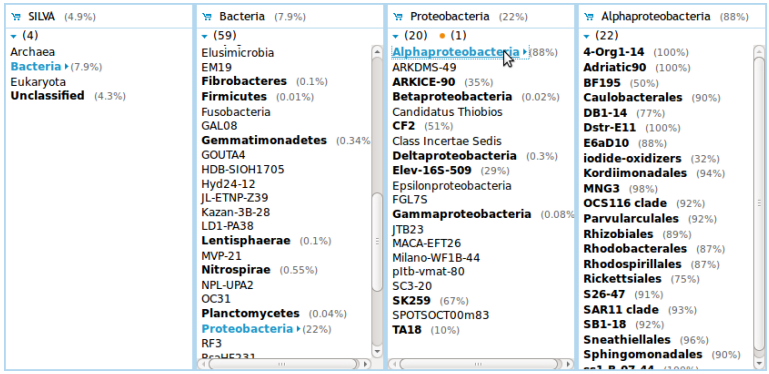


Figure 5.1 *Browser view of the SILVA taxonomy with the taxon* Alphaproteobacteria *opened and all sequences classified as* Alphaproteobacteria *by RDP selected. "Opened" taxa are rendered in blue, taxa containing selected sequences in bold face. The percentages show the coverage of the respective taxon. For example, 88% of all sequences classified as* Alphaproteobacteria *by SILVA are currently in the cart, thus 88% of the sequences classified as* Alphaproteobacteria *by SILVA are classified as such by RDP as well.*

contents can be modified and inspected using the browser and the search. Custom files can be generated from the cart in ARB and FASTA format. Optionally, gap columns or gap-only columns can be filtered from FASTA exports. The cart is also used to identify the target group for probe evaluation (see below).

The taxonomy browser shows the sequences hierarchically according to any of the taxonomies currently included in the respective database. If a taxon contains selected sequence entries, the percentage of selected sequence entries relative to the total number within that taxon is shown along with the absolute counts. Changing the selected taxonomy allows quickly inspecting the level of agreement between taxonomies (see Fig. 5.1). Taxa can be added to the cart and removed from the cart.

The database search allows filtering the database contents according to a number of criteria. Matching is currently offered based on organism name, sets of accession numbers, strain data, publication ID (DOI and Pubmed ID), publication description (title and authors), sequence length, the SILVA quality values (alignment quality, sequence quality, pintail score), taxonomic classification, sequence submission date and genome project ID. The search can be optionally confined to the Ref dataset and/or the contents of the cart. The search results can be added to the cart and removed from the cart. Searching for sequence similar to user sequences is possible using the alignment service.

The SINA (Chapter 6) alignment service allows submitting sequence for alignment in accordance with the SILVA multiple sequence alignments as well as for sequence similarity searches and sequence classification. File sizes are currently limited to 500 sequences of at most 6000 base pairs each. The aligned

sequences can be downloaded in ARB or FASTA format. In the case of FASTA, meta data generated during the alignment process can be exported as brace-enclosed key-value pairs in the FASTA header, as key value-pairs on FASTA comment lines, situated between header and sequence and beginning with a semicolon, or via a separate file in CSV format. The SILVA SEED databases are used as an alignment reference for SINA. The sequence search and classification stages of SINA are also accessible via the web interface. The search result can be confined by a maximum number of best matching sequences as well as a minimal identify with the respective submitted sequence. Sequence comparison is executed based on the alignment. An identity of 1 is defined as all base pairs in the submitted sequence occurring in the database sequence at the same alignment position. The search results can be added to the cart.

The probe evaluation service TestProbe allows assessing probe sensitivity and specificity based on the SILVA databases. TestProbe considers the contents of the cart as the designated target group. It visualizes sensitivity and specificity as a pie charts showing the fractions of matched target sequences and matched non-target sequences (see Table 5.2). Probe matching can be configured to allow none to five mismatches and optionally use weighted mismatches as implemented by the ARB PT server. If probe matching is not limited to the Ref data set, sequences within the target group may not be matched because the region targeted by the probe was not sequenced. TestProbe therefore derives a canonical probe match position as the most frequently matched alignment column and displays the fraction of not matched target sequences having insufficient length along with the sensitivity chart. If a number of allowable mismatches different from none or weighted mismatches are configured, the distribution of the number of mismatches is displayed along with the specificity chart. The thirty most frequently occurring matched sub-sequences are summarized in a table, allowing inspection of typical mismatches. A detailed overview of all matched positions is also displayed as a table and can be downloaded in CSV format. Matched sequences can be downloaded directly or added to the cart.

5.2.9 Languages, Frameworks and Tools

The website uses the Typo3 content management system for static content and secure user sessions, custom PHP scripts for access to the MySQL database hosting the sequence databases and ExtJS widgets and custom JavaScript components to display dynamic content. TestProbe, SINA alignment and custom file export rely on a message passing and job management software implemented in Python. This software interfaces with an internal cluster managed by the Oracle Grid Engine (OGE) in a way similar to current cloud techniques. To ensure

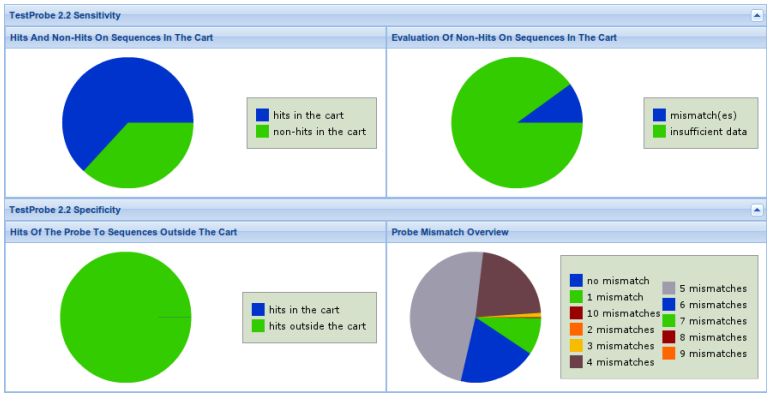


Figure 5.2 *Visualization of probe specificity and sensitivity by TestProbe. The cart contained all sequences classified as* Nevskia *by RDP. The probe evaluated was Nev656 [90].*

timely processing of user requests even under high cluster-load conditions, a dynamically adjusted number or cluster nodes continuously runs the SILVA job shepherd, which in turn schedules the execution of submitted jobs. TestProbe uses SQLite instances of the SILVA databases to allow scalable computation of probe evaluation results without relying on a central database server. The web server uses NGINX to deliver static pages, php5-fpm to schedule PHP script execution and apache2 with mod_phython to schedule Python script execution.

The pipeline modules are implemented in C++ and Python, the pipeline itself is implemented in BASH using the OGE for job execution as well as to maintain the pipeline state in a fashion resilient to failure of cluster nodes.

5.3 Results and Discussion

Assessment of candidate detection sensitivity is, to a degree, possible by comparison of annotation sources. At total SSU database size of 2,492,653 sequences, only 528 candidate sequences were detected solely by the RDP II white list. Of these, only 141 made it past the quality filtering for the Parc dataset. Taking into account that the white list was based on RDP 10.26 and the RDP release 10.27 current at the time of releasing SILVA r108 has grown by 308,116 sequences to include a total of 1,921,179 sequences, we can still conclude that only a minor fraction of sequences is missed during candidate detection. It is interesting to note that a total of 37,608 sequences contained within RDP 10.26 where excluded from the Parc dataset due to quality filtering. Of these, 22,774 can be explained by the lower minimum length requirement (250 instead of 300 bases) imposed by RDP. A further 8,481 were rejected due to high ambiguity content. Candidate sequence detection using RNAmmer yielded a total of only

Table 5.4 *Sequence retrieval and processing, SILVA releases 91 and 108*

	release 108		release 91	
	SSU	**LSU**	**SSU**	**LSU**
Total number of candidate sequences	**4 301 517**	**779 857**	**900 573**	**417 217**
Thereof, detected solely by RNAmmer profiles	42 631	8 191	-	-
Number of sequences rejected from Parc[1]				
<300 bases	-	-	320 327	297 218
not alignable (SINA)	177 668	41 694	49 063	13 081
<300 gene bases	1 291 577	436 016	25 961	7 510
>2% Ambiguities	14 868	4 086	8018	2 193
>2% Homopolymers	6 136	5 605	19 240	4 772
>2% / >5% Vector contamination	2 704	337	14 973	2 573
alignment quality or base pair score <30	21 926	10 739	6 583	3 390
blacklisted	1 651	15	-	-
Total sequences in Parc	**2 492 653**	**269 498**	**461 823**	**85 689**
Number of sequences excluded from Ref				
sequence or alignment quality or base pair score <50	12 467	-	-	-
not "full-length" (900/1200/1900 bp)	1 459 723	245 898	264 933	78 787
HSM project	360 751	-	-	-
MWM project	41 270	-	-	-
Total sequences in Ref	**618 442**	**23 600**	**196 890**	**6 902**

[1] Note that rejected sequence counts are not directly comparable between releases 91 and 108 due to changes in the quality assessment stage (see Section 5.2.4).

339,774 sequences. Of these, 292,334 overlapped with EMBL and/or RDP annotations and were removed. Of the remaining 47,440 candidate sequences, 42,631 passed the quality assurance and were included in the Ref. We therefore conclude that keyword based detection is working extremely well, but that room for improvement exists in sequence based candidate detection.

The peaks of the SSU sequence length distribution follow the prominent primer sets used to sequence specific conserved regions on the 16S/18S rRNA gene [181] (Fig. 5.3). The large number of sequences with 300 and 500 bases is typical for diversity studies that use single reads or fingerprinting techniques. It is interesting to note that up to SILVA release 94, the 500 base peak clearly dominated over the full length sequences. Recent releases show a trend towards the submission of higher quality, nearly full length rRNA sequences.

It has to be emphasized that the primary intention of the SILVA project is to provide reliable rRNA datasets with an informative set of processing and quality values assigned to each sequence. Such quality values enable users to easily evaluate sequences in order to create subsets of sequences for specific applications, or to identify sequences that need further attention with respect to sequence and/or alignment quality or anomalies. The alternative taxonomies and type strain information, as well as the latest nomenclature will facilitate the daily workflow of diversity analysis using classical clone based and high

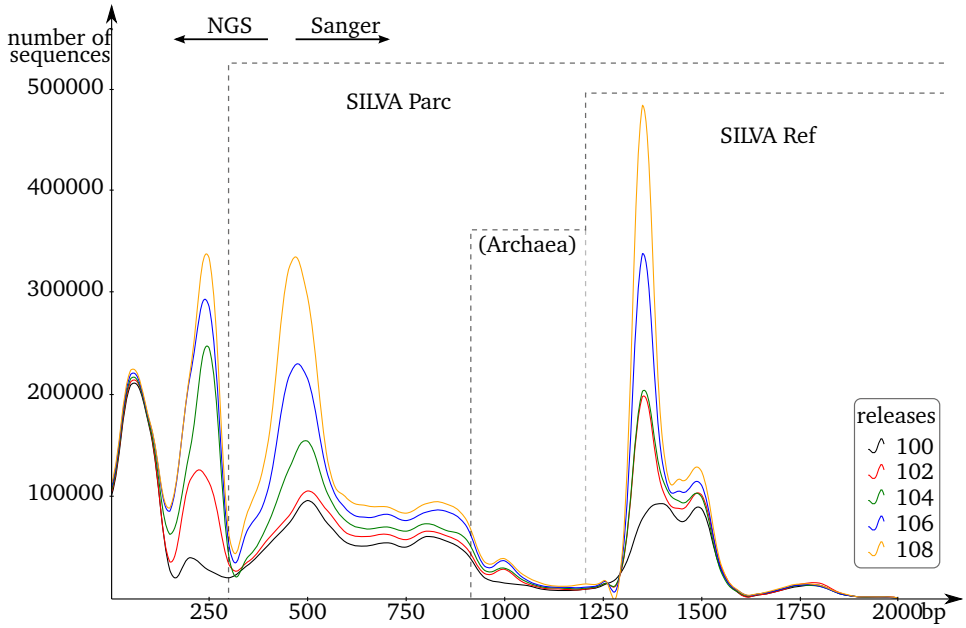


Figure 5.3 *Development of candidate sequence length distribution of rRNA genes in the SILVA SSU databases from release 100 to release 108. Sequences were grouped into buckets of 50bp width. Clearly visible are peaks attributable to the technologies used in large scale sequencing projects.*

throughput sequencing approaches. Additionally, SILVA provides two LSU databases to support the increasing use of molecular markers with a higher resolution than the SSU rRNA [175]. A taxonomic breakdown of the LSU Parc database contents shows that 91% of the sequences are of eukaryotic origin. A closer look indicates that the LSU rRNA is becoming more and more attractive for the molecular identification of e.g. Fungi.

Contributions

The curation of the SILVA taxonomy was done by Pelin Yilmaz. Curated habitat data was provided by Pier Luigi Buttigieg. The integration of typestrain information from RDP II and StrainInfo.net was implemented by Christian Quast and Karin Dietrich. The new download and task management system was implemented by Timmy Schweer. The TestProbe service was implemented initially by Daniel Pletzer and further developed by Timmy Schweer. The revised sequence quality management was developed by Timmy Schweer [251]. The

RefNR datasets were prepared by Jörg Peplies. The first incarnation of the HMM based candidate sequence detection tool was written by Felix Schlesinger in Perl, the second by the author in Python and the third by Arne Böckmann in C++. The cart system, browser and search were written by the author. The students Arne Böckmann, Timmy Schweer and Felix Schlesinger acted under joint supervision of Christian Quast, Frank Oliver Glöckner and the author. Concepts and ideas were a product of joint discussions, making precise individual attributions impossible.

CHAPTER 6

SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes

Authors: Elmar Pruesse, Jörg Peplies and Frank Oliver Glöckner

Status: Published: Bioinformatics, 2012; doi:10.1093/bioinformatics/bts252[1]

ABSTRACT

Motivation: In the analysis of homologous sequences, computation of multiple sequence alignments (MSAs) has become a bottleneck. This is especially troublesome for marker genes like the ribosomal RNA (rRNA) where already millions of sequences are publicly available and individual studies can easily produce hundreds of thousands of new sequences. Methods have been developed to cope with such numbers, but further improvements are needed to meet accuracy requirements.

Results: Here we present the SILVA Incremental Aligner (SINA) used to align the rRNA gene databases provided by the SILVA ribosomal RNA project. SINA employs a combination of k-mer searching and partial order alignment (POA) to maintain very high alignment accuracy while satisfying high throughput performance demands. SINA was evaluated in comparison with the commonly used high throughput MSA programs PyNAST and mothur. The three BRAliBase III benchmark MSAs could be reproduced with 99.3%, 97.6% and 96.1% accuracy. A larger benchmark MSA comprising 38,772 sequences could be reproduced with 98.9% and 99.3% accuracy using reference MSAs comprising 1000 and 5000 sequences. SINA was able to achieve higher accuracy than PyNAST and mothur in all performed benchmarks.

Availability: Alignment of up to 500 sequences using the latest SILVA SSU/LSU Ref data-sets as reference MSA is offered at http://www.arb-silva.de/aligner. This page also links to Linux binaries, user manual and tutorial. SINA is made available under a personal use license.

[1]Chapter updated to puduring proof

6.1 Introduction

Multiple sequence alignment (MSA) is a core building block in the analysis of biological sequence data. Phylogenetic tree reconstruction, structure prediction or hidden Markov modeling require multiple sequence alignment to infer residue-level homology or structural or functional identity. The ubiquitous need for MSA computation has made this field an active research topic with over 100 methods published in the past 30 years and numerous review papers discussing their relative merits and deficiencies [134, 209, 220].

The dependency of the subsequent analysis methods on the results of the MSA stage and the drastic effect differing MSAs can cause [169, 197] make alignment accuracy the primary benchmark for novel and improved methods. The task of computing the optimal alignment (as determined by the Sum-of Pairs (SP) score) was shown to be non-deterministic polynomial (NP)-complete [299], and is therefore only feasible for very few sequences. For sets of sequences comprising several thousand or more sequences heuristic algorithms are used. The most prevalent algorithms are based on the progressive alignment [73] technique, which builds the MSA via a series of pairwise alignments of sequences and partial alignments along the branches of a guide tree.

Sequence data volumes are growing exponentially. This was already observed almost twenty years ago [234] and the effect has not diminished since [157]. Multiple sequence alignment has long been largely unaffected, because the numbers in which homologous gene sequences were available remained low. For many genes, however, this situation is changing. Especially frequently sequenced marker genes, such as the ribosomal RNA, are rapidly becoming available in volumes exceeding the scalability of traditional alignment techniques. In 2007, the first release of the SILVA SSU database contained over 353,366 small subunit rRNA (SSU) gene sequences [226]. Until September 2011, that database grew more than sevenfold to contain 2,494,582 sequences. The two other large rRNA databases, greengenes [55] and RDP [42], are of similar size [7]. The large subunit rRNA (LSU), provided only by SILVA, grew only slightly slower: In 2007, the database contained 46,979 sequences. Currently, it contains almost six times as many sequences (269,498).

While each of these databases uses a different tool to compute their alignments, the employed methods share one important characteristic: Rather than computing an alignment *de novo*, the alignment of each individual sequence is derived from a static reference MSA. The reference MSA implicitly defines a fixed set of alignment columns into which the bases comprising the query sequence are placed. By avoiding mutual comparisons between the sequences considered for inclusion in the final MSA (candidate sequences), the alignment process becomes inherently scalable. Furthermore, the MSA offered by

the database provider can be easily extended by database users in the same manner in which the MSA was originally constructed. This, in turn, allows employing established alignment-based methods to analyze even large-volume NGS data-sets.

The MSA provided by RDP II is computed using Infernal, which implements a model based approach using a special form of stochastic context free grammar (SCFG) termed covariance models (CM). These are similar to Hidden Markov Models (HMM) but are able to capture the co-variations caused by the highly conserved secondary structure of rRNAs [204, 205]. The Infernal model used by RDP II is computed from a set of several hundred carefully chosen sequences which were manually aligned to match the well-known secondary structure of the 16S rRNA. The nearest alignment space termination (NAST) method DeSantis et al. [54] employed by greengenes uses BLAST [2] to obtain a pairwise alignment between the candidate sequence and the best match in the reference MSA. The alignment is then used to map the candidate sequence into the reference MSA via a series of gap character reintroduction and removal operations. Improved implementations of the same principle have been published as PyNAST [33] and as part of mothur [249]. PyNAST uses UCLUST [65] instead of BLAST whereas mothur relies on its own implementations of a *k*-mer search to select the reference sequence and a Needleman-Wunsch type alignment algorithm to perform the pairwise alignment.

Here we describe the SILVA Incremental Aligner (SINA) which is part of the rRNA gene processing pipeline of the SILVA ribosomal databases project.

6.2 Algorithm

Our algorithm is based on the assumption that the the sequences contained in the reference MSA are more likely to have a sibling relationship with the candidate sequence than to be direct ancestors or descendants. Because each sibling will have diverged differently from the common ancestor, some parts of the candidate sequence may be resembled most closely by one of the siblings while other parts are more similar to different siblings. Instead of seeking the optimal alignment with a single, best reference sequence (as is done by NAST) or optimizing the SP-score between the candidate and all of its siblings, we attempt to align each part of the candidate with the most similar counterpart found in any sibling. In order to prevent arbitrary alignment in hypervariable regions, we further demand that consecutive “parts” must be joined by at least one mutually aligned, identical base.

The optimal sequence of parts and the optimal alignment of the candidate with these parts can be found at the same time using dynamic programming. The algorithm employed by SINA for this purpose is essentially equivalent to

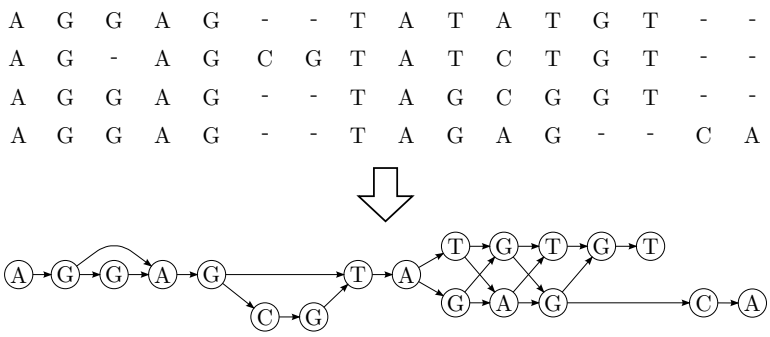


Figure 6.1 *The alignment of the selected reference sequences is converted from RC-MSA representation (top) to PO-MSA representation (bottom).*

partial order alignment (POA) as described in Lee et al. [155]. The reference MSA is reduced to a directed acyclic graph (DAG) as shown in Figure 6.1. Each node of the graph represents an evolutionarily unique base. That is, all identical bases sharing a column in the reference MSA are coalesced into one node. Gaps and the order of bases are represented by the graph topology: Two nodes are connected exactly if there is a sequence in which the two bases they represent occur consecutively. Thus, there is exactly one path through the graph for each combination of "parts" as defined above. By applying a Needleman-Wunsch [206] modified to allow a DAG along one axis we obtain the least costly alignment of the candidate as well as the corresponding path.

The time and space complexity of the alignment stage is decoupled from the size of the entire reference MSA by prefixing a sequence selection stage. This stage chooses a small set of sequences from the results of a heuristic similarity search. The DAG used as alignment template is constructed from these sequences only.

The fixed-column constraint necessary to allow concatenation of the individually aligned sequences into a joint MSA is maintained during DP alignment using a further modification of the Needleman-Wunsch algorithm.

6.2.1 Reference Sequence Selection

The sequences to be used in building the alignment template are assembled from the result of a k-mer sequence search on the reference MSA. SINA does not implement this search itself but utilizes a component from the ARB software package called the PT server [177]. The PT server offers several parameters to configure the k-mer search, all of which are exposed by the SINA command line interface. These parameters are: 1) the value of k, 2) a number of allowable mismatches at arbitrary positions within each k-mer, 3) a range of alignment columns to which the search for shared k-mers is restricted, 4) a fast

mode which searches only for k-mers beginning with "A", 5) a "non-relative" mode which computes the fractional k-mer count by dividing the number of shared k-mers by the query length rather than by the minimum of the lengths of query and matched sequence.

Based on the findings in Edgar [63], we apply a logarithmic transformation to obtain a measure in approximately linear relationship with fractional identity. Here F is the fractional k-mer count, L_q the length of the query sequence and Y the obtained measure.

$$Y = 1 - \frac{log\frac{F+1}{L_q}}{log\frac{1}{L_q}} \tag{6.1}$$

After executing the search, SINA iterates through the matches in order of descending identity and decides according to the following rules and parameters which sequences are to be kept and passed into the alignment template construction stage. 1) The first *fs-min* sequences are always kept. 2) Up to *fs-max* sequences are kept if their similarity to the candidate is at least *fs-msc*. 3) Further sequences of at least *fs-full-len* bases length are kept independent of their match score until the set of selected sequences contains at least *fs-req-full* such sequences. 4) Further sequences are kept if they cover the start and end of the gene as determined by the alignment positions *gene-start* and *gene-end* until at least *fs-cover-gene* such sequences have been found. The latter two rules are designed to ensure that the outer edges of the alignment are covered even if the reference alignment contains partial sequences.

As a performance optimization, the candidate sequence is compared to all sequences in the reference set. If it is found to be contained in one of them, the candidate sequence is aligned by simply copying the matching part of the alignment of the reference sequence. An explaining remark is made in the log and the remaining alignment stages are skipped.

6.2.2 Construction of Alignment Template

We use a directed acyclic graph (DAG) to represent the selected set of aligned reference sequences. The nodes of this graph correspond to unique base-column combinations in the reference sequences. The nodes are linked by edges if the corresponding bases occur consecutively in any of the reference sequences (see figure 6.1). Consider the aligned reference sequences as lists of base-column pairs. Then, for each such sequence, there is a path in the graph comprising an equivalent list of nodes. This type of graph is described as "partial order MSA" (PO-MSA) by Lee et al. [155]. The term expresses that the structure itself only imposes a partial order on the bases comprising the ali-

gnment, whereas the traditional "row-column MSA" (RC-MSA) representation imposes a total order. When storing a list of sequence identifiers with each edge, exact conversion between the two representations is possible.

Our method of constructing a PO-MSA from a RC-MSA and the data stored within the nodes differs slightly from the method described in Lee et al. [155]. We preserve the frequency of the represented base in its column to be used as a weight during the alignment process. Also, we do not construct the PO-MSA by iteratively adding sequences and merging those nodes that represent homologous bases. Instead, we use a scan-line algorithm passing horizontally through the input RC-MSA: For each sequence S_i in the RC-MSA the last created node N_i is remembered. We then pass through all alignment columns j. In each column, one node is created for each non gap character encountered. For each sequence S_k in which the character was encountered, an edge from the last remembered node N_k is created to the new node and the new node is remembered as N_k. After all columns have been processed, duplicate edges are removed.

6.2.3 Dynamic Programming Alignment

In order to align a candidate sequence with an alignment template in PO-MSA format, we extend the dynamic programming recursion from the Needleman-Wunsch algorithm. Our extension is similar to that employed by POA. In Needleman-Wunsch and its derivative algorithms, two sequences A and B are aligned by computing a matrix H such that the value of $H_{i,j}$ is the optimal score for the alignment of the prefixes $A_1 \ldots A_i$ and $B_1 \ldots B_j$ of lengths i and j of the sequences A and B. The value of each cell $H_{i,j}$ is defined as a function of the scores of the three prefix pairs where either one or both of the prefixes is one item shorter. Given a function $S(i,j)$ defining the matching score for A_i and B_j and using g as the score for a gap, we have:

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + S(i,j) \\ H_{i,j-1} + g \\ H_{i-1,j} + g \end{cases} \tag{6.2}$$

This recursion is generalized to allow using a PO-MSA instead of one of the sequences by replacing the notion of "prefix of length i" with "path leading up to node A_i". Leaving B as a sequence, $H_{i,j}$ then becomes the optimal score of the alignment of the prefix of B of length j with any path in A leading to A_i. Using $A_p \rightarrow A_i$ to denote that an edge from A_p to A_i exists, we arrive at:

$$H_{i,j} = \max_{p:A_p \to A_i} \begin{cases} H_{p,j-1} + S(i,j) \\ H_{i,j-1} + g \\ H_{p,j} + g \end{cases} \tag{6.3}$$

Affine Gap Penalties

To support affine gap penalties of the form $g_k = g_{open} + (k-1)g_{extend}$, SINA uses a further extension of this induction, modified in the same way as was shown by Gotoh for the original induction [92]:

$$P_{i,j} = \max_{p:A_p \to A_i} \begin{cases} H_{p,j} + g_{open} \\ P_{p,j} + g_{extend} \end{cases} \tag{6.4}$$

$$Q_{i,j} = \max \begin{cases} H_{i,j-1} + g_{open} \\ Q_{i,j-1} + g_{extend} \end{cases} \tag{6.5}$$

$$H_{i,j} = \max_{p:A_p \to A_i} \begin{cases} H_{p,j-1} + S(i,j) \\ P_{i,j} \\ Q_{i,j} \end{cases} \tag{6.6}$$

6.2.4 Scoring

Although SINA supports the use of arbitrary substitution matrices to define $S(i,j)$, the default is to use 2 as the score for matching bases and -1 for mismatching bases. IUPAC encoded ambiguities are treated as a match if a match is conceivable (i.e. "N" matches anything).

SINA also implements two methods for weighting $S(i,j)$ according to the variability in the reference MSA: 1) The score is multiplied with the frequency with which the base A_i occurs among the selected reference sequences in column i according to a configurable scaling factor. 2) The score is multiplied with a per-column conservation indicator derived from a conservation profile computed within ARB ("positional variability by parsimony (PVP)", see supplementary materials). After POA sequence alignment, the total score is normalized via division by the sum of the weighted rewards for a match in each template column contributing to the alignment.

6.2.5 Treatment of Sequence Ends

SINA uses what is sometimes referred to as "overlap" alignment. While global alignment allows no unaligned sequence tails and local alignment allows both sequences to have unaligned tails, overlap alignment allows only one unaligned tail at either end. At both ends, either the candidate sequence or the template is aligned until its last base. The cost-free terminal gap is achieved by initializing $H_{0,j}$ and $H_{i,0}$ with 0 and choosing the best scoring cell $H_{i,j}$ where at least A_i or B_j has no successor to start the backtracking through the alignment matrix.

Three policies are provided for dealing with the unaligned sequence tails: 1) The unaligned bases may be omitted from the final alignment. 2) The unaligned bases may be placed consecutively following the outermost aligned base. 3) The unaligned bases may be placed at the out-most columns of the MSA.

6.2.6 Treatment of Insertions

The alignment of the candidate with the PO-MSA yields column positions only for substitution events (matches and mismatches). While deletions in the candidate with respect to the reference sequences pose no problem, appropriate column positions must be determined for inserted bases. If the number of alignment positions between the two bases enclosing an insertion, that is the size of the gap in the reference alignment, is larger than the insertion, the insertion is placed right-bound in this gap. SINA offers three choices for dealing with insertions that cannot be accommodated by the reference MSA: 1) The insertion may be shortened as required by erasing bases. 2) The bases surrounding the insertion may be shifted outwards. 3) A modified DP algorithm may be used that disallows insertions not mappable to the reference MSA.

Our base shifting algorithm is a greedy search for free alignment positions to the left and right of the insertion which we believe to be equivalent to nearest alignment space termination (NAST). If the gap closest to the insertion is of insufficient size, the bases between this gap and the original insertion are included in the insertion and the process repeated until the insertion can be placed.

As an alternative option, we further extended the DP alignment to observe constrained alignment space by only considering gap open and gap extension events that can be accommodated by the reference MSA. For a node A_i in the template DAG, the amount of free columns f_i to the right of it is defined as the difference between its alignment position and the lowest alignment position of its immediate successor nodes minus one. Ignoring gap extension, the induction defining H becomes:

$$H_{i,j} = \max_{p:A_p \to A_i} \begin{cases} H_{p,j-1} + S(i,j) \\ H_{i,j-1} + g \quad \text{if } f_i > 0 \\ H_{p,j} + g \end{cases} \tag{6.7}$$

Note that this is equivalent to using a cost function for gaps which assigns an infinite penalty for inserting a gap into the reference alignment. However, the Gotoh optimization for DP alignment with affine gap penalties requires the cost for extending gaps to be monotonically decreasing [92]. Nonetheless, we have implemented an analogous extension, aware that the induction we use constitutes a loss of optimality where alignment space is insufficient. $F_{i,j}$ is set to f_i when $Q_{i,j}$ is based on a gap open event and set to $F_{i,j-1} - 1$ if $Q_{i,j}$ is based on a gap extension.

$$Q_{i,j} = \max \begin{cases} H_{i,j-1} + g_{open} \quad \text{if } f_i > 0 \\ Q_{i,j-1} + g_{extend} \quad \text{if } F_{i,j} > 0 \end{cases} \tag{6.8}$$

$$H_{i,j} = \max_{p:A_p \to A_i} \begin{cases} H_{p,j-1} + S(i,j) \\ P_{i,j} \\ Q_{i,j} \quad \text{if } f_i > 0 \end{cases} \tag{6.9}$$

6.3 Implementation

SINA has been implemented in C++ making heavy use of generic programming techniques. External components employed include several BOOST libraries, the ARB database library and the ARB PT server. ARB and FASTA formats are supported for sequence input and output. Per sequence meta data can be exported via ARB database fields, FASTA headers, FASTA comments or a separate CSV file. The reference MSA must be in ARB format. Conversion of reference alignments from FASTA to ARB format is possible with SINA

6.3.1 Reverse Complement Detection

If instructed, SINA will execute the k-mer search multiple times using the reversed and/or complemented candidate sequence. If an orientation different to the original yields a better best scoring match, the candidate is transformed accordingly.

6.3.2 Sequence Search and Classification

We also implemented a simple search and classify stage. The search uses the alignment (as computed by SINA or by an external tool) to quickly determine fractional identities. Both an exhaustive search and a quick search considering only the best matches from a *k*-mer search can be performed.

Also, a least common ancestor (LCA) classification can be performed if the searched database contains taxonomy data in materialized path format. LCA classification can be relaxed to allow a percentage of outliers.

6.3.3 Visualization of Alignment Differences

Manual inspection of the alignment differences (resulting for example from different tools, changed parameters or modifications to the reference MSA) is supported via a differencing function. This function prints a colored RC-MSA representation of the sections of the alignment in which the reference alignment and the alignment to be inspected differ. Columns containing only gap characters are removed from this view. The reference sequences used to construct the PO-MSA template are listed together with the new and the original alignment. If the SINA alignment stage was bypassed, the SINA search stage can be used to select suitable sequences for display in combination with the two different alignments of the candidate. Rows are consolidated such that only unique alignments remain.

6.3.4 Parameter Tuning

The default parameter settings in SINA were tuned for the alignment of SSU rRNA gene sequences. In order to simplify determining correct parameters for other genes, SINA offers automated evaluation of alignment accuracy using a leave-query-out approach. In this mode, each sequence in the reference alignment is newly aligned (excluding the sequence itself from the set of selected reference sequences), the result compared to the original alignment and the average scores reported. Alignment parameters such as match and mismatch scores, gap penalties or *k*-mer length can then be adjusted to maximize this score.

In order to simulate more difficult alignment cases where the candidate sequence is distant to the closest match in the reference MSA, reference sequence selection may be constrained using a maximum identity parameter. The identity of each sequence considered during reference sequence selection with the candidate sequence is computed using their original alignments. Sequences with an identity higher than the configured threshold are discarded and not included in computing the alignment template .

6.4 Evaluation of SINA

MSA computation methods are generally validated by quantifying their ability to accurately reproduce benchmark MSAs known to be of high quality. The degree to which a tool was able to reproduce the benchmark MSA is measured by determining the fraction of exactly reproduced alignment columns (CS score [282]) and the fraction of correctly aligned residue pairs (Q score [64], also called SP-score [283]). This measure was used in the evaluation of SINA. However, we expect significantly higher scores than commonly achieved by *de novo* methods (see Discussion).

For evaluation, we used the three MSAs provided with BRAliBase III (5S rRNA, tRNA and U5) and the manually aligned subsets of the MSAs provided by SILVA (SSU and LSU). The SILVA alignments where chosen because they are the largest manually created alignments available to us. The BRALiBase alignments were chosen to complement the SILVA alignments with test data from a source not affiliated in any way with the authors of this paper. The SSU and LSU test data was generated by excluding all sequences in the SILVA databases that were themselves aligned by SINA, leaving only manually aligned sequences from the SILVA seed. This test data is equal to the published subsets of the SILVA seed alignments. The SILVA seed alignments are based on alignments published by the ARB project in 2004. During construction and maintenance of the SILVA seed, sequences were removed if they could not be aligned unambiguously and new sequences added to enhance phylogenetic coverage. All sequences in the seed (and therefore in the test data) were aligned manually by rRNA alignment experts. The alignment itself is guided strongly by the secondary and tertiary structure of the respective rRNA. The SSU and LSU test data are made available at ftp.arb-silva.de/SINA/test_data/.

We compared SINA to the NAST implementations by mothur and PyNAST. The align.seqs command from mothur (version 1.19.1) was used with default parameters. PyNAST (version 1.1, UCLUST version v1.2.exportedq, cogent version 1.5.0) was used with identity threshold below which it refuses alignment lowered to 0.0001. Minimal reference sequence length was set to 50 for SINA and PyNAST. SINA (version 1.2.8) was also configured with appropriate values for full-length sizes (5S rRNA: 120, tRNA: 80, U5: 80, SSU: 1400, LSU: 2900). The k-mer size used by SINA was lowered to eight for the tRNA and U5.

Three different benchmarks were performed, one using the four smaller MSAs and two using the large SSU MSA. The three benchmarks differ in the way the benchmark MSA is split into the set of sequences to be used as a reference MSA and the set to be used for measuring the alignment accuracy. Since all three tools expect sizable reference MSAs, the benchmark based on the four smaller MSAs follows a “leave-query-out” scheme: Every sequence

in the benchmark MSA is aligned using all other sequences as reference MSA (benchmark 1). The SSU MSA is large enough to create reference MSAs of different size by randomly sampling sequences. Sampling was repeated 100 times, once using 1000 sequences and once using 5000 sequence. Candidate sequence sets of equal size were sampled from the remaining sequences.

The typical identity between each candidate and its best matching reference sequence remains very high, even when sampling a reference MSA of only 1000 sequences. In order to obtain more difficult test cases having lower rates of identity, we constrained the reference sequence selection algorithm to exclude sequences above a cut-off value (see 6.3.4). Using 21 cut-off values between 50% and 100% at 2.5% intervals (100% being equivalent to leave-query-out benchmarking), we examined the accuracy in relation to the identity of the candidate with the reference. This benchmark was repeated for numerous sets of parameter settings and also used for parameter optimization (benchmark 2).

Lastly, we repeated benchmark 2 with an alternative alignment template implementation relying on column profiles rather than a PO-MSA for comparison. All other settings including the selection of reference sequences remained identical to the original benchmark 2.

6.5 Results

Table 6.1 shows that SINA performed better than both mothur and PyNAST for all MSAs used in the leave-query-out benchmarks. Friedman rank tests using the results for each sequence as blocks showed significant P-values ($<2*10^{-5}$) for all pairs of tools in each data-set except PyNAST vs mothur in the U5 data-set (0.55).

Table 6.2 shows the results for the benchmarks using candidate sequences and reference MSAs sampled from the SSU data-set. We show the average Q scores from all successfully aligned sequences, although this slightly inflates the scores for PyNAST which failed to align all candidate sequences. Lowering the identity threshold below which alignment is refused by PyNAST to 0.0001 reduced the number of failed alignments. However, of the 100,000 sequences aligned using 1k reference MSAs, PyNAST still failed to align 547. Of the 500,000 sequences aligned using 5k reference MSAs PyNAST failed to align 2750 sequences. The average Q scores achieved by mothur for the sequences refused by PyNAST were 91.36% and 94.8%, respectively. The average Q scores achieved by SINA for these sequences were 97.45% and 98.46%.

In addition to the average Q scores we show the standard deviation between averages computed for each of the 100 samples. The variance between tests is much lower than the differences between tools, indicating that the

Table 6.1 *Results from Leave-Query-Out benchmarks*

data-set	5S rRNA	tRNA	U5	SILVA LSU
sequences	597	1113	232	1588
PyNAST	98.6%	96.4%	94.0%	98.9%
mothur	97.5%	92.1%	93.3%	98.9%
SINA	99.3%	97.6%	96.1%	99.2%

The reported percentages are the average Q scores. Only sequences aligned by all three tools where considered.

Table 6.2 *Results using test data sampled from the SILVA SSU data-set*

	all SSU samples		< 80% identity	
reference size	1000	5000	1000	5000
sequences	100,000	500,000	5443	8811
mean identity	92.34%	95.24%	75.71%	75.9%
(PyNAST[1])	96.7%	97.6%	90%	89%
	(0.20%)	(0.08%)	(1.7%)	(1.5%)
mothur	96.6%	97.8%	88%	88%
	(0.23%)	(0.07%)	(2.0%)	(1.3%)
SINA	98.9%	99.3%	94%	93%
	(0.12%)	(0.03%)	(1.2%)	(1.1%)

The average Q scores shown were obtained by randomly sampling sequences from the SILVA SSU based test data to create 100 reference MSAs and benchmark sets. This was repeated once with a reference MSA size of 1000and once with a size of 5000. The standard deviation between Q score averages from each of the 100 reference MSAs is shown in parentheses. The two columns on the right show the results when considering only difficult cases where the candidate sequences have less than 80% identity with all sequences in the respective reference MSA.

[1]PyNAST failed to align 0.5% of the sequences.

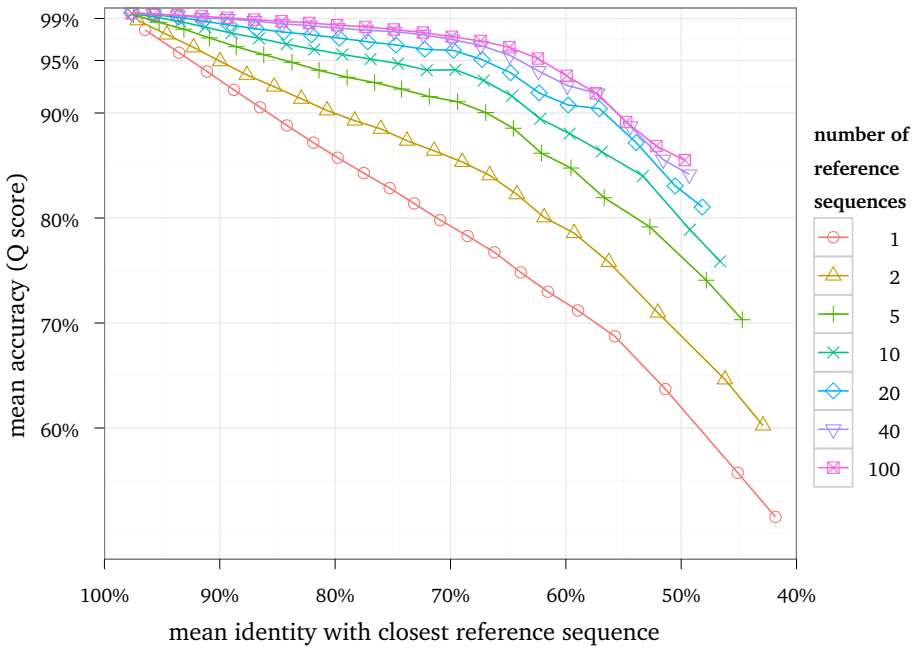


Figure 6.2 *Alignment accuracy decreases almost linearly with the shared fractional identity of candidate and reference when using one reference sequence (red line). Using larger numbers of reference sequences markedly increases accuracy.*

reported Q score averages are sufficiently robust for comparing the tools. Pearson rank tests using the per sample averages as blocks showed P-values below $2*10^{-5}$ for all pairs of tools except PyNAST vs. mothur in the 1k reference MSA benchmarks.

The second benchmark showed marked differences in alignment accuracy for varying reference sequence set sizes. The average Q scores rises over all identity thresholds with each increase in the number of reference sequences used. Above 40 sequences, the effect tapers off (figure 6.2, supplementary figure S1). The same can be observed for the average fraction of bases that were part of an insertion with respect to the template PO-MSA (figure S2). Configuring SINA to use a column profile as alignment template yielded lower accuracy (figure 6.3). Especially when candidate and reference sequences share a lower fractional identity, alignment accuracy drops significantly. Increasing the reference set size beyond five had a detrimental effect.

At a reference set size of 40 sequences, and match/mismatch scores of 2 and -1 (figure S3), a gap open penalty of 5 and a gap extension penalty of 2 was found to perform best (figure S4). Enforcing the inclusion of at least one sequence of at least 1400 bases in the reference set improved results at

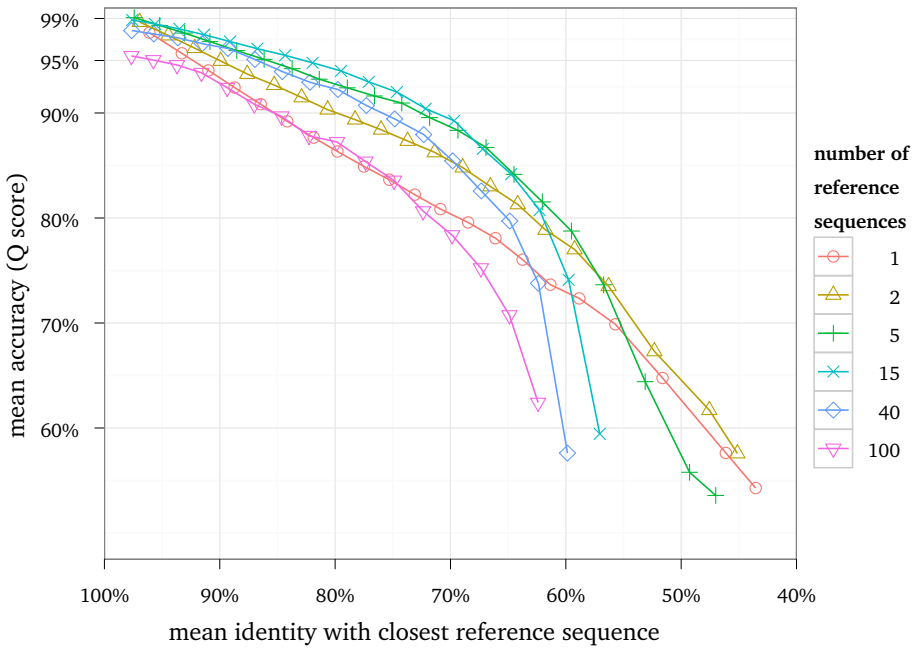


Figure 6.3 *An alternative implementation which used simple column-profiles built from the selected reference sequences showed overall lower accuracy. Increasing the number of reference sequences quickly led to a degradation in accuracy.*

identity thresholds lower than 0.9 visibly (figure S5). Using the modified DP algorithm to maintain fixed columns gave a slight improvement over using base shifting (figure S6). Varying the *k*-mer size had little impact, values between 8 and 10 were found to produce best results (figure S7). Using only *k*-mers beginning with 'A' resulted in a slight accuracy degradation (figure S8). Among the three methods for weighting match/mismatch scores per column using the base frequency in the reference set performed best by far, improving Q scores by almost 0.5% points (figure S9).

We did not benchmark speed and memory requirements specifically as these depend heavily on sequence length, reference MSA size and parameter settings. In the tests using reference MSAs sampled from the SSU data-set we observed mothur to align roughly 20 sequences per second per core and SINA to align roughly 2 sequences per second per core. PyNAST was as fast as mothur in the benchmark using a reduced width alignment and matched SINA when using the full 51,000 column MSA. Tests were executed on a non-dedicated heterogeneous cluster comprising current 2, 4 and 8-way servers equipped with Intel and AMD quad core CPUs.

6.6 Discussion

We reported the average Q scores because they are commonly used as accuracy indicator for sequence alignment. However, the values are not directly comparable to results obtained for *de novo* methods as these lack the benefit of a guiding reference alignment. Given a consistent reference alignment, selecting a reference sequence closely resembling the candidate sequence and transferring the alignment positions of the shared segments suffices to perfectly align those shared segments. The identity between the candidate sequences and the available reference sequences should therefore be considered as a baseline when interpreting the results. This also affects the precision with which accuracy can be measured. As can be seen in Table 6.2, the variance among sampled test cases was extremely low. When considering only those sequences that had an identity with the reference sequences of less than 80%, variance increased by an order of magnitude. We therefore believe that assessing alignment accuracy to a precision of 0.1% is permissible for the benchmarks we performed.

In interpreting the results, it may also be more informative to consider error rates, rather than the fraction of correctly aligned bases. For example, PyNAST achieves 98.55% accuracy (Q) on the BRAliBase 5S rRNA data-set whereas SINA achieves 99.23%. This amounts to error rates of 1.45% and 0.77%, thus SINA placed only half as many bases in the wrong columns. Since sequence alignment is only one of many sources for error in sequence alignment, the permissible margin of error depends on many factors. We can, however, determine an upper bound at which it is more sensible to forgo extension of the reference MSA and instead use a homology search to map candidate sequences to results based solely on the reference MSA. In this case the error would be equivalent to the distance between candidate and best matching reference because both error and distance are measured as a fraction of differing base positions. The average distance may therefore be used as a point of reference for the permissible error. Methods expecting a MSA as input do not commonly incorporate measures to deal with errors in the MSA. They will also make mutual comparisons between the aligned candidate sequences. Demanding that the error be at least an order of magnitude lower than the distance therefore seems prudent.

According to the SSU benchmark, the distance between candidates and references averages to 7.66% using 1000 reference sequences and 4.76% using 5000 reference sequences. The same benchmark shows error rates for the NAST based methods of above 3.37% and 2.19%. In absolute numbers, this means that when using a 5k reference MSA, the candidates and their best matching reference sequences where on average distinguished by 71 positions

(according to the original alignment). 32 positions where misaligned by NAST. SINA fares much better. At 0.74% error rate (or 11 misaligned positions), its error was only a third of that produced by PyNAST and mothur. While the aforementioned order of magnitude difference between error and distance would demand at most seven misaligned positions, we may have reached the resolution of the benchmark.

When manually inspecting the positions comprising the error, we found that most cases were related to extensions of homo-polymers, conflicts between primary and secondary structure alignment or inconsistencies in the reference MSA. From the SILVA rRNA gene data-sets and the online SINA alignment service, both of which having been available for several years now, we were able to gather user feedback on these shortcomings. In general, users stated that the changes they made in manually refining the SINA alignment were related to the secondary structure. However, we were unable to collect sufficient problematic sequences in which secondary structure awareness would clearly improve alignment accuracy to build a data-set for benchmarking. We therefore concur with the observation made by Kemena [134] that much larger, high quality benchmark MSAs are needed, especially for improving and evaluating the accuracy of high throughput MSA methods. While the data-set extracted from the SILVA SSU Ref database used in the evaluation of SINA is of high quality, it is merely a subset of the SILVA SSU seed. As such, it lacks a representative distribution of distances between sequences and would require further refinement and extension to become a good benchmark. Furthermore, a benchmark MSA explicitly constructed to comprise fewer columns than a correct alignment demands would be required to test the performance of alternative methods for constraining the number of columns. Since we expect that many other genes besides the RNAs will soon become available in numbers surpassing what can be feasibly aligned using *de novo* techniques, we also see a need for advanced interactive tools to support building and curating large MSAs to be used as benchmark or reference MSAs. Once benchmarks of sufficient resolution at high alignment accuracy levels become available, it may be interesting to investigate whether improving the POA based stage in SINA with methods employed by *de novo* MSA tools such as Infernal, MUSCLE or MaFFT can further enhance alignment accuracy.

6.7 Conclusion

We have shown that combining a *k*-mer distance search with POA incremental multiple sequence alignment to integrate candidate sequences into an existing MSA yields highly accurate results. Using multiple reference sequences as a basis for the alignment of the candidate sequences significantly improves ali-

gnment quality. Dynamically selecting a low, fixed number of sequences from which the alignment template is constructed rather than basing the alignment on a global template built from all reference sequences allows the use of very large reference MSAs, lowering the number of bases remaining unaligned because they do not occur in the reference MSA. Furthermore, suboptimal alignment behavior for groups of novel candidate sequences can be easily corrected by manually optimizing the alignment of one of these sequences and adding it to the reference MSA.

With SINA we provide a versatile and flexible tool for accurate high throughput multiple sequence alignment that has proven its reliability and robustness over several years of testing in the context of the SILVA project.

Acknowledgments

We would like to thank all those SINA users who have over the past years provided us with invaluable feedback, without which SINA would neither be as accurate nor as robust as it is today. We would also like to thank Ralf Westram for his help in interfacing SINA with ARB and his great work on ARB in general and Dr. Wolfgang Ludwig for providing us with deep insights into the topics of alignment and rRNA. This work was funded by the Max Planck Society.

CHAPTER 7

ARB: A Software Environment for Sequence Data

Authors: Ralf Westram, Kai Bader, Elmar Pruesse, Yadhu Kumar, Harald Meier, Frank Oliver Glöckner and Wolfgang Ludwig

Status: Published in *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, 2007, de Bruijn, F. J., editor, pages 393–398. John Wiley & Sons.

7.1 INTRODUCTION

Comparative sequence analysis of evolutionary conserved marker molecules nowadays is the standard procedure for assigning organisms to phylogenetic groups and/or taxonomic units. The current prokaryotic taxonomic framework is mainly based on rRNA-based phylogenetic conclusions [170, 173]. This approach provides the basis for identification or new description in pure culture investigations or culture-independent studies of complex environmental samples [6]. Furthermore, comparative analysis of appropriate markers allows assigning contigs to taxa in metagenomics studies.

Powerful interoperating bioinformatics tools are prerequisites for sound utilization of the data flood for identification and phylogenetic inference in the genomics era. Such tools were missing or only available as standalone programs when the ARB project was initiated about 16 years ago [177]. Given this situation, two major goals were formulated in the early days of the ARB project and are maintained to the present: (1) the maintenance of a structured integrative secondary database combining processed primary structures and any type of additional data assigned to the individual sequence entries and (2) a comprehensive selection of software tools directly interacting with one another as well as the central database which are controlled via a common

graphical interface. Initially, the ARB package was designed for handling and analyzing rRNA data. Later, it was extended by developing and/or including software tools for managing protein sequences as well as contigs and genomes.

Currently, the ARB project is maintained by members of the institutions with which the authors of this chapter are affiliated. The ARB package [145, 146, 175, 177] as well as expert-curated rRNA databases [226] are freely available via `http://www.arb-home.de` and `http://www.arb-silva.de`.

7.2 THE ARB SOFTWARE PACKAGE

The ARB software package provides a set of cooperating tools for database maintenance and managing as well as data handling and analysis. These tools directly interact with a central database of processed sequence and various types of sequence associated meta data. A common graphical user interface allows data access, modification, and analysis. The database structure as well as the mode and parameters of interaction of the software tools are customizable by the user to a large extent.

7.2.1 The ARB Main Window

After database selection and ARB program start, the ARB main window provides the turnip for accessing the various software tools and facilities of the ARB package via the respective menus and buttons (Fig. 7.1). Furthermore, a user-selected tree is shown in radial or (two different) dendrogram formats. Primary data and metadata can be visualized at the terminal nodes. Compression of the view is possible by depicting user defined (phylogenetic) groups as triangles or rectangles in radial trees or dendrograms, respectively. Alternatively, these data can be shown by simple listing. Datasets for further analyses can be selected by mouse button directed “marking” of the respective internal or terminal nodes. Opening a slave window for tree comparisons is also possible. The respective trees can be exported to xfig – a simple open source graphics program (`http://www.xfig.org`) – for further modification and/or transformation into various formats.

7.2.2 The Central Database

The central component of the ARB package is a special hierarchical and highly compressed database. During operation, it is loaded in the main memory ensuring rapid access by the peripheral software tools. The sequences representing

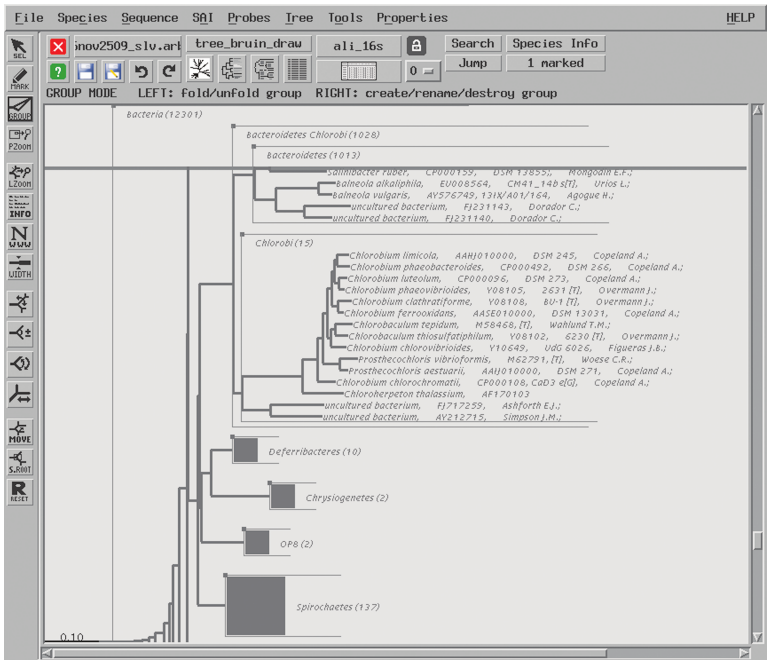


Figure 7.1 *The ARB main window. Buttons in top and left panels provide access to the various ARB tools. Phylogenetic groups are indicated by brackets, and condensed groups are represented by rectangles along with numbers of terminal nodes hidden. NDS (node display settings)-controlled database field entries at terminal nodes indicate the names, accession numbers, strain designations of the respective organisms (master entries), and first authors of the respective bibliography.*

organisms, genes, or gene products are stored in individual database fields. Different sequences (genes, contigs, nucleic acid, and protein sequences) of the same organism can be stored in individual containers (alignments) assigned to the same master entry (organism). A unique identifier (short_name) is automatically generated and assigned to each master entry under the control of a "name server." Following the ARB concept of an integrative database, any type of additional data can be assigned to the individual master entry and stored within default or user defined database fields. Besides a set of default database fields, additional ones can be created, deleted, and renamed by the user. The metadata can either be intrinsic parts of the database or linked to it via local networks or the internet. In the latter case the path to the respective file or the URL of an external database—optionally including commands and search strings—have to be defined using the ARB WWW (world wide web) tool. The default hierarchy of the database entries is according to the phylogeny of the organisms derived from the respective sequence data. However, it can also be changed according to other criteria defined by database field entries. This hierarchy is used by special algorithms for highly effective data compression. Different protection levels (0–6) can be assigned to the individual database fields. Database as well as security management is facilitated by this tool. Data import and export is possible in various common flat file formats. Default or user-defined parsing filters control the storage or extraction of data and features into and from defined ARB database fields, respectively. A versatile merge tool allows data merging and exchanging between different ARB databases. A similar tool can be used for exporting of data subsets in the ARB format.

7.2.3 Data Access and Visualization

Multiple alternative ways provide data access, selection, visualization, modification, and analysis using the ARB package. As mentioned above (see Section 7.2.1), the tree or list shown in the ARB main window can be used for browsing the data. Phylogenetic trees generated by intrinsic ARB tree reconstruction tools or imported from external sources are stored in the database and can be visualized in different formats within the ARB main window. Any (combination of) database field entries can be visualized at the terminal nodes of the tree currently shown (Fig. 7.1). Selection and order of data entries, the results of data analysis, or extraction to be visualized are defined by the NDS (node display settings) tool. Irrespective of the visualization mode used, the ARB SRT (search and replacement tool), ACI (ARB command interpreter), and RGE (regular expressions) tools can be used for extraction of combinations of (sub)strings as well as for analysis of database field entries, respectively.

A powerful search tool allows simple (strings and combination of strings)

and complex (default or user-defined algorithms) searches in one or more (up to three) of the database fields. The matching master entries are shown in a hit list along with restricted information on the respective hits. Selecting from this list provides access to the information in all or user-defined selections of database fields.

The "info" window – the standard tool for data visualization – lists the database fields along with the respective stored information for one master entry. Database field selection and order in this list can be customized by the user. Furthermore, editing of the field entries is possible using this tool. Multiple windows can be opened allowing simultaneous data access for different master entries. Besides this standard procedure, raw and processed data visualization is possible via "user masks." The layout of the visualization windows (i.e., selection, size, and positioning) of database field entries can be customized by the user. Furthermore, simple algorithms for modifying and analyzing of database field entries (SRT, ACI, RGE) can be included when designing "user masks."

7.2.4 Sequence Editors

A powerful editor provides versatile user access to primary structure (nucleotide or amino acid sequences) visualization, arrangement, and modification (Fig. 7.2). The set of sequences to be displayed can be interactively defined as well as stored in user-defined "configurations." The arrangement of the primary structures depends on the tree displayed in the ARB main window or is taken from a "configuration" selected while starting the editor. The original data as well as virtually transformed (e.g. purine-pyrimidine, in silico translated amino acid sequences, or simplified amino acid presentation) data are displayed in user-defined color codes. Keyboard customization is possible for data entry and modification. Two different editing modes can be selected. The "Align" mode allows inserting/removing alignment gaps and moving sequence characters or stretches, while character changes are possible by switching to the "Edit" mode. The rights to overcome protection of the individual sequence entries can be given for the two modes independently. This helps to prevent unwanted character changes when manually modifying the sequence data or the alignment. A set of hot keys in combination with (alignment, sequence, reference, or helix specific) cursor positioning facilities support easy navigation. Block operations are available for modifying the respective primary structure or alignment regions. Sets of search strings can be defined and optionally stored. Perfect or partial matches can be visualized within the displayed sequences by user-defined background colors (Fig. 7.1). Virtual compression—removal of alignment gaps common to all or a certain fraction of the displayed sequences—is

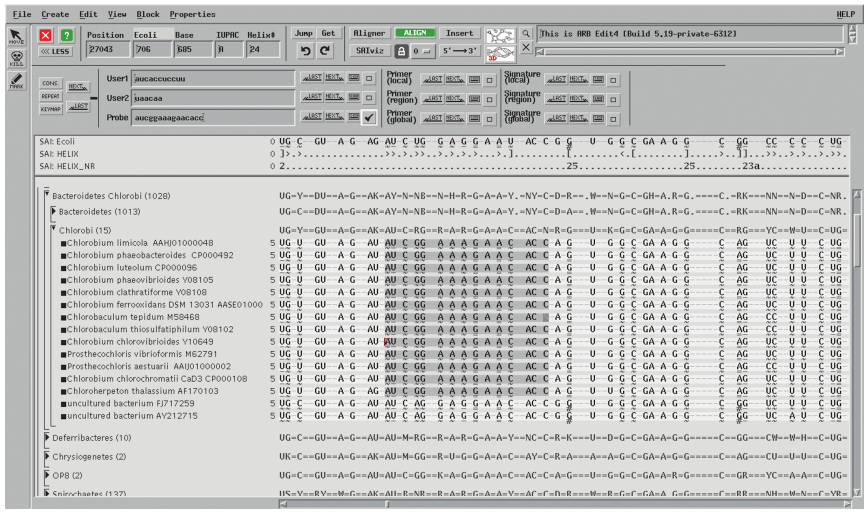


Figure 7.2 *The ARB primary structure editor. Buttons in top and left panels provide access to the various editor-associated ARB tools. Subwindows in the upper part indicate cursor positioning, error messages, and search strings. SAI (sequence-associated information) lines show the* E. coli *reference sequence as well as secondary structure mask and helix numbering. Condensed groups as shown in Figure 7.1 are represented by the respective consensus. The "Probe" search string is highlighted in the respective primary structures. Positional base pairing (~,-,+,=) or consensus secondary structure violation (#) is indicated below the base symbols.*

possible. This makes data inspection and editing more convenient in case of large insertions occurring in only part of the sequences. Groups of sequences can be interactively defined or are automatically shown if defined according to the tree selected while starting the editor. Consensus sequences are determined for each defined group of sequences according to default or user defined criteria and optionally visualized along with or instead of the individual sequences. This consensus can be edited, and changes made concern any sequence in the group. A special feature of the editor is the simultaneous secondary structure check if rRNA (gene) data are visualized. Symbols indicating the presence or absence as well as the character of base pairings are shown below the individual nucleotide symbols and immediately refreshed during sequence editing. A (three-domain) consensus secondary structure mask established according to commonly accepted secondary structure models [32] functions as a guide for this tool.

The ARB (nucleotide) secondary structure editor fits any sequence selected by cursor positioning in the primary structure editor into the common consensus model (Fig. 7.3). The layout of the structure—that is, color coding of base paired, nonpaired, and loop positions as well as the arrangement,

shape, and size of helices and loops—can be customized according to the user's preferences. Any of the search strings or SAIs ("sequence associated information"; see above) activated in the primary structure editor can be visualized by background colors in the secondary structure model [145]. The structure can be exported to xfig – a simple open source graphics program (`http://www.xfig.org`) – for further modification and/or transformation into various formats.

Three-dimensional (3-D) presentation of the respective sequence optionally with search string and SAI visualization is also possible [145]. Color coding can be customized as described for the secondary structure editor. The 3-D structure is based on x-ray structure data for the rRNA molecules of Escherichia coli [17, 288].

The primary structure editor contains a "protein viewer" component allowing in silico translation and virtual presentation of database inherited nucleic acid sequences in selected or all frames. Two- and three-letter as well as user-defined color code presentation is possible. This tool helps when performing primary structure quality checking and optimizing the respective alignment. For further analyses of the in silico translated amino acid, sequences have to be stored in a separate protein sequence alignment (database field; see Section 7.2.1). The respective nucleic and amino acid alignments can be synchronized (see Section 7.2.8).

7.2.5 Profiles, Masks, and Filters

Conservation or base composition profiles, higher-order structure masks, and filters including or excluding particular alignment positions are important tools for sequence data analyses, especially for phylogenetic inference [170, 222]. The ARB package provides tools for determining such profiles based upon the full database or user-defined subsets. These profiles, masks, and filters are stored in the central database as so-called SAIs and can be visualized and modified by the primary structure editor. The filter selection tool not only allows us to choose sets of particular filters but also allows to perform a fine tuning with respect to the inclusion or exclusion of alignment positions in case of multiple character filters. Besides SAIs derived from the primary structures, any other information that can be assigned to sequence/alignment positions or regions can be stored and used as SAIs. Examples are rRNA–protein interaction sites or "in situ" accessibility maps for FISH (fluorescence in situ hybridization) [6, 145] probes.

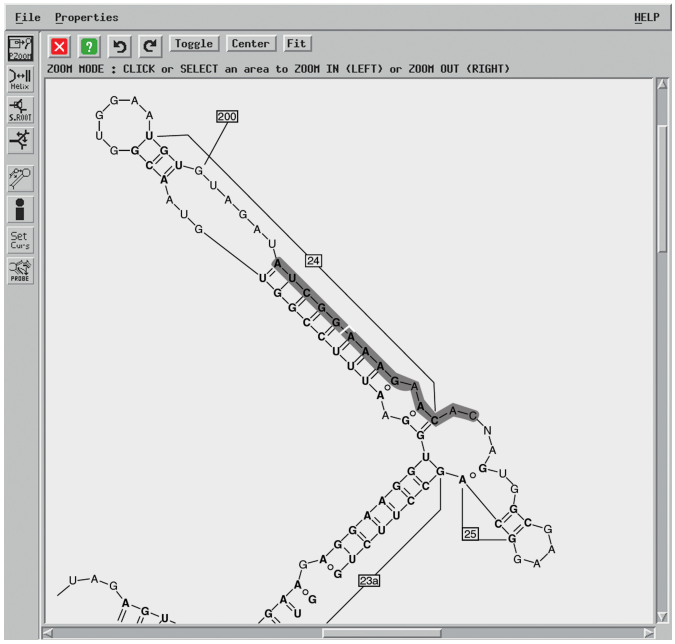


Figure 7.3 *The ARB secondary structure editor. Buttons in top and the left panels provide access to the editor associated layout tools. The "Probe" search string is highlighted (see Fig. 7.1).*

7.2.6 Phylogenetic Treeing

Software tools for nucleotide and amino acid sequencebased tree reconstruction according to the three most commonly used approaches (i.e. distance matrix, maximum likelihood, and maximum parsimony-based procedures) are incorporated in the package. They cooperate as intrinsic tools with the respective ARB components and database elements such as alignment and filters.

The central treeing tool of the package—ARB parsimony—is a special development for the handling of several thousand sequences (more than 500.000 in the current small subunit (SSU Ref) rRNA SILVA database [226]. New sequences are successively added to an existing tree according to the parsimony criterion. A special software component superimposes branch lengths to the parsimony generated tree topology. These branch lengths reflect the significance of the individual "tetra-furcations" by expressing the difference of the most and the two less parsimonious solutions when performing NNI (nearest-neighbor interchange of adjacent branches or sub trees). These relative distances are normalized according to a distance matrix deduced from primary structure comparison. Thus branch lengths in ARB-parsimony-generated trees in the first instance visualize the significance of topologies, while in the second

instance they reflect a degree of estimated sequence divergence. A special feature of ARB parsimony allows adding sequences to an existing tree without permitting any changes in the initial tree. This enables the user to include partial, low-quality or preliminary aligned sequences without perturbing the topology of an optimized tree based upon optimally aligned fulland high-quality data. Another peculiarity of this treeing software concerns the tree optimization by performing cycles of NNI (nearest-neighbor interchange) and KL [136] topology modifications. These optimizations can not only be performed for the complete tree but also can be confined to user-selected subtrees. Thus tree optimization is possible by applying the appropriate filters for the respective phylogenetic levels and groups.

The ARB-neighbor tool for generating distance matrix trees is an accelerated and improved version of the respective component of Felsenstein's [72] PHYLIP package.

Selected stand-alone tools of the former package can be used in the ARB environment in combination with all respective ARB features.

The various facilities of the currently most powerful maximum likelihood program RAxML [265] can also be operated from the ARB user interface applying parameters and filters generated by the respective ARB features. Besides RaxML, also TREE-PUZZLE [250] and PhyML [97] versions can be used for ARB controlled tree reconstruction.

A "concatenation" tool allows merging alignments of different genes or gene products for multiple markerbased phylogenetic studies. The full spectrum of filter and parameter setting is available for analyzing or controlling the influences of the individual markers in the concatenated set.

7.2.7 The Positional Tree Server

Once established, the ARB PT server (positional tree) allows rapid and exact searching for sequence identity or peculiarity. Thus, it represents the central tool for fast searching of closest relatives for automated sequence alignment or to define diagnostic sequence stretches for primer and probe design. Establishing a prefix tree server of any oligonucleotide sequence up to 100-mers occurring in the underlying database and assignment of the individual oligonucleotides to the sequences or organisms containing them is the basis for these procedures. PT-server-based analyses do not rely upon aligned sequences. The PT server is not provided with the ARB program package or ARB database. It has to be established for the respective database locally. The PT server is used for rapid finding of the most similar reference sequences indicating the closest relative of the query organism. This also helps finding appropriate templates for adding new sequences to existing alignments (see Section 7.2.8). The PT

server is also used for finding (taxon- or group-specific) diagnostic sequence stretches for probe and primer design and evaluation (see Section 7.3).

7.2.8 Sequence Alignment and Quality Checks

For de novo-generating nucleic or amino acid sequence alignments, ClustalW [280] was added to the peripheral tools of the ARB package. However, in the context of database maintenance, new sequence entries have to be integrated in an already existing database of aligned sequences. For this purpose the ARB fast aligner was developed. This aligner uses a (set of) selected aligned reference sequences as template(s) for rapid integration of a (set of) unaligned sequence(s). Individual entries—that is, sequences or consensus defined by the user or automatically determined by PT-server-based search for most similar reference sequences—are used as template.

In case of protein coding nucleic acid sequences, the alignment usually is optimized on the amino acid level (given that the phylogenetic information is stored there) [222]. The underlying nucleic acid alignment can then be adapted to the amino acid alignment by a back-translation based tool taking into consideration all known codon usages.

Once a reasonable data set of high quality and optimally aligned primary structures is reasonably structured (grouped) according to the results of careful phylogenetic analyses, further sequence and alignment quality checking is possible using the respective ARB tools. A component of the primary structure editor takes into account SAIs (see Section 7.2.4) expressing positional variability as well as phylogenetic tree topologies for estimating reasonability of a certain monomer (nucleotide or amino acid) at a certain alignment position. The degree of "(miss)-fit" is optionally indicated by user defined background colors in the editor window. Another tool determines a quality score for the individual sequences by estimating degrees of deviation from group specific primary and secondary structure consensus, conservation profiles, sequence sizes, and completeness.

7.3 Probe Design and Evaluation

Taxon- or gene-specific probes or primers certainly play a central role in many molecular biological research and analysis projects—for example, the identification and detection of organisms in complex environmental samples or expression studies within the scope of genome projects. The ARB "Probe Design" and "Probe Match" tools are searching the PT server identify short (10–100 monomers) diagnostic sequence stretches that are evaluated against the background of all sequences in the database the PT server has been built from. In

principle, no alignment of the sequence data is needed for specific probe design. However, in the case of taxon-specific probes, alignment and phylogenetic analyses are necessary for defining groups of phylogenetically (taxonomically) related organisms as the targets of specific probes. The design of taxonspecific oligonucleotide probes with ARB is performed in three steps. First, the (group of) target organism(s), gene(s), or sequence(s) has to be defined ("marked," see Section 7.2.1). Second, potential target sites are searched by the "Probe Design" tool with the aid of a PT server. The results are shown in a ranked list of proposed targets, probes, and additional information. The ranking is according to in silico-predicted probe quality. Third, the proposed oligonucleotide probes are evaluated against the whole database by using the program "Probe Match." Local alignments are determined between the probe target sequence(s) and the most similar reference sequences (optionally from 0 to 5 mismatches) in the respective database. Furthermore, these sequence strings can automatically be visualized in the primary and secondary structure editors (see Section 7.2.4). A special advancement is the ARB multiprobe software component. It determines sets of up to five probes optimally identifying the target group. Color-coded visualization of target master entries (see Section 7.2.2) and matching probe combinations is possible in the ARB main window.

7.3.1 Further Useful ARB Tools

A large fraction of sequences in the currently available rRNA sequence databases [226] comprises clusters of highly similar to identical primary structures most often retrieved by culture independent environmental studies. Commonly, such "sequence clouds" are represented as OTUs (operational taxonomic units) in further data analyses. Such OTUs are defined either manually or by applying respective software tools [249]. Using the ARB package OTUs can be defined and automatically grouped in the selected tree by a newly developed component. The OTU definition according to user provided parameters is deduced from topology of a selected tree. A (best) representative is proposed by the software.

ARB can also function as a simple genome viewer allowing comparison of annotated contigs or genomes. Data access is possible by "search" and "info" tools, alternatively via genome maps similarly as described in Section 7.2.2. Extraction of (sets of) genes into ARB gene databases can also be managed by this ARB facility.

7.3.2 Availability and Training

The ARB software has been designed for Linux operating systems. Tested versions for SuSE and Ubuntu Linux distributions are available at `http://www.arb-home.de` and `http://www.arb-silva.de`. The binaries, source code, and some documentation are provided in the download area of these web pages. The latter URL also provides access to the current release of the SILVA LSU and SSU rRNA databases. Furthermore, there is a user group for the world wide ARB community. Subscription is needed for those interested in joining (`subscribe@arb-home.de`). Basic and advanced ARB training courses are offered by the company Ribocon GmbH in Bremen (Germany, `http://www.ribocon.com`). Mac users interested in ARB should contact `http://www.haloarchaea.com/resources/arb/`.

7.4 CONCLUDING REMARKS

The ARB software package provides a powerful and comprehensive set of directly cooperating software tools for managing and analyzing integrative databases of sequences. It is in use worldwide. The ARB software and database maintaining teams try to keep it up to date and compatible with the ongoing hardware developments. Given more than 16 years of ARB development by different computer scientists and a large number of students of computer science, the huge and heterogeneous source code needs to be cleaned and at least partially redesigned. However, it is difficult to get funding or sponsoring for software redesign.

INTERNET RESOURCES

ARB software (`http://www.arb-home.de`)
ARB databases (`http://www.arb-silva.de`)

Acknowledgments

ARB software and database maintenance is partially supported by the Deutsche Forschungsgemeinschaft and the Bayerische Forschungstiftung, as well as the Max Planck Society.

Part III

Applications

CHAPTER 8

Analysis of 23S rRNA genes in metagenomes – a case study from the Global Ocean Sampling Expedition

Authors: Pelin Yılmaz, Renzo Kottmann, Elmar Pruesse, Christian Quast and Frank Oliver Glöckner

Status: Published in Systematic and Applied Microbiology, 2011, Vol. 34, Issue 6, pages 462-469.

ABSTRACT

As an evolutionary marker, 23S ribosomal RNA (rRNA) offers more diagnostic sequence stretches and greater sequence variation than 16S rRNA. However, 23S rRNA is still not as widely used. Based on 80 metagenome samples from the Global Ocean Sampling (GOS) Expedition, the usefulness and taxonomic resolution of 23S rRNA were compared to those of 16S rRNA. Since 23S rRNA is approximately twice as large as 16S rRNA, twice as many 23S rRNA gene fragments were retrieved from the GOS reads than 16S rRNA gene fragments, with 23S rRNA gene fragments being generally about 100 bp longer. Datasets for 16S and 23S rRNA sequences revealed similar relative abundances for major marine bacterial and archaeal taxa. However, 16S rRNA sequences had a better taxonomic resolution due to their significantly larger reference database.

Reevaluation of the specificity of previously published PCR amplification primers and group specific fluorescence *in situ* hybridization probes on this metagenomic set of non-amplified 23S rRNA sequences revealed that out of 16 primers investigated, only two had more than 90% target group coverage. Evaluations of two probes, BET42a and GAM42a, were in accordance with previous evaluations, with a discrepancy in the target group coverage of the GAM42a probe when evaluated against the GOS metagenomic dataset.

8.1 Introduction

Metagenomics, the study of community genomes taken directly from the environment, allows the cultivation-independent access to the diversity and functional information of microbial communities in their natural habitats [103]. For marine habitats, at least 51 metagenome studies are currently available [162]. One of the largest and geographically most comprehensive is the Global Ocean Sampling (GOS) Expedition. The initial dataset consisted of 6.3 billion bp of Sanger sequence reads obtained from 41 surface water samples. These 41 samples covered a region from the North Atlantic to the South Pacific [239]. Furthermore, the publicly available GOS dataset has recently been augmented by samples from the Atlantic, Pacific and Indian Oceans [315].

The taxonomic diversity of the GOS metagenomic dataset has been assessed previously based on 16S ribosomal RNA (rRNA) gene fragments [25, 239]. The distribution of 23S rRNA gene sequences in the GOS and other metagenomes remains unexplored. Although the 16S rRNA gene has been established as the standard molecule for analyzing the taxonomic diversity in metagenomes [286, 307], 23S rRNA offers advantages over 16S rRNA. With an average length of 2900 bases, it is almost twice as long as the 16S rRNA and, therefore, is theoretically a more informative phylogenetic marker than the 16S rRNA gene [170, 171, 174]. The 23S and 16S rRNA molecules share the same properties in terms of molecule-ubiquity, as well as sequence and structure conservation. Furthermore, phylogenetic trees based on 16S rRNA and on 23S rRNA genes have comparable topologies [172, 235].

A disadvantage of the 23S rRNA gene is the relatively low number of sequences available in the public databases as compared to 16S rRNA genes. Currently (March 2011), only 231,356 23S/28S sequences are publicly available, compared to 1,962,952 16S/18S sequences [226]. Furthermore, the low number of 23S/28S rRNA sequences (20,959) longer than 1900 bases (full-length) limits the assessment of taxonomic diversity due to reduced resolution in taxonomic assignments. The lower number of available 23S rRNA gene sequences can historically be explained by the technical difficulty and higher cost of sequencing the larger molecule with Sanger sequencing technology. However, with new technologies and constantly decreasing sequencing costs, these difficulties are becoming less.

This study is a systematic analysis of 23S rRNA gene sequences in unassembled reads of 80 GOS samples, with the focus on the quantity of retrieved fragments, the fragment length distribution, and the high level taxonomic classification of the fragments. In order to evaluate and validate the classification results obtained using 23S rRNA sequences, a comparison of the bacterial and archaeal diversity of the GOS sites was undertaken based on 23S rRNA and 16S

rRNA gene classifications. Additionally, previously reported 23S rRNA primers and probes have been evaluated based on the extended dataset.

8.2 Materials and methods

8.2.1 Retrieval, alignment and taxonomic classification of 23S/28S and 16S/18S rRNA fragments

Unassembled metagenomic reads for 80 GOS sample datasets were downloaded as a FASTA file from the CAMERA website [253] in September, 2009. A total of 10,085,737 reads, with an average read length of 822 bp, were processed with the SILVA pipeline [227] in order to retrieve 23S/28S and 16S/18S rRNA gene fragments. Aligned fragments were imported into the ARB software suite for further analysis [177]. The fragments were added to the guide trees of the large subunit (LSU (23S/28S)) and small subunit (SSU (16S/18S)) datasets of the SILVA Reference (Ref) release 102 using the ARB Parsimony tool. Fragments with 300-600 aligned bases within the 23S/28S rRNA gene boundaries, and 100-500 aligned bases within the 16S/18S rRNA gene boundaries were added to the guide tree using positional variability filters (an all domain filter for 23S/28S; individual *Bacteria*, *Archaea* and *Eukarya* filters for 16S/18S) excluding highly variable positions indicated by numbers between 1 and 7, which resulted in 2903 out of 3546 valid positions for 23S/28S rRNA sequences, and 1391 out of 1444 positions for 16S/18S rRNA sequences. Fragments with more than 600 aligned bases for 23S/28S rRNA, and 500 aligned bases for 16S/18S rRNA sequences were added with the same positional variability filters but excluding highly variable positions between 1 and 9, leaving 2345 and 1224 valid positions for 23S/28S and 16S/18S rRNA sequences, respectively. Taxonomic assignments are based on membership of the fragments to the existing clades of the SILVA taxonomy, as represented by the guide trees of the high quality SILVA Ref datasets [227]. Taxonomic path assignments were stored in the "tax_slv" field of ARB files using the taxonomy(n) function of ARB Command Interpreter (ACI).

A "Best-BLASTN (Nucleotide BLAST) hit" approach of 23S rRNA fragments was also performed for comparison with the ARB-parsimony approach [2]. Unaligned 23S/28S rRNA fragments retrieved by the SILVA pipeline were used to query the reference dataset of SILVA LSU release 102, using the Tera-BLASTN algorithm (Tera-BLAST™, TimeLogic Inc., Carlsbad, CA, USA). The parameters used for the BLASTN algorithm were as follows: word size=11, extension threshold=20, nucleic match=1, nucleic mismatch=-3, gap open penalty=-5,

Table 8.1 *Percentage of 23S and 16S rRNA gene fragments that can be classified up to Domain, Phylum, Class, Order, Family and Genus levels. Total number of fragments classified are 20,036 and 12,491 for 23S and 16S rRNA, respectively, excluding Eukarya and fragments with less than 300 aligned bases for LSU and less than 100 aligned bases for SSU.*

	23S rRNA gene fragments (%)	16S rRNA gene fragments (%)
Domain	99.9	100.0
Phylum	96.6	100.0
Class	94.4	99.1
Order	78.8	96.3
Family	35.4	80.0
Genus	16.6	31.2

gap extension penalty=-2. Best-BLASTN hits were selected as the top-scoring hit from a group of hits having an expect value of less than 0.00001, and an identity to the query of more than or equal to 97%. The taxonomy of the best hit in the reference dataset was assigned to the query sequence. Further processing of data for taxa abundance counts and method comparisons was performed using MegDB [27].

8.2.2 Primer and probe matching

Sequence Associated Information (SAI) filters corresponding to binding sites of the primers and probes (Supplementary material 1) were manually constructed. These filters were used to count the number of bases within the primer/probe binding sites of all 23S rRNA sequences found in the GOS and SILVA LSU release 102. The target group sequences were chosen from all sequences having a full-length primer/probe-binding region according to these counts. The sizes of primer/probe target groups for GOS and LSU Parc, as well as sequences of the primers and probes are given in Table 8.1 and Supplementary material 1, respectively. Primer/probe matching was carried out manually using the PROBE MATCH module of the ARB software package with the “zero mismatches” and “no weighted mismatches” criteria. Results were parsed and the group coverage in each target group was calculated as the relative number of probe and primer hits to the total number of sequences in the respective target group.

8.2.3 Data Access

23S/28S rRNA sequences retrieved from the GOS metagenomes that were analyzed in this study are publicly available from `http://www.arb-silva.de/download/archive/GOS_diversity/` in ARB format, as well as unaligned and aligned FASTA files. The ARB file contains fields created for the purpose of the primer/probe matching procedure; specifically, fields named with the primer or probe name (example, 129f) contain the PROBE MATCH results as 'pos' if the results reported were positive. The fields carrying the primer name and the suffix '_len' (example, 129f_len) contain the length of the primer/probe binding regions.

8.3 Results and Discussion

8.3.1 Summary of rRNA gene fragment retrieval

A total of 29,581 23S/28S rRNA (0.3% of total reads), and 142,783 16S/18S rRNA (1.4% of total reads) gene fragments were retrieved and aligned using SINA. Fragments with less than 100 aligned bases within the 23S/28S or 16S/18S rRNA gene boundaries were excluded from further analysis, which reduced the dataset to 22,575 23S/28S (76% of total 23S/28S) and 12,742 16S/18S (9% of total 16S/18S) rRNA fragments. For the majority of the excluded sequences (>98%) less than 50 bases could be aligned. Excluding these sequences from the analysis increased the reliability of taxonomic assignments, since sequences this short do not carry sufficient phylogenetic information. Ten GOS sample datasets (GS038-GS046, and GS050) had less than five rRNA gene fragments of sufficient length (Figs. 8.1A and 8.2B) and were excluded from further analysis. These sites contained, on average, only 700 total reads, explaining the low fragment retrieval. Furthermore, no rRNA fragments were retrieved from the MOVE858 sample, which was obtained using 0.002-0.22 μm filters, representing the viral metagenome fraction.

The 23S/28S rRNA gene is twice the length of the 16S/18S rRNA gene, hence the probability of retrieving a 23S/28S rRNA gene fragment should be proportionately higher. This expectation was supported by the results of this study, since ratios of almost 2:1 were observed at sites GS000d (904 23S/28S vs. 438 16S/18S), GS029 (351 23S/28S vs. 162 16S/18S), or GS112a (227 23S/28S vs. 113 16S/18S) (Fig. 8.1A). This two-fold difference was also reflected by the average number of fragments retrieved per site, which was 301 for 23S/28S rRNA and 177 for 16S/18S rRNA. Furthermore, 23S/28S rRNA gene fragments were considerably longer than 16S/18S gene fragments (Fig. 8.1B). Where an average 23S/28S rRNA fragment had 836 aligned bases

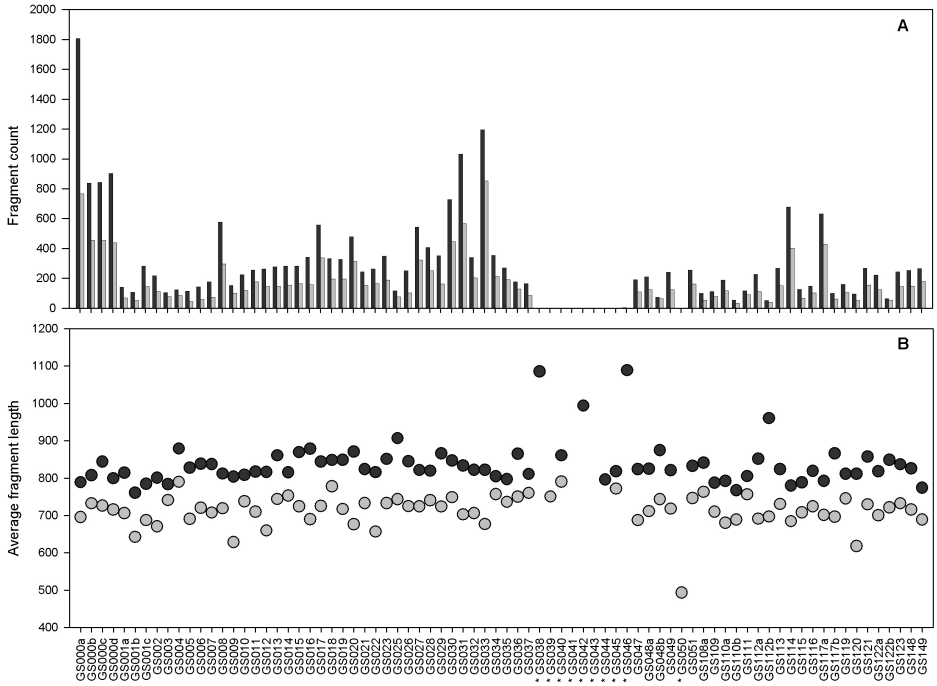


Figure 8.1 *A) Comparison of number of 23S/28S (dark grey bars) and 16S/18S (light grey bars) rRNA fragments retrieved from each GOS sample dataset. B) Average length of 23S/28S (dark grey circles) and 16S/18S (light grey circles) rRNA fragments from each GOS sample dataset in terms of number of aligned bases within rRNA gene boundaries, excluding any fragment (23S/28S or 16S/18S) that contained less than 100 aligned bases. Sites marked with a '*' indicate that less than five fragments were retrieved.*

within the rRNA gene boundaries, a 16S/18S rRNA fragment had 713 aligned bases. More abundant and larger rRNA gene fragments may provide additional information in assessing taxonomic diversity, both with phylogeny and operational taxonomic unit based methods, as well as increasing the chances to affiliate other gene fragments with specific lineages. Both 23S/28S and 16S/18S rRNA fragments were randomly distributed over the rRNA gene regions, meaning that no specific sequence region was over- or under-represented (Supplementary material 3).

8.3.2 Taxonomic diversity based on 23S and 16S rRNA genes

Few eukaryotic sequences (340 28S rRNA and 251 18S rRNA) were retrieved from samples obtained from 0.22-0.8 μm, 0.8-3 μm and 3-20 μm size fractions. These were excluded from further analyses due to the inconsistent taxonomic classification of eukaryotic sequences in databases and to allow greater focus on the bacterial and archaeal fraction. As a result, a total of 20,036 23S rRNA (>300 bases) and 12,491 16S rRNA (>100 bases) gene sequences were classified. Percentages of both 23S and 16S rRNA fragments associated with major marine bacterial and archaeal taxa showed good agreement with each other and with previous studies [83, 89, 224] (Figs. 8.2A and 8.2B). Specifically, based on 23S rRNA assignments, 43% of the retrieved rRNA fragments were found to be associated with *Alphaproteobacteria*, followed by 17% *Gammaproteobacteria*, 9% *Actinobacteria*, 8% *Cyanobacteria*, 8% *Bacteroidetes*, 3% *Betaproteobacteria*, 2% *Euryarchaeota*, and 0.4% *Crenarchaeota* (Fig. 8.2A). However, less agreement in the assignment of 23S rRNA and 16S rRNA fragments was observed with less abundant marine taxa. For example, *Chloroflexi* and *Deferribacteres* associated fragments were not observed in the 23S rRNA gene-based classification, which may be ascribed to the lack of annotated clades for these taxa. In such cases, 16S rRNA gene-based classifications appear to provide better estimations.

Similar trends were observed in sample-by-sample distribution of taxa at the "Class" level for both 23S and 16S rRNA-based assignments, as compared to the previous overall assessment (Figs. 8.3A and 8.3B, Supplementary material 2). *Alphaproteobacteria*, followed by *Gammaproteobacteria*, *Actinobacteria*, *Cyanobacteria*, *Flavobacteria* and *Betaproteobacteria* were the most abundant taxa in the majority of sample datasets. However, differences were observed in the occurrence or relative abundance of minor groups, such as *Planctomycetacia* or *Aquificae*. For example, *Planctomycetacia* associated 16S rRNA fragments were found in 15 sample datasets, whereas only 13 sample datasets contained *Planctomycetacia* associated 23S rRNA fragments. The differences in relative

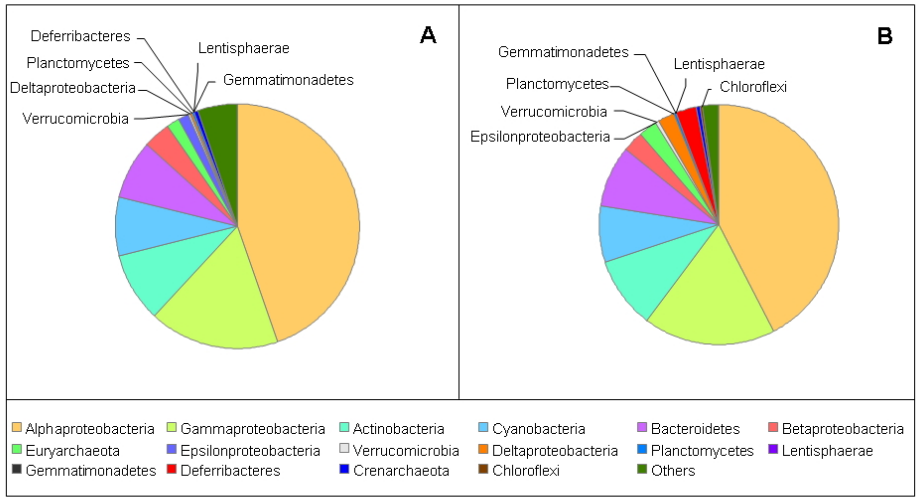


Figure 8.2 *Percentage of 23S (A) and 16S (B) rRNA fragments associated with major marine bacterial and archaeal taxa among all GOS sample datasets, except GS038-GS046 and GS050. Percentages were calculated based on absolute numbers of fragments associated with a given taxa.*

abundance observed with 16S or 23S rRNA-based assignments in these sample datasets were up to six-fold (GS000a). Surprisingly, in certain cases, 23S rRNA-based assessments predicted higher relative abundances or occurrence in sample datasets for other taxa. Up to 12-fold more *Epsilonproteobacteria* associated 23S rRNA fragments were found in sample dataset GS000b compared to 16S rRNA fragments. Additionally, *Lentisphaeria*, which appeared to be present in ten sites according to 23S rRNA classifications, were observed only at two sites according to 16S rRNA gene classifications.

The former case, where 16S rRNA-based assignments estimated more taxa in more sample datasets, demonstrated the current drawback of 23S rRNA-based classification (i.e. its lack of resolution due to insufficient full-length reference sequences). On the other hand, the latter observations demonstrated that when reference sequences are present for a taxon, the higher number of 23S rRNA fragments retrieved can capture what is missed with 16S rRNA fragments.

An evaluation of the suitability of 23S rRNA-based diversity assessments can be obtained by comparing the community composition of contrasting habitats. Subtle differences in contrasting marine habitats are evident and comparable to each other and to general expectations for both 23S and 16S rRNA-based diversity assessments (Figs. 8.3A and 8.3B). For example, *Gammaproteobacteria* were less frequent in estuarine and freshwater habitats compared with coastal and open ocean habitats (GS000a vs. GS020). On the contrary, *Actinobacteria* and *Betaproteobacteria* were more abundant in estuarine and freshwater habitats than in coastal or open ocean habitats (GS011 vs. GS119), underlining previously reported trends [44, 91, 113]. Additionally, a distinct composition was evident in non-open ocean GOS habitats (GS033-hypersaline, GS030-mangrove).

Investigating relative abundances at lower taxonomic levels can shed light on more prominent habitat-specific diversity patterns. However, with the current size and content of LSU rRNA reference databases, the 23S rRNA has a distinct disadvantage in achieving this. As summarized in Table 8.1, the percentage of 23S rRNA gene fragments that can be classified to a certain taxa is comparable to the 16S rRNA gene-based classification at Domain, Phylum or Class levels. A decrease in percentage of classified 23S rRNA fragments was observed at lower levels, from 95% at the Class level, down to even 17% at the Genus level. This can be explained by the 23,197 sequences of taxonomically classified cultured organisms in the SILVA Ref release 102 SSU dataset versus only 3,602 sequences in the LSU Ref dataset.

In addition to the comparison of tree guided taxonomic classification methods, a comparison of the parsimony classification approach to a Best-BLAST hit approach was performed for 23S/28S rRNA gene fragments. BLAST, or mod-

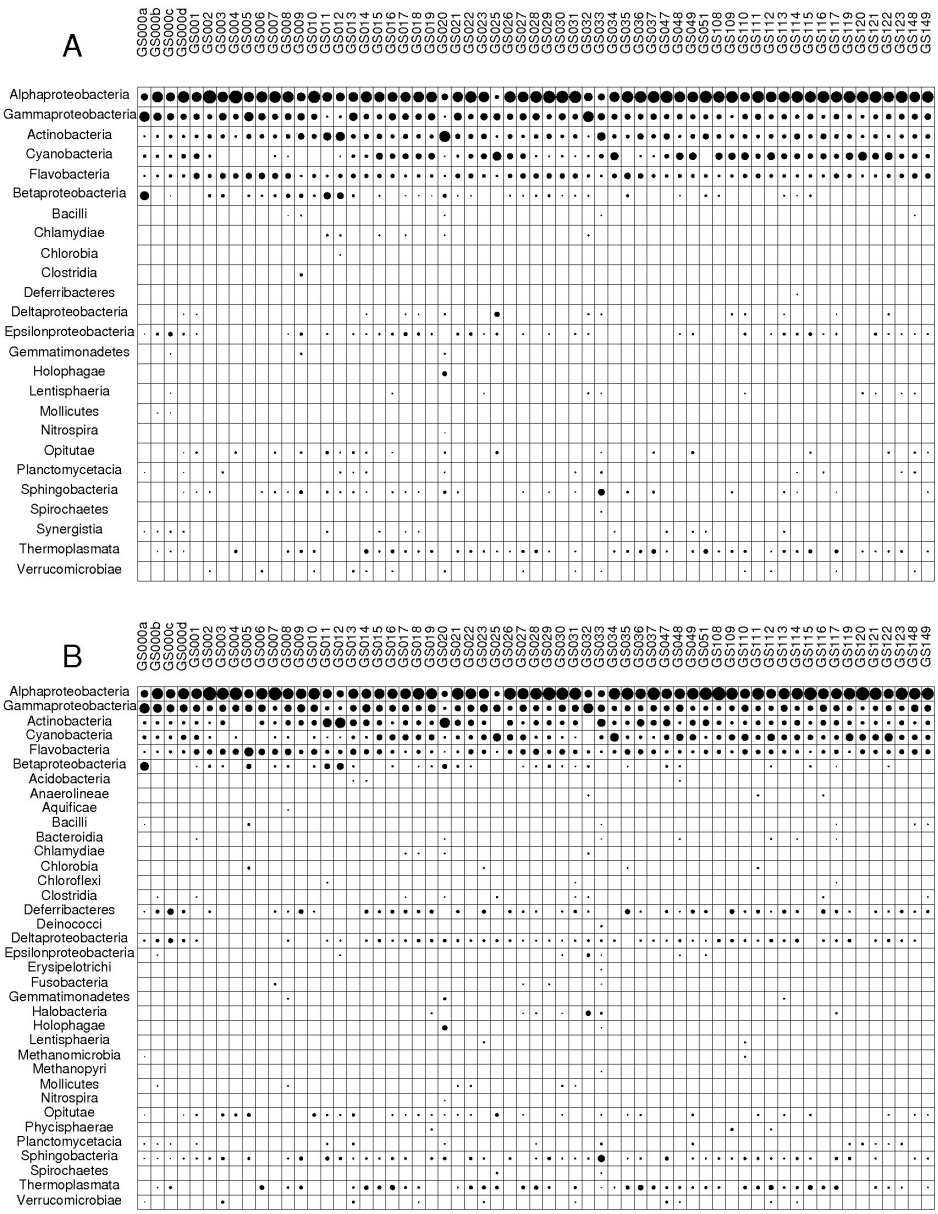


Figure 8.3 *The relative abundance of 23S (A) and 16S (B) rRNA fragments associated with different taxa (rows) at each GOS sample dataset (columns). Presence of a spot indicates the presence of fragments associated with a given taxa, and the area of a spot represents the relative abundance. Relative abundances are based on absolute counts of all fragments from a given site associated with a certain taxa, which are then normalized according to the total fragment counts from that site. Abundances are not normalized with respect to single copy genes, and since rRNA operons can occur multiple times in a genome, the numbers do not represent cell abundances. The taxa shown here are on the 'Class' level, except Cyanobacteria, which is at the 'Phylum' level.*

ifications of this method, are increasingly popular in assessing the taxonomic diversity of high-throughput metagenomic datasets and rRNA surveys. This is due to BLAST being faster than phylogenetic methods, such as ARB Parsimony, and it also provides a means of a multiple-alignment free taxonomic classification approach [121, 122, 191, 261].

A total of 15,798 (excluding 86 Eukaryotic sequences) out of 29,581 unaligned GOS 23S/28S rRNA fragments could be classified using the Best-BLASTN hit approach. Sequences below 300 nucleotides were rejected, revealing a total of 14,656 classified sequences. The BLASTN approach was successful in classifying 5,380 sequences, which were not classified by ARB Parsimony. However, the identity to the target sequence was below the chosen thresholds. The differences between the two methods could be settled by a sufficiently high bit score for the Best-BLASTN hit approach as the sole criterion for assigning taxonomy [39].

In the next step, the taxonomic assignments between Best-BLASTN hit and ARB parsimony were investigated. In summary, 97% of the 14,656 common sequences were assigned identical taxonomy by both methods. The remaining 3.4% (499) of the sequences, which had different taxonomic paths, fell into three different cases: 1) the taxonomic path assigned by the Best-BLASTN hit was at a lower rank compared to ARB Parsimony, 2) the taxonomic path assigned by ARB Parsimony was at a lower rank compared to the Best-BLASTN hit, and finally, 3) the assigned taxonomies were entirely different below a certain rank. For the majority of the sequences (408), the Best-BLASTN hit provided classification at a lower rank (case 1). This is an expected outcome because the taxonomic path is assigned directly from the next relative of the target sequence by the Best-BLASTN hit approach. On the contrary, in the ARB Parsimony approach the taxonomy is assigned based on a group membership, and a sequence can be placed close to, but outside, a group. At a lesser amount, with 28 sequences, a classification to a lower rank was achieved with the ARB Parsimony approach. Finally, 63 sequences had different taxonomic assignments, which could be broken down into 36 sequences assigned to different genera, 15 to different orders, and 12 to different classes.

The relatively small differences in taxonomic assignments between the two methods were encouraging, especially regarding concerns about the suitability of large multiple alignments for taxonomic classification. In response to these concerns, it is important to point out that the SILVA alignment has been rated as having the 'best-quality' within similar projects [247]. Furthermore, the SILVA alignment is based on a reference seed alignment, hence it is not subjected to the many drawbacks of large-scale multiple *de novo* alignments. Finally, with this comparison, it was shown that both Best-BLAST hit and phylogenetic approaches, such as ARB Parsimony, can provide comparable and very similar

results. This methodological comparison showed that if a congruent dataset for taxonomic classification is used, very similar results are obtained, regardless of the algorithms behind the taxonomic classifications.

8.3.3 Specificity of common 23S rRNA primers and probes

The addition of GOS 23S rRNA sequences increases the size of the current 23S/28S rRNA databases (based on SILVA 102 LSUParc) by 12%. Furthermore, they have not undergone PCR amplification, and hence provide a unique opportunity for testing the coverage of previously described universal amplification primers, as well as widely used class-specific probes.

The most recently developed primer sets (129f, 189f, 457r, 2490r) [120], as well as primer 2241r [151], showed reasonable group coverage in the GOS 23S dataset sequences with an average of 85% (Table 8.2), and the results were comparable to those obtained from matching the primers against the SILVA release 102 LSU Parc dataset with only a ±2% difference. The reference dataset used by Hunt et al. [120] was smaller with 2,176 sequences than both the LSU Parc (average of 11,000 target group sequences) and the GOS 23S (average of 5,400 target group sequences) datasets used in this study. However, the authors have included environmental shotgun sequences from the Sargasso Sea pilot study [293] in their dataset, which would account for the comprehensiveness of these primers also in the GOS 23S dataset.

Contrary to these results, the primers developed for the amplification of variable regions of bacterial 23S rRNA sequences (11a-97ar) [290] showed very poor group coverage in the GOS 23S dataset sequences, with generally less than 50% coverage of the target group. A 90% group coverage was only observed for 69ar (Table 8.2). Although the primer binding sites were highly conserved, this was obviously counteracted by the very small dataset that these primers were based on [100]. Surprisingly, primers 53a to 97ar were observed to have higher group coverage within the GOS 23S rRNA sequences than within LSU Parc.

The two archaeal primers (LSU190-F and LSU2445a-R) [51] showed very low group coverage in the GOS 23S dataset (Table 8.2), with 14% and 5%, respectively. Nevertheless, while the percentages were higher in the LSU Parc, they did not exceed 50%.

For the BET42a probe [180], 79% group coverage was found. This, as well as the number of outgroup hits within the GOS 23S dataset, was close to that reported by a previous evaluation [4] (Table 8.2). Group coverage within LSU Parc (87%) was in accordance with Amann and Fuchs [4] (Table 8.2), although considerably more outgroup hits, 348 in LSU Parc vs. 62, were observed.

Table 8.2 *Specificities of selected primers and probes, evaluated on the 23S/28S rRNA gene fragments retrieved from the GOS metagenomes having more than 300 aligned bases within the rRNA gene boundaries, and on the SILVA Parc release 102 LSU dataset. Outgroup hits are the sum of both Archaea and Eukarya in case of bacterial primers, both Bacteria and Eukarya in case of archaeal primers, only Eukarya in case of bacterial and archaeal primers, and non-Betaproteobacteria and non-Gammaproteobacteria for BET42a and GAM42A probes.*

		GOS 23S/28S			LSU Parc		
Primer/probe	Target group	Size of target group	Group coverage (%)	Outgroup hits	Size of target group	Group coverage (%)	Outgroup hits
129f [1]	*Bacteria*	4853	74%	0	10640	82%	4
189f [1]	*Bacteria*	5285	87%	0	11508	87%	0
457r [1]	*Bacteria*	5551	86%	4	11177	83%	279
2241r [2]	*Bacteria*	5832	84%	10	11457	86%	3967
2490r [1]	*Bacteria*	5734	94%	0	10821	98%	0
11a [3]	*Bacteria*	5256	20%	0	11478	39%	0
23ar [3]	*Bacteria*	5619	23%	0	10526	49%	4
43a [3]	*Bacteria*	5633	6%	0	10999	44%	0
53a [3]	*Bacteria*	5320	3%	0	10594	1%	0
62ar [3]	*Bacteria*	5540	8%	0	11455	5%	0
69ar [3]	*Bacteria*	5731	90%	0	11443	87%	0
93a [3]	*Bacteria*	5737	62%	0	10322	55%	0
93ar [3]	*Bacteria*	5731	63%	0	10327	56%	2
97ar [3]	*Bacteria*	4969	55%	0	9165	29%	38
LSU190-F [4]	*Bacteria & Archaea*	5348	14%	0	11741	24% / 28%	0
LSU2445a-R [4]	*Archaea*	142	5%	0	262	28%	0
BET42a [5]	*Betaproteobacteria*	209	79%	63	570	87%	348
GAM42a [5]	*Gammaproteobacteria*	980	42%	1	2877	78%	10

[1]Hunt et al. [120], [2]Lane [151], [3]Van Camp et al. [290], [4]DeLong et al. [51], [5]Manz et al. [180]

The GAM42a probe coverage in the GOS 23S dataset (Table 8.2) was almost half (42%) of the value reported previously (76%) [4], and the corresponding evaluation of the LSU Parc (78%) dataset. Since the mismatches could result from sequencing errors, the alignments of sequences with mismatches to the probe GAM42a were manually inspected. A few cases were likely to be sequencing errors, and were mainly observed in fragments obtained from ends of sequencing reads. The majority of the mismatches revealed consistent, class-specific mismatches. These mismatches were up to four bases, and were found mainly between *E. coli* positions 1030 to 1040. Although this evaluation of the GAM42a probe was based on a single environment, the surface ocean, limitations and anomalous results with the GAM42a probe have been reported previously for other environments as well, which were found to be mainly due to polymorphisms at *E. coli* position 1033 [20, 313]. Our observation confirms these reports, by adding additional polymorphisms before and after this position. Consequently, the limitations of the GAM42a probe might be more severe than previously thought, and therefore we recommend the design and testing of novel *Gammaproteobacteria* probes.

8.4 Conclusions

This study exemplifies the possibility and power of using 23S rRNA genes for biodiversity surveys by providing a comparative overview of 16S and 23S rRNA fragments retrieved from the GOS metagenomes. High quality taxonomic classification for biodiversity analysis, as well as primer and probe design, depends on the size and extent of the reference dataset used. The advantage of using the larger 23S rRNA genes for biodiversity analysis, especially for the marine system, has been shown previously [221]. Additionally, a recent study assessing the diversity of paralogous 23S rRNA genes has shown that significant sequence diversification was observed in 184 species, further supporting the suitability of this molecule for taxonomy [219]. Although an obvious limitation faced during this study was the small size of the 23S rRNA gene reference datasets, this is likely to be overcome in the near future with the contribution of (meta-) genomic sequences from mega-sequencing projects, such as the Human Microbiome Project [289], the TerraGenome [295], Tara Oceans (see `http://oceans.taraexpeditions.org/`) or the Genomic Encyclopedia of *Bacteria* and *Archaea* [308]. Moreover, studies assessing the characteristics and sequence diversity of 23S rRNA genes in bacterial and archaeal genomes, in combination with efforts to design, test and re-evaluate universal and group specific primers and probes [120], can renew the interest and utilization of this molecule. Application of continually advancing, cheaper sequencing technologies to the undiscovered fraction of the 23S rRNA gene sequences can result in

a higher appreciation of this valuable phylogenetic marker.

Acknowledgements

We would like to thank Mar Fernández Méndez and Petra Pjevac for their assistance in phylogenetic and taxonomic analysis of the dataset. We would also like to thank Pier Luigi Buttigieg, Jörg Peplies and Hannah Marchant for their critical reading of the manuscript and helpful suggestions. This study was supported by the Max Planck Society.

CHAPTER 9

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Authors: Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Philippe Rocca-Serra, Peter Sterk, Manimozhiyan Arumugam, Mark Bailey, Laura Baumgartner, Bruce W Birren, Martin J Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D Bushman, Pier Luigi Buttigieg, Patrick S G Chain, Emily Charlson, Elizabeth K Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis, Noah Fierer, Jed A Fuhrman, Rachel E Gallery, Dirk Gevers, Richard A Gibbs, Inigo San Gil, Antonio Gonzalez, Jeffrey I Gordon, Robert Guralnick, Wolfgang Hankeln, Sarah Highlander, Philip Hugenholtz, Janet Jansson, Andrew L Kau, Scott T Kelley, Jerry Kennedy, Dan Knights, Omry Koren, Justin Kuczynski, Nikos Kyrpides, Robert Larsen, Christian L Lauber, Teresa Legg, Ruth E Ley, Catherine A Lozupone, Wolfgang Ludwig, Donna Lyons, Eamonn Maguire, Barbara A Methé, Folker Meyer, Brian Muegge, Sara Nakielny, Karen E Nelson, Diana Nemergut, Josh D Neufeld, Lindsay K Newbold, Anna E Oliver, Norman R Pace, Giriprakash Palanisamy, Jörg Peplies, Joseph Petrosino, Lita Proctor, Elmar Pruesse, Christian Quast, Jeroen Raes, Sujeevan Ratnasingham, Jacques Ravel, David A Relman, Susanna Assunta-Sansone, Patrick D Schloss, Lynn Schriml, Rohini Sinha, Michelle I Smith, Erica Sodergren, Aymé Spor, Jesse Stombaugh, James M Tiedje, Doyle V Ward, George M Weinstock, Doug Wendel, Owen White, Andrew Whiteley, Andreas Wilke, Jennifer R Wortman, Tanya Yatsunenko & Frank Oliver Glöckner

Status: Published in Nature Biotechnology, 2011, Vol 29, No.5, pages

415-420.

ABSTRACT

Here we present a standard developed by the Genomic Standards Consortium (GSC) for reporting marker gene sequences – the minimum information about a marker gene sequence (MIMARKS). We also introduce a system for describing the environment from which a biological sample originates. The 'environmental packages' apply to any genome sequence of known origin and can be used in combination with MIMARKS and other GSC checklists. Finally, to establish a unified standard for describing sequence data and to provide a single point of entry for the scientific community to access and learn about GSC checklists, we present the minimum information about any (x) sequence (MIxS). Adoption of MIxS will enhance our ability to analyze natural genetic diversity documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

9.1 Introduction

Without specific guidelines, most genomic, metagenomic and marker gene sequences in databases are sparsely annotated with the information required to guide data integration, comparative studies and knowledge generation. Even with complex keyword searches, it is currently impossible to reliably retrieve sequences that have originated from certain environments or particular locations on Earth—for example, all sequences from 'soil' or 'freshwater lakes' in a certain region of the world. Because public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprising DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) and GenBank (`http://www.insdc.org/`)) depend on author-submitted information to enrich the value of sequence data sets, we argue that the only way to change the current practice is to establish a standard of reporting that requires contextual data to be deposited at the time of sequence submission. The adoption of such a standard would elevate the quality, accessibility and utility of information that can be collected from INSDC or any other data repository.

The GSC has previously proposed standards for describing genomic sequences – the "minimum information about a genome sequence" (MIGS) – and metagenomic sequences – the "minimum information about a metagenome sequence" (MIMS) [74]. Here we introduce an extension of these standards for capturing information about marker genes. Additionally, we introduce 'environmental packages' that standardize sets of measurements and observations

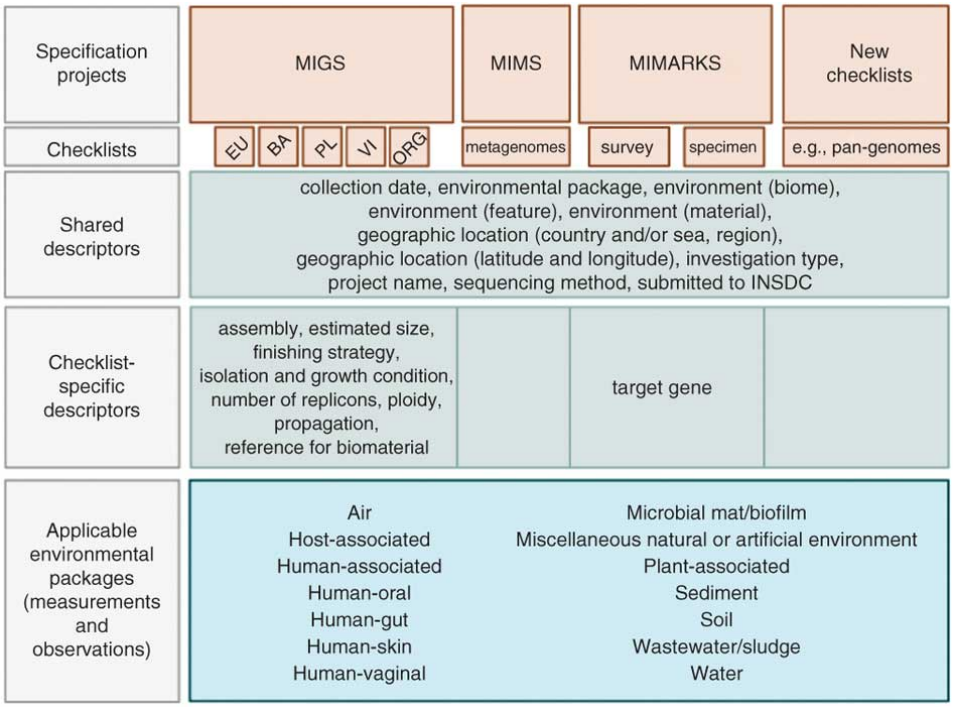
describing particular habitats that are applicable across all GSC checklists and beyond [277]. We define 'environment' as any location in which a sample or organism is found, e.g., soil, air, water, human-associated, plant-associated or laboratory. The original MIGS/MIMS checklists included contextual data about the location from which a sample was isolated and how the sequence data were produced. However, standard descriptions for a more comprehensive range of environmental parameters, which would help to better contextualize a sample, were not included. The environmental packages presented here are relevant to any genome sequence of known origin and are designed to be used in combination with MIGS, MIMS and MIMARKS checklists.

To create a single entry point to all minimum information checklists from the GSC and to the environmental packages, we propose an overarching framework, the MIxS standard (`http://gensc.org/gc_wiki/index.php/MIxS`). MIxS includes the technology-specific checklists from the previous MIGS and MIMS standards, provides a way of introducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages. A schematic overview of MIxS along with the MIxS environmental packages is shown in Figure 1.

9.1.1 Development of MIMARKS and the environmental packages

Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed spacer gene sequences (ITS) from *Bacteria*, *Archaea* and microbial Eukaryotes have provided deep insights into the topology of the tree of life [175, 176] and the composition of communities of organisms that live in diverse environments, ranging from deep sea hydrothermal vents to ice sheets in the Arctic [52, 57, 84, 87, 113, 119, 167, 193, 215, 230, 264, 301]. Numerous other phylogenetic marker genes have proven useful, including RNA polymerase subunits (rpoB), DNA gyrases (gyrB), DNA recombination and repair proteins (recA) and heat shock proteins (HSP70) [175]. Marker genes can also reveal key metabolic functions rather than phylogeny; examples include nitrogen cycling (amoA, nifH, ntcA) [80, 316], sulfate reduction (dsrAB) [192] or phosphorus metabolism (phnA, phnI, phnJ)[86, 184]. In this paper we define all phylogenetic and functional genes (or gene fragments) used to profile natural genetic diversity as 'marker genes'. MIMARKS (Table 9.1) complements the MIGS/MIMS checklists for genomes and metagenomes by adding two new checklists, a MIMARKS survey, for uncultured diversity marker gene surveys, and a MIMARKS specimen, for marker gene sequences obtained from any material identifiable by means of specimens. The MIMARKS extension adopts and incorporates the standards being developed by the Consortium for the Barcode

Figure 9.1 *Schematic overview about the GSC MIxS standard (brown), including combination with specific environmental packages (blue).*



Shared descriptors apply to all MIxS checklists; however, each checklist has its own specific descriptors as well. Environmental packages can be applied to any of the checklists. EU, eukarya; BA, bacteria/archaea; PL, plasmid; VI, virus; ORG, organelle.

of Life (CBOL) [105]. Therefore, the checklist can be universally applied to any marker gene, from small subunit rRNA to cytochrome oxidase I (COI), to all taxa, and to studies ranging from single individuals to complex communities.

Both MIMARKS and the environmental packages were developed by collating information from several sources and evaluating it in the framework of the existing MIGS/MIMS checklists. These include four independent community-led surveys, examination of the parameters reported in published studies and examination of compliance with optional features in INSDC documents. The overall goal of these activities was to design the backbone of the MIMARKS checklist, which describes the most important aspects of marker gene contextual data.

9.1.2 Results of community-led surveys

Four online surveys about descriptors for marker genes have been conducted to determine researcher preferences for core descriptors. The Department of Energy Joint Genome Institute and SILVA [226] surveys focused on general descriptor contextual data for a marker gene, whereas the Ribosomal Database Project (RDP) [42] focused on prevalent habitats for rRNA gene surveys, and the Terragenome Consortium[295] focused on soil metagenome project contextual data (Supplementary Results 1). The above recommendations were combined with an extensive set of contextual data items suggested by an International Census of Marine Microbes (ICoMM) working group that met in 2005. These collective resources provided valuable insights into community requests for contextual data items to be included in the MIMARKS checklist and the main habitats constituting the environmental packages.

9.1.3 Survey of published parameters

We reviewed published rRNA gene studies, retrieved from SILVA and the ICoMM database MICROBIS (The Microbial Oceanic Biogeographic Information System, `http://icomm.mbl.edu/microbis/`) to further supplement contextual data items that are included in the respective environmental packages. In total, 39 publications from SILVA and >40 ICoMM projects were scanned for contextual data items to constitute the core of the environmental package subtables (Supplementary Results 1).

In a final analysis step, we surveyed usage statistics of INSDC source feature key qualifier values of rRNA gene sequences contained in SILVA (Supplementary Results 1). Notably, <10% of the 1.2 million 16S rRNA gene sequences (SILVA release 100) were associated with even basic information such as latitude and longitude, collection date or PCR primers.

Table 9.1 *The core items of the MIMARKS checklists, along with the value types, descriptions and requirement status (part 1)*

Item	Description	Report Type: survey	Report Type: specimen
Investigation			
Submitted to INSDC [boolean]	Depending on the study (large-scale, e.g., done with next-generation sequencing technology, or small-scale) sequences have to be submitted to SRA (Sequence Read Archives), DRA (DDBJ Sequence Read Archive) or through the classical Webin/Sequin systems to GenBank, ENA and DDBJ	M	M
Investigation type [mimarks-survey or mimarks-specimen]	Nucleic Acid Sequence Report is the root element of all MIMARKS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIMARKS survey or MIMARKS specimen	M	M
Project name	Name of the project within which the sequencing was organized	M	M
Environment			
Geographic location (latitude and longitude [float, point, transect and region])	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system	M	M
Geographic location (depth [integer, point, interval, unit])	Please refer to the definitions of depth in the environmental packages	E	E
Geographic location (elevation of site [integer, unit]; altitude of sample [integer, unit])	Please refer to the definitions of either altitude or elevation in the environmental packages	E	E
Geographic location (country and/or sea [INSDC or GAZ]; region [GAZ])	The geographical origin of the sample as defined by the country or sea name. Country, sea or region names should be chosen from the INSDC list (`http://insdc.org/country.html`), or the GAZ (Gazetteer, v1.446) ontology (`http://bioportal.bioontology.org/visualize/40651`)	M	M
Collection date [ISO8601]	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated, that is, all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; except for 2008-01 and 2008, all are ISO6801 compliant	M	M
Environment (biome [EnvO])	In environmental biome level are the major classes of ecologically similar communities of plants, animals and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing and other factors like climate. Examples include desert, taiga, deciduous woodland or coral reef. Environment Ontology (EnvO) (v1.53) terms listed under environmental biome can be found at `http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO:00000428`	M	M
Environment (feature [EnvO])	Environmental feature level includes geographic environmental features. Examples include harbor, cliff or lake. EnvO (v1.53) terms listed under environmental feature can be found at `http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO:00002297`	M	M

Table continues on page 123.

Table 9.2 *The core items of the MIMARKS checklists, along with the value types, descriptions and requirement status (part 2)*

Item	Description	Report Type	
		survey	specimen
Environment (material [EnvO])	The environmental material level refers to the matter that was displaced by the sample, before the sampling event. Environmental matter terms are generally mass nouns. Examples include: air, soil or water. EnvO (v1.53) terms listed under environmental matter can be found at `http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO:00010483`	M	M
MIGS/MIMS/MIMARKS extension			
Environmental package [air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water]	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported	M	M
Nucleic acid sequence source			
Isolation and growth conditions [PMID, DOI or URL]	Publication reference in the form of PubMed ID (PMID), digital object identifier (DOI) or URL for isolation and growth condition specifications of the organism/material	-	M
Sequencing			
Target gene or locus (e.g., 16S rRNA, 18S rRNA, nif, amoA, rpo)	Targeted gene or locus name for marker gene study	M	M
Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)	Sequencing method used, e.g., Sanger, pyrosequencing, ABI-solid	M	M

Items for the MIMARKS specification and their mandatory (M), status for both MIMARKS-survey and MIMARKS-specimen checklists. Furthermore, "–" denotes that an item is not applicable for a given checklist. E denotes that a field has environment-specific requirements. For example, whereas "depth" is mandatory for the environments water, sediment or soil, it is optional for human-associated environments. MIMARKS-survey is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. MIMARKS-specimen, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIMARKS-survey and specimen checklists can be used for any type of marker gene sequence data, ranging from 16S, 18S, 23S, 28S rRNA to COI, hence the checklists are universal for all three domains of life. Item names are followed by a short description of the value of the item in parentheses and/or value type in brackets as a superscript. Whenever applicable, value types are chosen from a controlled vocabulary (CV) or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (`http://www.obofoundry.org/`). This table only presents the very core of MIMARKS checklists, that is, only mandatory items for each checklist. Supplementary Results 2 contains all MIMARKS items, the tables for environmental packages in the MIGS/MIMS/MIMARKS extension and GenBank structured comment name that should be used for submitting MIMARKS data to GenBank. In case of submitting to EBI-ENA, the full names can be used.

9.1.4 The MIMARKS checklist

The MIMARKS checklist provides users with an 'electronic laboratory notebook' containing core contextual data items required for consistent reporting of marker gene investigations. MIMARKS uses the MIGS/MIMS checklists with respect to the nucleic acid sequence source and sequencing contextual data, but extends them with further experimental contextual data such as PCR primers and conditions, or target gene name.

For clarity and ease of use, all items within the MIMARKS checklist are presented with a value syntax description, as well as a clear definition of the item. Whenever terms from a specific ontology are required as the value of an item, these terms can be readily found in the respective ontology browsers linked by URLs in the item definition. Although this version of the MIMARKS checklist does not contain unit specifications, we recommend all units to be chosen from and follow the International System of Units (SI) recommendations. In addition, we strongly urge the community to provide feedback regarding the best unit recommendations for given parameters. Unit standardization across data sets will be vital to facilitate comparative studies in future. An Excel version of the MIMARKS checklist is provided on the GSC web site (`http://gensc.org/gc_wiki/index.php/MIMARKS`).

9.1.5 The MIxS environmental packages

Fourteen environmental packages provide a wealth of environmental and epidemiological contextual data fields for a complete description of sampling environments. The environmental packages can be combined with any of the GSC checklists (Fig. 9.1 and Supplementary Results 2). Researchers within The Human Microbiome Project [289] contributed the host-associated and all human packages. The Terragenome Consortium contributed sediment and soil packages. Finally, ICoMM, Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites and the Max Planck Institute for Marine Microbiology contributed the water package. The MIMARKS working group developed the remaining packages (air, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated and wastewater/sludge). The package names describe high-level habitat terms in order to be exhaustive. The miscellaneous natural or artificial environment package contains a generic set of parameters, and is included for any other habitat that does not fall into the other thirteen categories. Whenever needed, multiple packages may be used for the description of the environment.

9.1.6 Examples of MIMARKS-compliant data sets

Several MIMARKS-compliant reports are included in Supplementary Results 3. These include a 16S rRNA gene survey from samples obtained in the North Atlantic, an 18S pyrosequencing tag study of anaerobic protists in a permanently anoxic basin of the North Sea, a pmoA survey from Negev Desert soils, a dsrAB survey of Gulf of Mexico sediments and a 16S pyrosequencing tag study of bacterial diversity in the western English Channel (SRA accession no. SRP001108).

9.1.7 Adoption by major database and informatics resources

Support for adoption of MIMARKS and the MIxS standard has spread rapidly. Authors of this paper include representatives from genome sequencing centers, maintainers of major resources, principal investigators of large- and small-scale sequencing projects, and individual investigators who have provided compliant data sets, showing the breadth of support for the standard within the community.

In the past, the INSDC has issued a reserved 'barcode' keyword for the CBOL7. Following this model, the INSDC has recently recognized the GSC as an authority for the MIxS standard and issued the standard with official keywords within INSDC nucleotide sequence records [24]. This greatly facilitates automatic validation of the submitted contextual data and provides support for data sets compliant with previous versions by including the checklist version as a keyword.

GenBank accepts MIxS metadata in tabular format using the sequin and tbl2asn submission tools, validates MIxS compliance and reports the fields in the structured comment block. The EBI-ENA Webin submission system provides prepared web forms for the submission of MIxS compliant data; it presents all of the appropriate fields with descriptions, explanations and examples, and validates the data entered. One tool that can aid submitting contextual data is MetaBar [104], a spreadsheet and web-based software, designed to assist users in the consistent acquisition, electronic storage and submission of contextual data associated with their samples in compliance with the MIxS standard. The online tool CDinFusion (`http://www.megx.net/CDinFusion`) was created to facilitate the combination of contextual data with sequence data, and generation of submission-ready files.

The next-generation Sequence Read Archive (SRA) collects and displays MIxS-compliant metadata in sample and experiment objects. There are several tools that are already available or under development to assist users in SRA

submissions. The myRDP SRA PrepKit allows users to prepare and edit their submissions of reads generated from ultra-high-throughput sequencing technologies. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to checklists such as MIMARKS. The Quantitative Insights Into Microbial Ecology (QIIME) web application (`http://www.microbio.me/qiime`) allows users to generate and validate MIMARKS-compliant templates. These templates can be viewed and completed in the users' spreadsheet editor of choice (e.g., Microsoft Excel). The QIIME web-platform also offers an ontology lookup and geo-referencing tool to aid users when completing the MIMARKS templates. The Investigation/Study/Assay (ISA) is a software suite that assists in the curation, reporting and local management of experimental metadata from studies using one or a combination of technologies, including high-throughput sequencing [236]. Specific ISA configurations (`http://isa-tools.org/tools.html`) have been developed to ensure MIxS compliance by providing templates and validation capability. Another tool, ISAconverter, produces SRA.xml documents, facilitating submission to the SRA repository. MIxS checklists are also registered with the BioSharing catalog of standards (`http://biosharing.org/`), set to progressively link minimal information specifications to the respective exchange formats, ontologies and compliant tools.

Further detailed guidance for submission processes can be found under the respective wiki pages (`http://gensc.org/gc_wiki/index.php/MIxS`) of the standard.

9.1.8 Maintenance of the MIxS standard

To allow further developments, extensions and enhancements of MIxS, we set up a public issue tracking system to track changes and accomplish feature requests (`http://mixs.gensc.org/`). New versions will be released annually. Technically, the MIxS standard, including MIMARKS and the environmental packages, is maintained in a relational database system at the Max Planck Institute for Marine Microbiology Bremen on behalf of the GSC. This provides a secure and stable mechanism for updating the checklist suite and versioning. In the future, we plan to develop programmatic access to this database to allow automatic retrieval of the latest version of each checklist for INSDC databases and for GSC community resources. Moreover, the Genomic Contextual Data Markup Language is a reference implementation of the GSC checklists by the GSC and now implements the full range of MIxS standards. It is based on XML Schema technology and thus serves as an interoperable data exchange format for infrastructures based on web services [142].

9.1.9 Conclusions and call for action

The GSC is an international body with a stated mission of working towards richer descriptions of the complete collection of genomes and metagenomes through the MIxS standard. The present report extends the scope of GSC guidelines to marker gene sequences and environmental packages and establishes a single portal where experimentalists can gain access to and learn how to use GSC guidelines. The GSC is an open initiative that welcomes the participation of the wider community. This includes an open call to contribute to refinements of the MIxS standards and their implementations.

The adoption of the GSC standards by major data providers and organizations, as well as the INSDC, supports efforts to contextually enrich sequence data and complements recent efforts to enrich other (meta) 'omics data. The MIxS standard, including MIMARKS, has been developed to the point that it is ready for use in the publication of sequences. A defined procedure for requesting new features and stable release cycles will facilitate implementation of the standard across the community. Compliance among authors, adoption by journals and use by informatics resources will vastly improve our collective ability to mine and integrate invaluable sequence data collections for knowledge- and application-driven research. In particular, the ability to combine microbial community samples collected from any source, using the universal tree of life as a measure to compare even the most diverse communities, should provide new insights into the dynamic spatiotemporal distribution of microbial life on our planet and on the human body.

CHAPTER 10

Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next generation sequencing based diversity analysis

Authors: Anna Klindworth, Elmar Pruesse, Jörg Peplies, Christian Quast, Christine Klockow, Bernhard Fuchs and Frank Oliver Glöckner

Status: in preparation

ABSTRACT

16S ribosomal RNA gene (rDNA) analysis remains the standard approach for the cultivation independent determination of microbial diversity. Polymerase chain reaction (PCR) amplicon-based high throughput sequencing offer a fast and cost effective way to generate massive amounts of 16S rDNA fragments. However, accurateness of such analyses of biodiversity strongly depends on the choice of primers. The overall coverage and phylum spectrum of 54 primers and 96 selected pairs of primers generating fragment length $<$1000 bases were evaluated *in silico* based on the SILVA 16S rDNA reference dataset. Arch20_F/Parch519_R and Arch20_F/A519_R showed the best results for the domain *Archaea*. For *Bacteria*, Bakt_341_F/Bakt_805_R provided sufficient coverage and a large phylum spectrum. The *in silico* evaluation was experimentally verified by comparing the taxonomic distribution of 16S rDNA amplicons from pyrosequencing with 16S rDNA fragments from directly sequenced metagenomes as well as quantitative data from single cell fluorescence *in situ* hybridization (FISH) studies. Additionally, 54 pairs of primers resulting in fragments $>$1000 bases were evaluated *in silico*. This study provides validated sets of 16S rDNA targeting PCR primers to successfully reveal the complexity of archaeal and bacterial diversity using classical, clone library based methods as well as next-generation sequencing methods.

10.1 Introduction

Understanding microbial diversity has been fascinating scientists for decades. Microbes are ubiquitous [118] and their habitats range from terrestrial [76, 237, 248] to oceans [9, 82, 109, 261, 269], any living body [56, 94, 102, 208] as well as plants [70]. They participate in the global cycles of energy and matters, use a wide range of substrates and feature unique metabolic pathways [56, 118]. Therefore understanding patters and function of microbial diversity is of particular importance. It is estimated that between 90-99% of the microbial diversity resists cultivation [6], most likely because of their inability to grow alone or under standard laboratory conditions [287]. Even for the extensively studied habitats, such as the human distal gut, only 20-40% of the bacterial population have been cultivated so far [121]. To overcome this 'cultivation-barrier', culture-independent surveys have been developed. In the past, the most commonly used approach relied on amplifying, cloning and sequencing of the 16S ribosomal RNA gene (rDNA) using conserved broad-range PCR primers [208]. With the advent of massive parallel sequencing technologies, direct sequencing of PCR amplicons became feasible [11, 182, 188]. This fast and cost effective sequencing allows unprecedented statistical analysis and uncovered the "rare biosphere" [261]. In 2006, Roche's 454 pyrosequencing [183] became the first high throughput sequencing technology to be successfully applied in large scale biodiversity analysis [261]. For this purpose, the hyper variable (HV) region 6 of the 16S rDNA was PCR-amplified and sequenced using the first generation of the 454 pyrosequencing platform [261]. With the release of the 454 FLX and Titanium system [59] the throughput and resolution of 16S rDNA sequencing improved further [165]. Consequently, Roche's 454 pyrosequencing technology has been used for microbial diversity analysis in a great range of different habitat types, such as soil [237], human [56, 95, 154, 208], arctic ocean [139] and Baltic sea [9, 139] to name just some. Despite the many advantages of high throughput sequencing approaches, the absence of clone libraries and the relatively short read length still hampers in depth phylogenetic analysis [208, 286, 306]. At present, with up to 1000 bases, the commercially available Roche's 454 pyrosequencing machines produce the longest read length. However, Pacific Bioscience (PacBio) recently introduced the single-molecule real-time (SMRT) sequencing technology [67], which has the potential to produce much longer reads [244].

Besides appropriate fragment length, accurate rDNA analysis heavily depends on the choice of primers [11]. Using suboptimal primers could lead to under-representation [16] or discrimination of single species or even whole groups [102, 286, 300]. For example, the general primer 384F discriminates *Verrucomicrobia* [8] and 967F matches only <5% of *Bacteriodetes* [261]. Using

primers with insufficient coverage consequently results in significantly different biological conclusions [8, 102, 164]. Therefore, careful choice of primer sets is a crucial step to ensure minimum-biased results.

In this study, the overall coverage of a wide range of primers described as 'universal' and/or recently used in diversity analysis was investigated *in silico*. The primer sequences were compared with all 16S rDNA sequences available in the SILVA reference database release 106 [226]. Additionally, selected primer pairs and combinations of primer pairs have been evaluated. *In silico* evaluation was carried out in terms of domain and phylum coverage for *Bacteria* and *Archaea*. With the optimal primer sets for *Bacteria* (fragment length <1000 bases) a field study at Helgoland Roads was performed and the results were compared with PCR independent metagenomic results and single cell fluorescence in situ hybridization (FISH) results.

10.2 Material and Methods

10.2.1 In silico evaluation of primers, primer pairs and combination of primer pairs

Primers were evaluated using the SILVA Reference database release 106 (April 2011) containing all sequences longer than 1,200 bases for *Bacteria* and *Eukarya* and above 900 bases for *Archaea* and an alignment quality value better than 50 [226]. All primers were resolved into wobble free oligos. A list of matches was retrieved via the probe match function of the ARB PT server [226]. No mismatches were allowed during probe matching. A canonical match position was derived at the alignment position at which the primer matched most frequently. Thereafter, the sequences where split into three sets: 1) a sequence is considered to be matched by the primer if one of its oligos matched the sequence exactly at the canonical match position; 2) a sequence is considered to have insufficient data if the alignment position of its first base is larger (or smaller for reverse primers) than the canonical match position of the primer; 3) all other sequences are considered to be not matched. The coverage of a taxon was computed as the matched fraction of sequences either matched or not matched.

In the evaluation of primer pairs, a sequence is considered matched if it matches both forward and reverse primers. A sequence is considered to have insufficient data if there were no matches for either forward or reverse primers. All scripts and SQL queries as well as database dumps and raw output data in CSV format are available in the supplementary materials.

In this study the term 'coverage' refers to the percentage of matches for a

given taxonomic path. The threshold was set at 50%. Thus, a coverage ≤50% refers to 'taxonomic path not covered'. Evaluation was carried out in order to find primers and primer pairs sensitive enough to match all target sequences (high sensitivity) and exclude all non target sequences (high specificity). The term 'phylum spectrum' refers to the number of matched phyla. For example, if a primer or primer pair is covering the majority of all phyla it will be described as 'large phylum spectrum'. In this study, he term 'universal primer' will be used to describe a primer targeting both, *Archaea* and *Bacteria*.

10.2.2 Selection criteria for primer pairs and combination of primer pairs suitable for 16S rRNA gene amplification using long range next generation sequencing methods like Roche's 454 pyrosequencing

Primer pairs were selected according to annealing temperature, coverage of variable regions and length of PCR fragments. Annealing temperatures were calculated with Oligo Calc[1]. Primers with annealing temperature differences of less than 5°C were accepted as pairs. Only primer pairs generating PCR-fragments with a minimum fragment length of 450 bases and a maximum read length of 1000 bases were chosen conforming to Roche's 454 pyrosequencing technology [59]. In addition, primers pairs generating fragments ≥1000 bp were chosen, which are of particular interest for the new emerging long-range, SMRT sequencing technologies [67].

Roche's 454 FLX Titanium System [59] was chosen as method of choice because of its longer reads length and comparatively low cost. New even long-ranging, SMRT sequencing technologies as provided by Pacific Bioscience (PacBio) were not commercially available at that time.

10.2.3 Sampling site and collection of water samples

Sample collection was done within in the MIMAS project[2]. Surface water was collected on 11th February 2009 and weekly from 31th of March 2009 till October 2009. Water samples (total volume of 360 L) from the Kabeltonne site at Helgoland Roads in the North Sea (54°11.18[2032?]N, 7°54[2032?]E) were collected at a depth of 0.5 m and processed immediately at the Biological Station of Helgoland. The water was pre-filtered through a 10 μm and 3

[1] `http://www.basic.northwestern.edu/biotools/oligocalc.html`
[2] `http://www.mimas-project.de`

µm pore-size filter. For sample collection a 0.2-μm-pore-size filter was used. Per sampling day eight filters for genomic DNA extraction were sampled containing biomass of 10 L and 15 L seawater respectively. For catalyzed reporter deposition FISH (CARD-FISH) procedure, 500 ml seawater was fixed in 1% paraformaldehyde and incubated for 4 hours at 4°C. Alternatively 1 h at room temperature. 0.2-μm-pore-size filter was used to filter three times 100 ml and 10 ml respectively. In the end, all filters were stored at -80°C until future usage.

10.2.4 DNA extraction

Genomic DNA was directly extracted from filter as described in Zhou et al. [318] with the following changes: all extraction steps were performed with 50 µl proteinase K (10 mg/ml), and after isopropanol precipitation the pellet of crude nucleic acids was obtained by centrifugation at 50,000 g for 30 min at room temperature.

The quantity and quality of the extracted DNA were analyzed by spectrophotometry using ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and by agarose gel electrophoresis. The genomic DNA was stored at -20°C until it was used for PCR amplification and metagenomics.

10.2.5 Amplification

Fragments of the bacterial 16S rRNA genes were amplified from the extracted DNA using the primer pair Bakt_341_F, 5'CCTACGGGNGGCWGCAG3'[111], and Bakt_805_R, 5'GACTACHVGGGTATCTAATCC3'[111]. The reaction was carried out in 50-ml volumes and contained 0.3 mg/ml BSA (Bovine Serum Albumin), 250 µM dTNPs, 0.5 µM of each primer, 0.02 U Phusion High-Fidelity DNA Polymerase (Finnzymes OY, Espoo, Finland) and 5x Phusion HF Buffer, which contains 1.5 mM $MgCl_2$. The PCR was run at the following cycling conditions: initial denaturation at 95°C for 5 min, followed by 25 cycles consisting of denaturation (95°C for 40 sec), annealing (55°C for 2 min) and extension (72°C for 1 min) and a final step at 72°C for 7 min. PCR products were purified with a QiaQuick PCR purification kit (QUIAGEN, Hilden, Germany). The quantity and quality of the extracted DNA were analyzed by spectrophotometry using ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and by agarose gel electrophoresis. The PCR products were stored at -20°C for future sequencing.

10.2.6 Sequencing

PCR amplified DNA fragments as well as genomic DNA for metagenome studies were sent for sequencing to LGC Genomics (Berlin, Germany). Sequencing was performed using Roche's GS-FLX 454 Titanium pyrosequencer (Roche, Mannheim, Germany). For PCR amplified DNA fragments 1/8 pico titer plate (PTP) was sequenced per sampling date. For metagenomics two full PTPs per sample were sequenced. Raw data was stored as FNA file. Sequences were submitted to XXX with accession numbers XXX.

10.2.7 Identification and taxonomic classification of 16S rRNA fragments

Unassembled sequence reads from both, SSU rRNA gene PCR amplicons and metagenome sequencing, were preprocessed (quality control and alignment) by the bioinformatics pipeline of the SILVA project [226]. Briefly, reads shorter than 200 nucleotides and with more than 2% of ambiguities or 2% of homopolymers, respectively, were removed. Remaining reads were aligned against the SSU rRNA seed of the SILVA database release 106[3] [226] whereupon non-aligned reads have not further been considered for downstream analysis. Using this strategy, contaminations/artifacts in the PCR amplicon pool could be identified as well as all putative SSU rRNA gene reads within the metagenomic data sets. Subsequently, reads of the filtered datasets were dereplicated, clustered and classified on a sample by sample basis. Dereplication (identification of identical reads ignoring overhangs) was done with cd-hit-est[4] [160] using an identity criterion of 1.00 and a wordsize of 8. Remaining sequences were clustered again with cd-hit-est using an identity criterion of 0.98 (same wordsize). The longest read of each cluster was used as a reference for taxonomic classification done by a local blastn search against the SILVA SSURef 106 NR dataset[5] using blast-2.2.22+[6] with standard settings. The full SILVA taxonomic path of the best blast hit has been assigned to the reads in case the value for (% sequence identity + % alignment coverage)/2 was at least 93.0. In the final step, the taxonomic path of each cluster reference read was mapped to the additional reads within the corresponding cluster plus the corresponding replicates, identified in the previous analysis step, to finally obtain quantitative information (number of individual reads representing a taxonomic path). All process data can be found in supplementary material.

[3] http://www.arb-silva.de/documentation/background/release-106

[4] http://www.bioinformatics.org/cd-hit

[5] http://www.arb-silva.de/projects/ssu-ref-nr/

[6] http://blast.ncbi.nlm.nih.gov/Blast.cgi

10.2.8 Catalyzed reporter deposition (CARD)-FISH

The CARD-FISH procedure was performed according to Pernthaler et al. [223]. Distribution of *Bacteroidetes, Alphaproteobacteria* and *Gammaproteobacteria* have been detected by probe CF319a [179], ALF968 [207] and GAM42a2 [180] (relative abundance, % of 4,6-diamidino-2-phenylindole (DAPI) counts) respectively.

10.3 Results and Discussion

10.3.1 *In silico* evaluation of 16S rDNA primers

We compiled and evaluated 54 forward and reverse primers described as 'universal' and/or recently used in bacterial and archaeal diversity analysis (Table 10.1). Primers were evaluated using the SILVA Reference database release 106 (April 2011) in order to find primers and primer pairs sensitive enough to match all target sequences (high sensitivity) and exclude all non target sequences (high specificity). The term 'coverage' refers to the percentage of matches for a given taxonomic path and is given for all three domains of life, although the focus of this study lies on *Archaea* (A) and *Bacteria* (B). In this study, the threshold was set at 50%. Thus, a coverage $\leq$ 50% refers to 'taxonomic path not covered'. The term 'phylum spectrum' refers to the number of matched phyla. For example, if a primer or primer pair is covering the majority of all phyla it will be described as 'large phylum spectrum'. In this study, he term 'universal primer' will be used to describe a primer targeting both, *Archaea* and *Bacteria*.

The evaluation revealed that the coverage of nine primers was $\leq$50% and for 18 primers between 50% and 79% for the domain *Archaea* and *Bacteria*, respectively (Table 10.1). For 27 primers coverage above 80% could be obtained. Additionally, twelve primers with coverage of $\leq$50% for 18S sequences of *Eukarya* (E) could be detected.

The majority of the analysed primers are domain specific (Table 10.1). However the primers Bakt_805_R (A: 92.6%; B: 91.9%), 806V_R (A: 92%; B: 91.8%), U1053_R (A: 60.1%, B: 51.9%) and GM4_R (A: 80.8%; B: 71.8%) show relatively high coverage for both, *Archaea* and *Bacteria*. Similar results could be obtained for U515_F, U519_F, Parch519_R, U529_R, Uni1392_R and UA1406_R with the difference that they also match a high amount of eukaryotic 18S rDNA gene sequences. The universality of U515_F and 806V_R has also been approved by Walters et al. [298]. The study of Wang and Qian [300] detected similar results for U515_F (A: 63.3%; B: 99.0%) but for U519F (A: 96.7%; B: 98.5%) much higher results were obtained.

Table 10.1 *Coverage rate of commonly used primers*

Primer	Reference	Sequence 5'-3'	Position*	Coverage rate (%)**		
				Archaea	*Bacteria*	*Eukarya*
bio-pB_R5'.SE_F	[124]	GAAGAGTTTGATCATGGCTCAG	6-27	0.0	6.9	0.4
GM3_F	[202]	AGAGTTTGATCMTGGC	8-23	0.0	**83.9**	0.3
Arch8_F	[278]	TCCGGTTGATCCTGCC	8-23	62.7	0.0	1.3
8_F	[81]	AGAGTTTGATCCTGGCTCAG	8-27	0.0	69.1	0.3
Arch20_F	[185]	TTCCGGTTGATCCYGCCRG	20-38	78.2	0.0	0.6
pB_R-V1.AS_F	rc [124]	AGTGGCGGACGGGTGAGTAA	101-120	0.0	20.3	0.0
Bakt_341_F	[111]	CCTACGGGNGGCWGCAG	341-358	0.2	**95.5**	0.0
347_F	[208]	GGAGGCAGCAGTRRGGAAT	347-365	0.0	86.9	0.0
Arch349_F	[273]	GYGCASCAGKCGMGAAW	349-365	81.5	0.0	0.0
U515_F	[298]	GTGCCAGCMGCCGCGGTAA	515-534	59.7	**96.3**	93.1
Arch516_F	[273]	TGYCAGCCGCCGCGGTAAHACCVGC	516-484	**84.7**	0.0	0.1
517_F	[300]	GCCAGCAGCCGCGGTAA	517-533	1.4	**96.6**	**93.1**
518_F	[201]	CCAGCAGCCGCGGTAAT	518-534	0.5	**89.1**	**90.1**
U519_F	[300]	CAGCMGCCGCGGTAATWC	519-537	54.4	**89.2**	**90.3**
A519_F	[300]	CAGCMGCCGCGGTAA	519-533	0.0	0.0	0.0
Ab779_F	[16]	GCRAASSGGATTAGATACCC	779-800	62.6	5.7	0.0
U779_F	[300]	GCTAASSGGATTAGATACCC	779-799	0.0	0.0	0.9
U789_F	[16]	TAGATACCCSSGTAGTCC	789-807	**88.1**	5.7	0.0
U906_F	[16]	GAAACTTAAAKGAATTG	906-923	**86.8**	0.4	77.3
Arch915_F	rc [263]	AGGAATTGGCGGGGGAGCAC	915-896	**83.1**	0.0	0.0
bio-pJBS-V3.SE_F	[124]	GCAACGCGAAGAACCTTACC	947-967	0.0	60.7	0.0
Arch958_F	rc [52]	AATTGGAKTCAACGCCGGR	938-958	50.6	0.0	0.0
Arch958B_F	in house	AATTGGABTCAACGCCGGR	938-958	68.9	0.0	0.0
967_F	[212, 261]	CAACGCGAAGAACCTTACC	967-985	0.0	60.8	0.0
968_F	[212]	AACGCGAAGAACCTTAC	968-984	0.0	60.9	0.0
U1053_F	[16]	GCATGGCYGYCGTCAG	1053-1068	42.0	51.9	0.0
pB_R-V1.AS_R	[124]	TTACTCACCCGTCCGCCACT	120-101	0.0	19.7	0.0
518_R	[201]	ATTACCGCGGCTGCTGG	534-518	0.4	**89.1**	**88.9**
A519_R	[300]	GGTDTTACCGCGGCKGCTG	537-519	**92.9**	0.1	0.8
Parch519_R	[214]	TTACCGCGGCKGCTG	533-519	**94.9**	**96.6**	**92.3**
U529_R	[16]	ACCGCGGCKGCTGGC	529-514	59.6	**97.4**	**92.3**
803_R	[208]	CTACCRGGGTATCTAATCC	803-785	8.8	**87.2**	0.0
805_R	[81]	GACTACCAGGGTATCTAAT	805-787	0.2	**82.6**	0.0
Bakt_805_R	[111]	GACTACHVGGGTATCTAATCC	805-785	**92.6**	**91.9**	0.0
A806_R	[300]	GGACTACVSGGGTATCTAAT	806-787	**89.8**	7.0	0.0
E806_R	[300]	GGACTACCAGGGTATCTAAT	806-787	0.2	**82.2**	0.0
806V_R	[298]	GGACTACHVGGGTATCTAAT	806-787	**92.0**	**91.8**	0.0
U906_R	rc [16]	CAATTCMTTTAAGTTTC	923-906	**86.8**	0.4	77.3
Arch915_R	[263]	GTGCTCCCCCGCCAATTCCT	915-896	**83.1**	0.0	0.0
907_R	[200]	CCGTCAATTCMTTTGAGTTT	926-907	0.0	**91.0**	0.1
bio-pJBS-V3.SE_R	rc [124]	GGTAAGGTTCTTCGCGTTGC	967-947	0.0	60.7	0.0
Arch958_R	[52]	YCCGGCGTTGAMTCCAATT	958-940	52.0	0.0	0.0
Arch958V_R	in house	YCCGGCGTTGAVTCCAATT	958-940	70.7	0.0	0.0
967_R	rc [52, 261]	GGTAAGGTTCTTCGCGTTG	985-967	0.0	60.8	0.0
B-V3.AS_R	[124]	ACGACAGCCATGCAGCACCT	1047-1027	0.0	39.4	0.0
1046_R	[261]	CGACAGCCATGCANCACCT	1046-1028	0.0	49.8	0.0
U1053_R	rc [16]	CTGACGRCRGCCATGC	1068-1053	60.1	51.9	0.0
GM12_R	[125]	CGTCATCCMCACCTTCCTC	1193-1175	0.0	62.6	0.0
Uni1390_R	[317]	GACGGGCGGTGTGTACAA	1390-1373	5.7	65.9	**94.4**
Uni1392_R	[150]	ACGGGCGGTGTGTRC	1392-1378	**80.2**	79.5	**95.2**
1401_R	[212]	CGGTGTGTACAAGACCC	1401-1385	0.0	14.1	0.0
UA1406_R	[151]	ACGGGCGGTGTGTRCAA	1406-1390	65.2	72.0	**94.6**
Arch1492_R	[278]	GGCTACCTTGTTACGACTT	1492-1474	42.8	35.1	0.9
GM4_R	[202]	TACCTTGTTACGACTT	1507-1492	**80.8**	71.8	5.8

The names of forward and reverse primer end with '_F' and '_R' respectively. 'U' as well as 'Uni' refers to known universal primer. The dashed line separates forward and reverse primer. *Numbering based on the *Escherichia coli* system of nomenclature [28]; **Evaluation is based on SILVA Reference database 106 [226]; *in house*: in house modification of corresponding primer; *rc*: reverse complement of corresponding forward and reverse primer respectively; bold numbers: coverage ≥ 80%.

Our evaluation cast doubt on the accuracy of some universal primers such as U779_F (A: 0.0%; B: 0.0%), U789_F (A: 81.1%; B: 5.7%), U906_F (A: 86.8%; B: 0.4%), U906_R (A: 86.8%; B: 0.4%) and Uni1390_R (A: 5.7%; B: 65.9%). In comparison with the study of Wang et al. (2009), major discrepancies arose. With 89% archaeal and 5% bacterial coverage U779_F was claimed to be specific for *Archaea* but according to our results this primer failed to detect this domain. However, because Ab779_F and U779_F differ only in one base, we believe that a simple spelling mistake might be the reason for this discrepancy. Furthermore U789_F and U906_F have previously been asserted to be universal for *Bacteria* and *Archaea* [300], but with only 5.7% and 0.4% coverage rate for *Bacteria*, respectively, our evaluation revealed a high bias against *Bacteria*. Our results rather confirm the original purpose of Ab779_F [29], 789_F [19] and 906_F [233], which were primarily designed as archaeal primer.

The highest coverage and specificity for the domain *Bacteria* could be detected for the forward primers GM3_F (83.9%), Bakt_341_F (95.5%) and 347_F (86.9%) and reverse primers 803_R (87.2%), 805_R (82.6%) and 907_R (91.0%). However, high coverage of a single primer does not automatically result in a wide phylum spectrum as revealed by the detailed analysis (see Supplementary Table 2 online). For example, if at least half of the sequences in a phylum need to be targeted by the primer, 907_R discriminates 12 out of 59 phyla such as *Chlamydiae* and *Verrucomicrobia* as well as the class *Epsilonproteobacteria*. In terms of phylum spectrum, the best results could be obtained for GM3_F. Although with 83.9% it did not reveal the highest coverage rate, this primer excels by discriminating only *Chlamydiae*, *Dictyoglomi* and *WCHB1-60*. This clearly shows that overall coverage and analysis of phylum spectrum needs to be taken into account for primer evaluation.

For the domain *Archaea*, the primer with the best values and domain specificity are Arch516_F (84.7%), U789_F (88.1%) and U906_F (86.8%) as well as A519_R (92.9%), A806_R (89.8%) and U906_R (86.8%). The high coverage of A519_R has also been reported by Wang and Qian [300]. Detailed analysis of the six best primers showed that all of them cover the two main archaeal phyla *Crenarchaeota* and *Euryarchaeota* (see Supplementary Table 1 online). However in terms of phylum spectrum the best results could be obtained for A519_R. This primer is covering seven out of eight achaeal phyla. Only *Nanoarchaeota* is discriminated, which is not surprising, since the majority of *Archaea* specific primers were designed prior to the discovery of the *Nanoarchaeota* [16].

Although specificity of primers has been analysed in previous studies our results are complementary or provide more details that will help to select an appropriate primer or primer pair. A direct comparison of our results with the

studies of Huws et al. [123] and Baker et al. [16] is not possible, because in the respective studies the coverage rates of the primers were not given. Nossa et al. [208] restricted their evaluation to a single habitat. Walters et al. [298] analysed only four primers. In Wang and Qian [300], phyla containing less than 100 sequences were not taken into account and their evaluation was relaxed by allowing a single mismatch, which can explain the discrepancies to our results. Although, it is assumed that a standard PCR can tolerate up to two mismatches between primer and its target [208], it has been shown that a primer mismatch results in a biased picture of the bacterial diversity [256]. Preferential amplification might lead to under-representation of important members of a community [16, 256]. Consequently, in our study no mismatch was allowed and all phyla as well as the *Proteobacteria* classes were included.

10.3.2 *In silico* evaluation of primer pairs for long-range next generation sequencing methods

The evaluation in the previous section revealed suitable primers for 16S rDNA gene amplification. When both forward and reverse primers are needed, the bias of the used primers accumulates. To minimize total bias, primers with similar bias pattern must be used. Therefore, the careful choice of primer pair is a crucial step to ensure minimum-biased results. In this study, we evaluated the combined coverage of 150 possible primer pairs for the three domains of life (Table 10.2).

Primarily, a set of 96 primers pairs generating fragments between 400-1000 bases, which is optimal for Roche's 454 FLX Titanium machines, has been analysed in more detail (Table 10.2). 34 of those primer pairs showed coverage above 50% for *Archaea* and *Bacteria*, respectively. None achieved more than 86%.

No primer pair could be found to qualify as universal for *Bacteria* and *Archaea*. Only primer pairs 519_R/UA1406_R (A: 34.4%, B: 65.3%), U519_F/Uni1392_R (A: 42.7%, B: 72.1%) and U519_F/Arch1492_R (A: 30.3%, B: 31.6%) target *Archaea* and *Bacteria*, but the combined coverage are all below our threshold of 50%. Thus we recommend amplifying archaeal and bacterial 16S rDNA sequences separately.

For the domain *Archaea*, Arch20_F/Parch519_R (75.9%) and Arch20_F/A519_R (74.9%) revealed relatively high overall coverage. A more detailed analysis of these primer pairs showed that all main and sequence-rich phyla are covered (see Supplementary Table 3 online). Again, only the phylum *Nanoarchaeota* is discriminated. Although the primer U1053F and U1053_R cover 75% of the known *Nanoarchaeota* sequences (see Supplementary Table 1

online), no suitable primer pair could be found to address this missing phyla (see Supplementary Table 3 online). However, if known *Nanoarchaeota* specific primers [119] are added, any of those two primer pairs could be recommended for archaeal 16S rDNA amplification.

For the domain *Bacteria*, the best results could be obtained for the pairs GM3_F/907_R (76.4%), GM3_F/806V_R (78.1%), Bakt_341_F/Bakt_805_R (88.1%), Bakt_341_F/803_R (84.3%), 347_F/Bakt_805_R (81.5%) and 341_F/803_R (78.1%). However, detailed analysis of the targeted phyla showed that several phyla are discriminated (see Supplementary Table 3 online). For example, GM3_F/907_R as well as GM3_F/806V_R neglects EM19, *Chlamydiae*, *Dictyoglomi* and *Candidate division OP11*. Furthermore, 347_F/Bakt_805_R and 341_F/803_R fail to detect e.g. *Verrucomicrobia*, *Planctomycetes* and *Chloroflexi*. Unfortunately all other 97 primer pairs tested showed similar insufficiencies (see Supplementary Table 4 online). The primer pair Bakt_341_F/Bakt_805_R with the highest coverage neglects only 8 out of 59 phyla like *Chloroflexi*, *Candidate division WS6* and *EM19*. Although complete phylum spectrum could not be accomplished, this primer pair shows relatively good results for domain and phylum coverage. These findings are in line with the conclusion of Baker and Cowan [15]. They claim that no primer of sufficient length exists or can be designed that matches all bacterial 16S rDNA sequences. However, additional primer pair can be used to complement the missing phyla (see Supplementary Table 4 online). For example, GM3_F/806V_R e.g. would cover five additional groups including *Candidate division WS6* and *OP10*.

Table 10.2 *Coverage rate of selected primer pairs*

Forward Primer	Reverse Primer	Fragment length (bp)	Covered HV regions*	coverage rate (%)**		
				Archaea	*Bacteria*	*Eukarya*
GM3_F	805_R	797	1-4	0,0	**70,8**	0,0
GM3_F	907_R	918	1-5	0,0	**76,4**	0,0
GM3_F	U906_R	915	1-5	0,0	0,3	0,3
GM3_F	967_R	977	1-5	0,0	**54,0**	0,0
GM3_F	A806_R	798	1-4	0,0	5,0	0,0
GM3_F	806V_R	798	1-4	0,0	**78,1**	0,0
Bakt_341_F	Bakt_805_R	464	3-4	0,2	**88,1**	0,0
Bakt_341_F	GM12_R	852	3-7	0,0	**60,7**	0,0
Bakt_341_F	bio-pJBS-V3,SE_R	626	3-5	0,0	**58,4**	0,0
Bakt_341_F	B-V3,AS_R	705	3-6	0,0	**38,2**	0,0
Bakt_341_F	U1053_R	727	3-6	0,1	**49,9**	0,0
Bakt_341_F	1046_R	705	3-6	0,0	**47,9**	0,0
Bakt_341_F	803_R	462	3-4	0,0	**84,3**	0,0
Bakt_341_F	Arch_915_R	574	3-5	0,2	0,0	0,0
Bakt_341_F	Arch958_R	617	3-5	0,1	0,0	0,0
U519_F	UA1406_R	887	3-8	34,4	**65,3**	**86,5**
U519_F	GM12_R	674	4-7	0,0	**56,1**	0,0
U519_F	1401_R	882	4-8	0,0	13,5	0,0
U519_F	B-V3,AS	528	4-6	0,0	35,9	0,0

Table continued on next page

Table 10.2 *Coverage rate of selected primer pairs*

Forward Primer	Reverse Primer	Fragment length (bp)	Covered HV regions*	coverage rate (%)**		
				Archaea	*Bacteria*	*Eukarya*
U519_F	Uni1390_R	871	4-8	3,7	**59,7**	**86,3**
U519_F	Uni1392_R	873	4-8	42,7	**72,1**	**87,0**
U519_F	Arch1492_R	973	4-9	30,3	31,6	0,7
518_F	UA1406_R	888	4-8	0,4	**65,2**	**86,3**
518_F	GM12_R	675	4-7	0,0	**56,0**	0,0
518_F	1401_R	883	4-8	0,0	13,5	0,0
518_F	B-V3,AS_R	529	4-6	0,0	35,9	0,0
518_F	Uni1390_R	872	4-8	0,1	**59,7**	**86,2**
518_F	Uni1392_R	874	4-8	0,5	**72,0**	**86,8**
518_F	Arch1492_R	974	4-9	0,6	31,5	0,7
8_F	Bakt_805_R	797	1-4	0,0	**64,6**	0,0
8_F	518_R	510	1-3	0,0	0,0	0,0
8_F	907_R	918	1-5	0,0	**62,7**	0,0
8_F	bio-pJBS-V3,SE_R	959	1-5	0,0	45,1	0,0
8_F	803_R	795	1-4	0,0	**61,0**	0,0
8_F	U529_R	521	1-3	0,0	**67,4**	0,3
8_F	Arch958_R	950	1-5	0,0	0,0	0,0
8_F	Arch958V_R	950	1-5	0,0	0,0	0,0
bio-pBR5',SE_F	Bakt_805_R	799	1-4	0,0	6,8	0,0
bio-pBR5',SE_F	GM1_R	512	1-3	0,0	0,0	0,0
bio-pBR5',SE_F	907_R	920	1-5	0,0	6,8	0,0
bio-pBR5',SE_F	bio-pJBS-V3,SE_R	961	1-5	0,0	6,6	0,0
bio-pBR5',SE_F	967_R	979	1-5	0,0	6,6	0,0
bio-pBR5',SE_F	803_R	797	1-4	0,0	6,8	0,0
bio-pBR5',SE_F	Arch958_R	952	1-5	0,0	0,0	0,0
bio-pBR5',SE_F	Arch958R-V	952	1-5	0,0	0,0	0,0
pBR-V1,AS_F	bio-pJBS-V3,SE_R	866	2-5	0,0	17,3	0,0
pBR-V1,AS_F	B-V3,AS_R	946	2-6	0,0	6,2	0,0
pBR-V1,AS_F	1046_R	945	2-6	0,0	8,0	0,0
pBR-V1,AS_F	Arch915_R	814	2-5	0,0	0,0	0,0
pBR-V1,AS_F	Arch958_R	857	2-5	0,0	0,0	0,0
pBR-V1,AS_F	Arch958V_R	857	2-5	0,0	0,0	0,0
U779_F	UA1406_R	627	4-8	0,0	0,0	0,9
U779_F	Uni1390_R	611	4-8	0,0	0,0	0,9
U779_F	Uni1392_R	613	4-8	0,0	0,0	0,9
U779_F	Arch1492_R	713	4-9	0,0	0,0	0,0
Ab779_F	UA1406_R	627	4-8	39,2	4,0	0,0
Ab779_F	Uni1390_R	611	4-8	2,0	3,8	0,0
Ab779_F	Uni1392_R	613	4-8	48,3	4,4	0,0
Ab779_F	Arch1492_R	713	4-9	16,6	1,5	0,0
U789_F	UA1406_R	617	4-8	**60,1**	4,0	0,0
U789_F	Uni1390_R	601	4-8	5,3	3,8	0,0
U789_F	Uni1392_R	603	4-8	**73,9**	4,4	0,0
U789_F	Arch1492_R	703	4-9	39,2	1,5	0,0
967_F	GM4_R	540	6-9	0,0	41,4	0,0
347_F	Bakt_805_R	458	3-4	0,0	**81,2**	0,0
347_F	GM12_R	846	3-7	0,0	**56,3**	0,0
347_F	bio-pJBS-V3,SE_R	620	3-5	0,0	**54,1**	0,0
347_F	B-V3,AS_R	700	3-6	0,0	35,1	0,0
347_F	U1053_R	721	3-6	0,0	44,7	0,0
347_F	1046_R	698	3-6	0,0	43,5	0,0
347_F	803_R	456	3-4	0,0	**78,1**	0,0
347_F	Arch915_R	568	3-5	0,0	0,0	0,0

Table continued on next page

Table 10.2 *Coverage rate of selected primer pairs*

Forward Primer	Reverse Primer	Fragment length (bp)	Covered HV regions*	coverage rate (%)**		
				Archaea	*Bacteria*	*Eukarya*
347_F	Arch958_R	611	3-5	0,0	0,0	0,0
Arch8_F	U529_R	521	1-3	32,1	0,0	1,3
Arch8_F	A519_R	529	1-3	**60,5**	0,0	0,0
Arch8_F	Parch519_R	525	1-3	**61,3**	0,0	1,3
Arch8_F	Arch915_R	907	1-5	48,3	0,0	0,0
Arch8_F	Arch958_R	950	1-5	31,4	0,0	0,0
Arch8_F	Arch958V_R	950	1-5	40,9	0,0	0,0
Arch349_F	Arch915_R	566	3-5	**69,8**	0,0	0,0
Arch349_F	Arch958_R	609	3-5	43,0	0,0	0,0
Arch349_F	Arch958V_R	609	3-5	**58,7**	0,0	0,0
Arch915_F	Uni1390_R	494	6-8	4,5	0,0	0,0
Arch915_F	Uni1392_R	496	6-8	**69,9**	0,0	0,0
Arch915_F	Arch1492_R	596	6-9	39,5	0,0	0,0
Arch958_F	Uni1390_R	452	6-8	1,9	0,0	0,0
Arch958_F	Uni1392_R	454	6-8	34,6	0,0	0,0
Arch958_F	Arch1492_R	554	6-9	13,6	0,0	0,0
Arch958 B _F	Uni1390_R	452	6-8	5,2	0,0	0,0
Arch958 B _F	Uni1392_R	454	6-8	**59,3**	0,0	0,0
Arch958 B _F	Arch1492_R	554	6-9	36,2	0,0	0,0
Arch20_F	A519_R	517	1-3	**74,9**	0,0	0,0
Arch20_F	Parch519_R	513	1-3	**75,9**	0,0	0,6
Arch 20_F	Arch915_R	895	1-5	**60,0**	0,0	0,0
Arch 20_F	Arch958_R	938	1-5	43,6	0,0	0,0
Arch20_F	Arch958V_R	938	1-5	**53,8**	0,0	0,0
bio-pB_R5',SE_F	B-V3,AS_R	1041	1-6	0,0	6,7	0,0
bio-pB_R5',SE_F	U1053_R	1062	1-6	0,0	6,8	0,0
bio-pB_R5',SE_F	GM12_R	1187	1-7	0,0	5,7	0,0
bio-pB_R5',SE_F	Uni1390_R	1384	1-8	0,0	7,0	0,5
bio-pB_R5',SE_F	Uni1392_R	1386	1-8	0,0	7,0	0,5
bio-pB_R5',SE_F	1401_R	1395	1-8	0,0	0,1	0,0
bio-pB_R5',SE_F	UA1406_R	1400	1-8	0,0	7,0	0,5
bio-pB_R5',SE_F	Arch1492_R	1486	1-9	0,0	1,4	0,0
GM3_F	GM4_R	1499	1-9	0,1	**74,9**	0,0
Arch8_F	B-V3,AS_R	1039	1-6	0,0	0,0	0,0
Arch8_F	U1053_R	1060	1-6	34,4	0,0	0,0
Arch8_F	GM12_R	1185	1-7	0,0	0,0	0,0
Arch8_F	Uni1390_R	1382	1-8	5,3	0,0	1,4
Arch8_F	Uni1392_R	1384	1-8	**60,8**	0,0	1,4
Arch8_F	1401_R	1393	1-8	0,0	0,0	0,0
Arch8_F	UA1406_R	1398	1-8	**60,6**	0,0	1,4
Arch8_F	Arch1492_R	1484	1-9	31,3	0,0	0,2
8_F	B-V3,AS_R	1039	1-6	0,0	20,2	0,0
8_F	U1053_R	1060	1-6	0,0	26,9	0,0
8_F	GM12_R	1185	1-7	0,0	40,7	0,0
8_F	Uni1390_R	1382	1-8	0,0	39,6	0,3
8_F	Uni1392_R	1384	1-8	0,0	**51,3**	0,3
8_F	1401_R	1393	1-8	0,0	9,6	0,0
8_F	UA1406_R	1398	1-8	0,0	**50,6**	0,3
8_F	Arch1492_R	1484	1-9	0,1	26,6	0,0
Arch20_F	B-V3,AS_R	1039	1-6	0,0	0,0	0,0
Arch20_F	U1053_R	1060	1-6	41,8	0,0	0,0
Arch20_F	GM12_R	1185	1-7	0,0	0,0	0,0
Arch20_F	Uni1390_R	1382	1-8	8,9	0,0	0,6

Table continued on next page

Table 10.2 *Coverage rate of selected primer pairs*

Forward Primer	Reverse Primer	Fragment length (bp)	Covered HV regions*	coverage rate (%)**		
				Archaea	*Bacteria*	*Eukarya*
Arch20_F	Uni1392_R	1384	1-8	**61,9**	0,0	0,6
Arch20_F	1401_R	1393	1-8	0,0	0,0	0,0
Arch20_F	UA1406_R	1398	1-8	**61,7**	0,0	0,6
Arch20_F	Arch1492_R	1484	1-9	31,3	0,0	0,1
pB_R-V1,AS_F	Uni1390_R	1289	2-8	0,0	10,9	0,0
pB_R-V1,AS_F	Uni1392_R	1291	2-8	0,0	14,4	0,0
pB_R-V1,AS_F	1401_R	1300	2-8	0,0	3,2	0,0
pB_R-V1,AS_F	UA1406_R	1305	2-8	0,0	13,0	0,0
pB_R-V1,AS_F	Arch1492_R	1391	2-9	0,0	5,3	0,0
pB_R-V1,AS_F	GM12_R	1092	2-7	0,0	10,6	0,0
Bakt_341_F	Uni1390_R	1289	3-8	0,0	**63,4**	0,0
Bakt_341_F	Uni1392_R	1291	3-8	0,2	**76,2**	0,0
Bakt_341_F	1401_R	1300	3-8	0,0	13,6	0,0
Bakt_341_F	UA1406_R	1305	3-8	0,2	**69,0**	0,0
Bakt_341_F	Arch1492_R	1391	3-9	0,1	34,0	0,0
347_F	Uni1390_R	1289	3-8	0,0	**58,7**	0,0
347_F	Uni1392_R	1291	3-8	0,0	**70,4**	0,0
347_F	1401_R	1300	3-8	0,0	12,4	0,0
347_F	UA1406_R	1305	3-8	0,0	**63,8**	0,0
347_F	Arch1492_R	1391	3-9	0,0	32,1	0,0
Arch349_F	Uni1390_R	1289	3-8	3,2	0,0	0,0
Arch349_F	Uni1392_R	1291	3-8	**66,9**	0,0	0,0
Arch349_F	1401_R	1300	3-8	0,0	0,0	0,0
Arch349_F	UA1406_R	1305	3-8	**53,9**	0,0	0,0
Arch349_F	Arch1492_R	1391	3-9	39.2	0,0	0,0

The names of forward and reverse primer end with '_F' and '_R' respectively, 'U' as well as 'Uni' refers to known universal primer, HV = hyper variable; The line separates primer pairs generating PCR fragments >1000 bp in length from the others; *Positions of hypervariable regions span nucleotides 69-99, 137-242, 433-497, 576-682, 822-879, 986-1043, 1117-1173, 1243-1294 and 1435-1465 for V1 through V9 respectively and are based on the E,coli system of nomenclature [59]; **Evaluation is based on SILVA Reference database 106 [226]; bold numbers: coverage ≥ 50%,

Additionally, a set of 54 primers that generate PCR fragments >1000 bp in length were tested *in silico*. These primer pairs could be of particular interest for classical, clone library based diversity analysis and the upcoming third generation SMRT technology. 15 of the 54 primer pairs (Table 10.2) showed coverage above 50% for *Archaea* and *Bacteria*, respectively. Again, no primer pair could be detected to qualify as universal.

For *Archaea*, no combination covers all nine HV regions in combination with high coverage. However, with 66.9% Arch349_F/Uni1392_R sticks out. This pair generates fragments of 1291 bp in length and spans HV regions 3-8. However, detailed analysis revealed that only the two sequence-rich phyla, *Crenarchaeota* and *Euryarchaeota*, are covered (see Supplementary Table 3 online). Despite lower domain coverage Arch20_F/Uni1392_R (61.4%) and Arch20nano_F (61.4%) only neglect *Marine Hydrothermal Vent Group 1 (MHVG-1)* and *Nanoarchaeota*. However, the high coverage and almost complete phylum spectrum of Arch20_F/Parch519_R and Arch20nano_F/Parch519_R could

not be achieved.

For the domain *Bacteria*, the highest coverage could be obtained for GM3_F/GM4_R (74.9%) and Bakt_341_F/Uni1392_R (76.2%). However, detailed analysis revealed a larger phylum spectrum for GM3_F/GM4_R (see Supplementary Table 4 online). This pair neglects only seven phyla, including *Chlamydiae*, *Candidate division WS6* and *Dictyoglomi*. Bakt_341_F/Bakt_805_R gained a similar wide phylum spectrum. GM3_F/GM4_R also generates almost full length 16S rDNA sequences and covers all HV regions. Again, additional primer pairs can be used to target the missing groups. For example, Bakt_341_F/UA1406_R covers five out of the seven missing phyla (see Supplementary Table 4 online).

This detailed evaluation of primers and primer pairs also demonstrates that a relatively good reverse and forward primer do not automatically result in a good pair. For instance, 967_F and GM4_R cover 60.8% and 71.8% of the domain *Bacteria*, respectively. For the pair 967_F/GM4_R the combined coverage rate decreases to 41.4%. Another example are the primers 347_F (86.9%) and 803_R (87.2%), which have been designed and approved by the Human Microbiome Project for analysing the foregut microbiome [208]. Based on the promising results for the human habitat, Nossa et al. [208] suggests that this pair may be a good candidate to access the bacterial diversity in any habitat. With 79.1% combined coverage for 347_F/803_R and discrimination of 37 bacterial phyla (see Supplementary Table 4 online) we are not in favour of this recommendation.

Furthermore we like to note that *in silico* evaluation studies, including this study, are perforce limited by the diversity of 16S rDNA sequences represented in the public repositories. Although we believe that the SILVA datasets provide one of the most comprehensive set of high quality rDNA sequence data currently available, it is not likely that they represent the 'real' microbial diversity. The majority of the sequences in the public databases are a result of prior PCR amplification, and taking our data into account, a bias must be expected. As soon as more metagenome studies become available re-evaluation of the primers is critical. Technically it has to be noted that the gained values for phylum coverage are affected by the numbers of sequences present in a phylum. If the majority of a small phylum is targeted, the coverage rate will probably be higher than for a member rich phylum. Similar effects occur for the phyla where only a small number of sequences contain full sequence information at the primer position of interest.

10.3.3 Experimental evaluation of GM3F/907R in combination with Bakt_341F/Bakt_805F

The primer pair Bakt_341_F/Bakt_805_R was applied on DNA extracted from a time series of four marine environmental samples at Helgoland Roads. For simplification we will refer to '16S tags' in this study. In addition PCR independent metagenomic studies and quantitative CARD-FISH analysis of the same samples were performed for comparison.

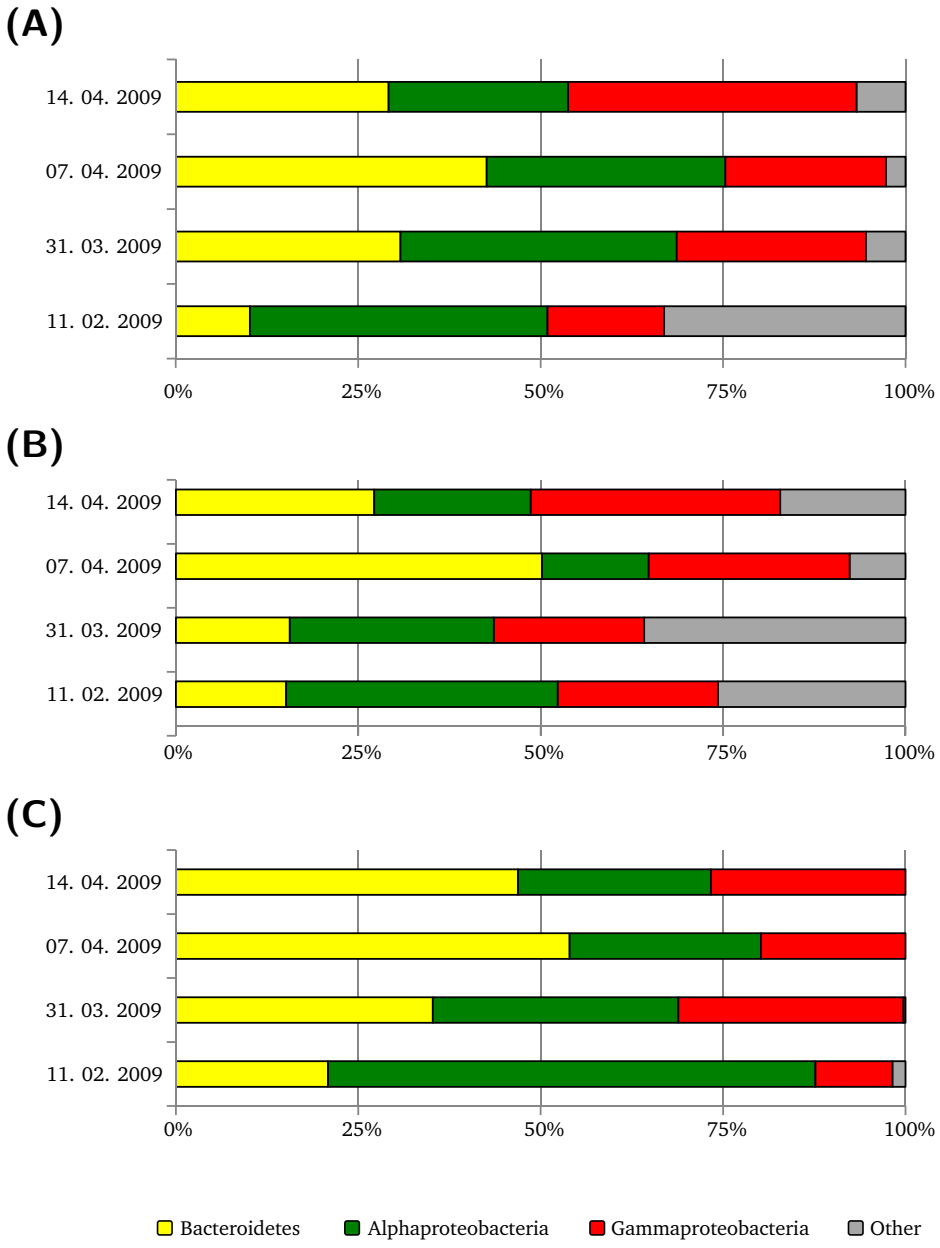
The results of the 16S tag analysis showed that the bacterial community is dominated by *Alphaproteobacteria*, *Bacteroidetes* and *Gammaproteobacteria* (Fig. 10.1A). *Bacteroidetes* appear to most abundant on 07.04.09 and 14.04.09 with a clear peak on 07.04.09. In contrary the abundance of *Gammaproteobacteria* increases relatively on the 14.04.09. *Alphaproteobacteria* appear to be most abundant in winter on the 11.02.09. The same trends could be obtained from the metagenome (Fig. 10.1B) and CARD-FISH studies (Fig. 10.1C), although small variations occurred. A quick evaluation using the TestProbe tool of the SILVA ribosomal RNA (rRNA) database project (`http://www.arb-silva.de/search/testprobe/`) reveals that the ALF968 FISH probe targets about 6,400 nearly full length sequences (>1200 bp) outside the alpha subclass of *Proteobacteria*. This can explain the overrepresentation of *Alphaproteobacteria* in the CARD-FISH results.

Additionally, 16S tag analysis provides an enhanced resolution up to the group or genus level. As an example, the distribution of the six most abundant taxonomic groups and genera have been examined in more detail (see Supplementary Fig. 2A online). Noticeable is for example the *Formosa* peak which goes along with the *Bacteroidetes* peak on the 07.04.09 (Fig. 10.1A) or that *Reinekea* has only been detected on 14.04.09. The same trend was confirmed by 16S rDNA studies from the corresponding metagenomes (see Supplementary Fig. 2B online). This demonstrates that Bakt_341_F/Bakt_805_R are able to provide a mostly unbiased picture of the bacterial diversity down to genus and group level.

10.4 Conclusion

16S rDNA analysis based on PCR is still the standard technology for the cultivation independent determination of bacterial and archaeal diversity. However, bias in phylogenetic analysis can occur through suboptimal choice of primer pairs. This could lead to discrimination or under representation of important members of the microbial community. We claim that the results of our study can guide the selection of optimal primer combinations, by providing coverage rates for all analysed primer and primer pairs for *Archaea* and *Bacteria* down to

Figure 10.1 *Taxonomic distribution of 16S rRNA gene sequences gained from four different surface water samples at Helgoland Roads in the North Sea. (A) 16S tags generated from PCR and sequenced with 454 pyrosequencing (relative abundance, % of total counts) (B) 16S tags gained from metagenome studies (relative abundance, % of total counts) (C) Results from catalyzed reporter deposition (CARD)-FISH studies, Distribution of* Bacteroidetes*,* Alphaproteobacteria *and* Gammaproteobacteria *as detected by probes CF319a, ALF968 and GAM42a2 (relative abundance, % of 4,6-diamidino-2-phenylindole (DAPI) counts), respectively.*



genus level (see Supplementary Table 1-4 online). Especially the evaluation of primer pairs can be used as a tool to find the most appropriate pair for specific research questions. However, with regard to comparability of results between studies and taking into account the high rate of unknown biodiversity in natural habitats, we think that a standard set of primers should be favoured over sets of habitat specific primers pairs.

For *Archaea*, primer pairs Arch20_F/Parch519_R and Arch20_F/A519_R stick out with high values and large phylum spectrum. But Nanoarchaeota specific primers [119] need to be added additionally to access the whole archaeal diversity. For the domain *Bacteria*, we recommend the primer pair Bakt_341_F/Bakt_805_R. Based on the computational and experimental analysis, we believe that this combination of primer pairs can be seen as a suitable set to successfully reveal the complexity of bacterial diversity using state of the art next-generation sequencing methods as it is currently represented by Roche's 454 FLX Titanium pyrosequencing method. It is interesting to note that GM3_F/GM4_R, which are commonly used primers for classical clone library based diversity analysis, will be most probably the primers of choice for the upcoming third generation SMRT technology. *In silico* evaluation shows that this combination provides a relatively unbiased picture of the bacterial diversity. However experimental analysis of this primer pair with respect to PCR-free methods is necessary. For *Archaea*, no suitable set of primers generating full length sequences could be verified, indicating the need for new optimized achaeal primer pairs.

Part IV

Concluding Discussion

CHAPTER 11

Summary

In Chapters 4 and 5 we have presented the SILVA project as it was originally published and the developments that have been made since then. With the SILVA pipeline we have created a software system for automated construction of rRNA gene databases. The system uses offline batch processing and a relational database management system (RDBMS) for persisting structured data. Volume sequence data is processed and filtered using a tool chain composited from both proprietary solutions and publicly available software. Descriptive data from multiple upstream sources can be incorporated into the resulting database. With the SILVA website we have created an easily accessible interface for querying the rRNA gene databases built in a regular schedule using the SILVA pipeline. The database can be browsed by taxonomic hierarchy and searched or filtered using string and numeric matching on combinations of descriptors. Both the full databases and custom subsets defined via search and browser can be downloaded in a variety of exchange formats. A facility for aligning and classifying user submitted sequences exist, which may also be used for sequence based search of the SILVA databases. The TestProbe tool allows evaluating probe specificity and sensitivity based on the full SILVA databases and a target group defined via browse and search mechanisms.

The alignment problem has been addressed by the development of the alignment tool SINA described in Chapter 6. The concept of using a *k*mer distance search to select closely related sequences from a large database of curated, mutually aligned sequences and transferring that alignment onto the target sequences using partial order alignment has been shown to work reliably. Compared to its most direct competitors, mothur and PyNAST, SINA is slower, yet more accurate. SINA also implements a fast sequence search that relies on the computed alignment. The number of pairwise alignment comparisons made during the search can be optionally reduced to the top results of a *k*mer based search. A lowest common ancestor (LCA) classification can be derived from the classifications of the search results.

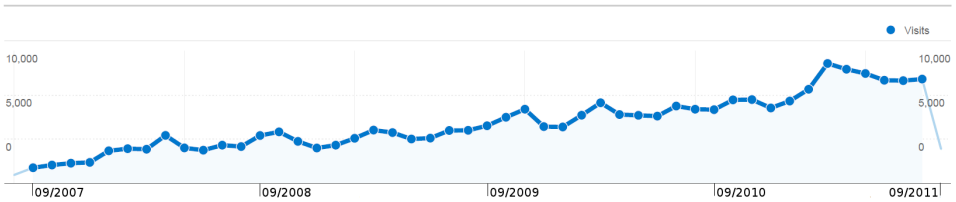


Figure 11.1 *SILVA adoption: Development of monthly visits on www.arb-silva.de according to google analytics.*

The ARB software suite for sequence analysis described in Chapter 7 has been enhanced to allow the use of the complete SILVA databases. The scalability issues caused by the limitation to 32-bit address space and the lack of modeling multiple SSU genes per accession number have been resolved. Well tested production releases of the ARB software have been made available. The entry barrier for novice users has been lowered by separating functions targeted at expert users from commonly required features. Several other improvements were made to improve user experience.

Based on these tools and databases and the experiences gained in their development, three flanking efforts have been completed. The MIMARKS standard described in Chapter 9 sets a baseline for the descriptors that should be included with newly sequenced marker genes. Extended use of the large subunit rRNA gene (LSU) in addition or alternative to the small subunit rRNA gene (SSU) has been promoted by the comparative case study described in Chapter 8. Suitable primer combinations for biodiversity studies relying on PCR amplicon sequencing have been determined and evaluated in Chapter 10.

The assessment that there is a high demand for a rRNA gene database such as SILVA and our success in meeting the requirements of dependent studies can be quantified bibliometrically and through website usage statistics. The SILVA project has been referenced by 529 studies [1]. The SILVA website currently receives well over a thousand visits per week (Fig. 11.1. Over the past six months, the alignment, custom export and TestProbe features have been used on average 338, 272 and 115 times per week, respectively.

[1] according to google scholar (accessed on Oct. 25th, 2011)

CHAPTER 12

Discussion

12.1 Infrastructure in Science

In this thesis, the focus lay on devising more comfortable and efficient ways to enact pre-existing, well established work-flows. This approach was motivated by the potential perceived in reducing the manual effort involved in these methods. Resources freed by eliminating repetitive tasks can be invested in other research activities. The work-flows can also be extended to larger volumes of primary data, thereby improving result significance and allowing work-flow application to scenarios previously intractable. The abstract means chosen to achieve overall effort reduction were automation and centralization. Automation, as implemented via a bioinformatics pipeline comprising tools for sequence preparation and data integration, resolves most repetitive tasks. Centralization reduces the overall effort incurred by tasks not easily automated. Such tasks include quality assurance, which always requires expert inspection, and large scale phylogenetic tree reconstruction, which is very compute intensive and must itself undergo quality assurance.

Centralization was achieved by providing processed data as a database service, rather than providing the means to build the database. This is a time honored approach which has been applied to art and literature throughout history by library institutions. Yet, in the position of the librarian, who builds, maintains, curates and indexes library collections, who cares for the library infrastructure and who acts as a teacher and a guide to those who would use the library, the drawbacks of the approach become visible. Biological databases require curation as much as any other document collection, their technical infrastructure must be continuously maintained and their users require guidance and support. Significant amounts of time have been invested into these tasks, both by the author and the other members of the SILVA team. However, filling the position of librarian, whether in terms of technical, support or curative

staff, by means of PhD students is not sustainable. While conceptual work on library or database infrastructure and content, such as involved in initially establishing the library or significantly enhancing the methods it employs, may be considered research, purely operational tasks must eventually be delegated to staff hired for this purpose. Operational tasks not only infringe unduly in their scale on the work expected from research staff, but their adequate execution also requires different qualifications. Enthusiasm, creativity and minds unencumbered by the routine of tried and tested processes are helpful when aiming for innovative research, but detrimental to the operation of robust and reliable services, where experience and a certain amount of stoicism are most sought for. Yet, at the time of writing, funding for adequate staff positions is extremely hard to acquire. Thus, we were faced with the alternatives of ignoring the operational tasks, effectively consigning the SILVA project to the grave, or assuming the burden ourselves. Accepting, that this would severely limit our capacity for further innovation, the entire SILVA team opted for the latter. Based on our experiences, we firmly believe it to be imperative for fruitful future research that this situation is remedied. As libraries have been the answer to preserving and disseminating the knowledge represented in the written word, databases in general are needed to realize the value of other data produced at significant cost. Although a system of such databases will itself require significant funding, we believe the allocation of such funds to be a very worthwhile investment. We also believe, that these databases should remain at their origin of development, rather than be moved to central locations. As we have also observed during the SILVA project, a database project is never complete. Every solved problem opens opportunities for further enhancements. This is therefore essential to continued research into ways of improving data and services hosted by individual database that these remain in the vicinity of the institution at which they were initially conceived. These beliefs are shared by many researchers in the life sciences, as illustrated by the founding of the European Life sciences Infrastructure for Biological Information (ELIXIR) project (`www.elixir-europe.org`) by EMBL-EBI. The mission statement of ELIXIR includes the acquisition of funding dedicated to infrastructure such as biological databases and the hubs-and-spokes model envisioned resembles the aforementioned wish for enabling the preservation of locally acquired knowledge. Beyond databases, ELIXIR also includes the areas of training, infrastructure for large scale computing and the development of tools dedicated to the tasks of data analysis in life-science.

12.2 Tool Development

The development of bioinformatical software tools faces the same difficulties as the provision of database services. Similar to statistical methods, bioinformatical methods are often opaque and difficult to understand for those not schooled in the discipline. It is therefore usually expected that novel methods or algorithms are accompanied by ready-to-use implementations. In this, a conflict of interest exists between the biological and bioinformatical domains. From a purely bioinformatical research point of view, a prototypical proof-of-concept implementation is entirely sufficient to complement the theoretical basis of the scientific contribution with its empirical evaluation. The domains applying bioinformatical methods, however, have justified interest in well designed, robust and easy to use tools to be immediately fielded in production environments. The natural way of reconciling these interests is the interjection of a commercialization step. Yet, this solution is not without caveats. Firstly, it is limited to applications for which a well funded market exists. Secondly, commercialization introduces business interests, which include the protection of innovations rather than their open exchange to further scientific progress. As with databases, we therefore believe that the life sciences would be best served by the allocation of funds to the development of scientific software by qualified programmers under free and open source licenses.

Although the practice of publishing algorithms and methods in the form of tools has lead to synonymous use of the terms “algorithm”, “method” and “tool”, it is important to realize the difference between the terms. An algorithm may be of interest even if it has no practical worth. A tool, on the other hand, is valued for its practical merit, even if the algorithms or methods it employs are inferior. Practical merits, such as the ease with which the tool can be applied, often even eclipse scientific precision in their relevance to the choice of tool made by researchers. For example, neighbor joining (NJ) is often applied because a NJ-tree can be easily acquired using web tools integrated with database searches. Although trees reconstructed with Maximum Parsimony (MP) and Maximum Likelihood (ML) methods more accurately reflect the true phylogenetic relationships, NJ is applied even to datasets were MP and ML based reconstruction would pose no significant computational problems. Web services use NJ mainly because it is much faster and therefore easy to implement without risking high computational costs even if offered to a large number of users.

Choosing a tool based on practical merit, however, is not inherently wrong. A tool that is simple to use and that produces results that are easy to interpret is less likely to be applied inappropriately. In producing tools that are ready to use, rather than methods that must be adapted to the research question,

bioinformatics implicitly takes on responsibility. This can be of great benefit to the interdisciplinary application of bioinformatical methods in biology or other life sciences, but only if the responsibility is taken seriously. There is also a conflict between the attributes “powerfull” and “flexible” on one hand, and the attributes “simple” and “safe” on the other hand. While it may be considered to be more responsible, to programmatically limit the application of a tool to what is judged correct or appropriate, such limits also reduce the flexibility and power of the tool. Our introduction of the “expert mode” into ARB was an attempt to fulfill both aims at once. By default, the features deemed potentially dangerous yet powerful and valuable are deactivated to safeguard them from accidental use. In our work on ARB, SINA and the SILVA website we have often found it to be very challenging to design user interfaces that are easily understood yet sufficiently powerful. It is easy to err on the side of flexibility as it is tempting to expose all functionality offered by the underlying engine. Catering only to standard work-flows, on the other hand, unduly inhibits creative application to novel research questions.

12.3 Alternatives to MSA Oriented Approaches

As our purpose was to seamlessly integrate with existing work-flows, rather than radically change them to allow continued feasibility at current data volumes, we considered the preparation of a MSA an axiomatic requirement. However, MSA, or even completely alignment free techniques exist for most of the methods stacked on top of MSA preparation. Phylogenies can be constructed from distance matrices alone, which in turn could be derived by pairwise alignment or other means of distance estimation. Alternatively, there are methods that will compute both MSA and phylogenetic tree at once. For sequence classification, naive Bayesian methods have recently become popular. Aside from the fact that it is a well researched topic, the most major benefits of using an MSA stage in sequence analysis are modularity and transparency. An MSA represents the interface between homology detection and homology interpretation. Irrespective of whether one is interested in determining sequence distance, phylogenetic relationship, covariance behavior or conservation, methods exist to do so based on a preexisting MSA. This MSA can be constructed by a variety of different tools or even manually, and the results can be easily inspected and compared. Thus, trust in this stage of analysis can be established and built upon in further stages.

One alternative to MSA based methods are those relying purely on pairwise alignments. Besides performance issues, the main argument brought forth by

proponents of pairwise alignment is that multiple sequence alignments (MSAs) overestimate the distance between sequences. We would, however, argue that pairwise alignments always underestimate evolutionary distance. Pairwise alignment strives to minimize the distance as defined by the applied metric, and in the algorithm by Gotoh a method exists to do so optimally for edit distance with affine gap penalties. The shortest path of weighted edit operations connecting one observed sequence to another, however, is not necessarily the one having actually occurred in nature. Off-roads will have been taken and although not all intermediate versions of the sequence need to have been functional, it is safe to assume that at least some were. Especially for large distances between functionally homologous sequences, it is unlikely that the path taken by evolution deviated far from the sequence space constrained by functional viability. Beyond the use of DNA evolution models to correct for hidden mutations, pairwise alignment based analysis must therefore always be aware of the increasing ambiguity of the computed results as sequence dissimilarity grows. Multiple sequence alignment is certainly not immune to this issue. The process followed in progressive alignment is derived exactly from the observation that an alignment will be most reliable if the pair of sequences aligned is most similar. Given sufficiently dense data, however, multiple sequence alignment has the opportunity of closely tracking the path actually taken by evolution and thus arriving at more precise results.

12.4 Redundancy with Competing Databases

In the RDP II and the greengenes databases, competitors to SILVA exist, raising the question, whether it is necessary for several projects with largely overlapping purposes to exist. We believe this to be most definitely the case. The possibility of establishing result independence from the choices made by the database providers alone would be sufficient justification. The more orthogonal the collection of methods used by the database providers is, the more trustworthy become results that can be equally derived from any of the databases. In this sense, we believe it to be highly valuable that the alignment method used by RDP II, Infernal, relies on a wholly different approach than SINA. Of course, the data quality in both RDP II and SILVA is sufficient to forgo comparative evaluation in most studies. Yet, occasionally such evaluation should be executed to ensure satisfactory precision in the data provided, thus maintaining the infrastructure character of both projects. All databases also feature “unique selling points”. SILVA is the only rRNA database to offer LSU sequences. It is also still the only one to provide eukaryotic sequences, although efforts are underway at RDP II to remedy this and a joint working group Eukaryotic Taxonomy Working Group (ETWG) has been formed. It will

be the purpose of ETWG to expedite establishing a reliable, curated, rRNA phylogeny based taxonomy for this domain. Only greengenes and SILVA include the taxonomies of the other databases, and only the SILVA web interface allows changing the active taxonomy without loosing the active set of sequences of interest. Only SILVA provides fully featured datasets in ARB format. On the other hand, only RDP features an assignment generator for lectures, an automated tree-generator (using Weighbor to build a weighted neighbor-joining tree) or an online interface to a pipeline for preprocessing data from pyrosequencing projects. Lastly, the existence of multiple databases addressing overlapping or identical needs created an environment of friendly competition. While some of the features shared by the databases were arrived at independently, others were pioneered by one database, whereupon the remaining databases moved to close the gap in usability. For example, SILVA was the first database to dare assigning quality values to the sequence it hosts. This, as well as the concept of a sequence cart (called SEQCART in RDP II and interest list by greengenes) have been subsequently been implemented by all databases. Similarly, RDP II set a new standard with their classifier and sequence search tools. These have been answered in SILVA by the search and classification features offered through SINA. Thus, we conclude, that it is highly beneficial for all three databases to exist. Even though they are to a certain degree mutually redundant, science is better served by the current diverse scenario than by a single, monolithic project.

12.5 Quality Assertion

In the SILVA pipeline, we have dared to assign quality measures to sequences in an attempt to allow researchers to focus on reliable sequence data, thereby reducing the noise that low quality sequences introduce into their analyses. We use the term “dared” because inferring sequence quality from the sequence data alone cannot be accomplished with certainty. We count ambiguous bases and homopolymers as these typically result from problems during sequence acquisition. In the case of ambiguities this is relatively safe, as it is clear that the actual sequence is not precisely represented. Homopolymers on the other hand may also occur naturally. We measure the fractional content of vector contamination by using BLAST to determine whether the parts of a sequence not found to be homologous with rRNA are homologous with vector sequences instead. In using this value as a quality indicator, rather than simply removing the contamination, we assume that the failure to remove the vector content by the original research indicates a lack of diligence that may also pertain to the remainder of the sequence. We use the alignment score and a value indicating to which degree the aligned sequence matches the secondary structure

associated with the alignment as “alignment quality”. More precisely defined, the value is an indication of the primary and secondary structure similarity of the query sequence with the sequences in our seed alignment. While this value does correlate highly with measured alignment accuracy, it is no actual estimation of expected alignment accuracy. It is also no true homology score, but it does show that a sequence is distant to the sequences contained within our seed alignment. The Pintail software measures to which degree the mutations spread over a sequence deviate from the typical distribution. Thus, it can only detect that a sequence is unusual, not whether it is faulty. Knowing these limitations in our quality estimation procedures, we use only very low cut-offs to reject sequences and expose the quality values prominently on the SILVA website.

CHAPTER 13

Conclusion & Outlook

In the combination of ARB, SILVA and SINA, we have presented powerful components to be used in rRNA gene analysis based work-flows. However, with the elimination of one bottleneck, another arises, as any answered question raises a number of new questions. Elements of the so-called "data-deluge", in particular those caused by next-generation sequencing technology, offer a multitude of new applications. Yet, as these applications can be expected to be data-heavy, much work remains to be done in the area of data analysis. Hardware scaling, such as through the use of cloud computing, can only address a small fraction of the problems surrounding the analysis of large volumes of complex data. Using buisiness accounting as an example of data analysis spanning orders of magnitude in data volume, the problems in data scaling become clear. While on household-scale, accounting can be accomplished easily by an unschooled individual using pen and paper methods. Small buisinesses may get by with a spread-sheet application and a personal computer. An international corporation, however, requires more than a mainframe or compute cluster. Rather, extremely complex (and expensive) custom software modelling the business logic is used to collect, persist, analyze and visualize data. And no matter how refined this software, entire buildings are filled with specifically schooled accountants and buisiness analysts. Consequently, we cannot assume that more data and more compute power will by itself result in furthered knowledge. Rather, extensive bioinformatical training needs to be included in the curriculum of the disciplines applying bioinformatical methods. Ultimately, a thorough understanding of the applied methods cannot be replaced by automation or simple and easy user interfaces. Yet, more refined methods and tools are of course required as well. In the remainder of this chapter we highlight some ideas on extending the capabilities of the tools developed in this thesis.

Tools assisting the curation of large taxonomies and the incorporation of information from phylogenetic trees are mandated by the growing volumes of sequence data and the scale of a taxonomy encompassing the entire tree of

life. The SILVA taxonomy in release 108 of the SSU database comprises 27,075 taxa. Including the other three taxonomies, the database contains a total of 70,163 taxa. These numbers illustrate the magnitude of the task of curating a complete taxonomy. Currently, this task is addressed in an almost entirely manual fashion. The SILVA taxonomy derives from a phylogenetic tree constructed iteratively using the ARB maximum parsimony add-to-tree function. This tree can be merely considered a guide tree, as the method is inferior and the tree must be modified to remove conflicts with established taxonomy. Especially in light of the recently formed ETWG, tools that bridge the gap between phylogenetic inference and taxonomic classification are sorely needed. At current scales, building a "phylogeny informed" taxonomy manually has become difficult to manage. The exact shape a useful tool should take is difficult to predict. A means to visualize the differences in topology between phylogenetic and/or taxonomic trees may be a worthwhile approach. Ultimately, the intention must be to incorporate the phylogenetic signal from many trees reconstructed using different methods and based on different genes. While resorting to "crowdsourcing" to accomplish the curation as a community effort may at first seem to be a good idea, this approach would endanger the reference character of the taxonomy. Even assuming that sufficient contributions could be collected, ensuring the quality of those contributions would be a very challenging task. Considering that databases such as SILVA are used in tag sequencing or barcoding approaches as a dictionary to translate sequence data into taxonomic identifications, the stability and accuracy of the taxonomy is imperative.

As SILVA already contains databases for two genes, an extension to a larger number of marker genes is therefore easily conceivable. Most of the software components comprising the SILVA system would support such an extension easily. However, two factors have thus far prevented the extension to more genes. The first is the gene specific data needed for pipeline operation. This includes the parameters that are tuned to each gene, the reference databases and the hidden Markov models used for sequence detection. However, the limiting factor is the additional manual effort required for curation and quality control. As long as the preparation of a database release continues to require manual effort, the extension of SILVA to further genes would bind to much resources.

Nevertheless, extending SINA to also allow the alignment of protein sequences may be worthwhile. As the PT server currently used for k-mer searches can only support DNA sequences, this would entail adding a k-mer search module to SINA itself. Such a module would also be desirable as we expect a simple word index to be faster than the suffix-tree solution. Further improvements to SINA would be detecting and handling introns appropriately, explicit calculation of the expected alignment accuracy and more refined search and

classification modules. Also, as SINA depends on a high quality reference MSA, features supporting the construction or improvement of such MSAs would be particularly useful. The strategies used in the iterative refinement stages of progressive alignment methods may contain concepts applicable to this problem.

We also see much room for improving the services offered via the website. The TestProbe tool should be extended to support the evaluation of primers, and primer or probe combinations. This could be complemented by a tool for primer and probe design. The sequence based search is currently implemented via the alignment facility. A more convenient implementation would be the integration of sequence based search directly into the search based on sequence descriptors. Currently, sequences can only be submitted for alignment and classification. In a more integrated service, these sequences would remain associated with a user account and passed through all components of the SILVA pipeline including quality screening, removal of redundant sequences and calculation of statistical summaries. A data entry module combined with a gateway for sequence submission to the primary sequence data archives would serve to further reduce the effort for completing rRNA studies. At the same time, such a module would offer an opportunity for increasing the consistency and completeness of the submitted descriptive data.

Ideas for further improving the ARB software include a revised plugin system, a database schema more suited to next generation sequencing (NGS) data, a modernization of the toolkit employed by the graphical user interface, a means for offloading compute heavy tasks to cloud or cluster infrastructure and perhaps a tighter integration with SILVA. The current plugin system was designed to be compatible with a software called “GDE”. While this system theoretically allows integrating external tools by simply replacing a configuration file, this has never been exploited to allow flexible extension of ARB. If system were revised to reflect the extension system used in the Mozilla products, ARB could be used as a graphical shell for command line based bioinformatic tools in a much broader fashion. The current database schema in ARB is sequence central and allows only key-value pairs to be added to sequence entries. Refactoring ARB to use a more normalized schema would reduce main memory requirements and increase database consistency. For example, publications should be stored as separate entities, rather than as series of author and title properties repeated in each sequence. The Motif toolkit used by ARB to render its graphical user interface (GUI) is outdated and lacks many useful widgets offered by more current tool kits. A port to Qt or Gtk would immediately modernize the look-and-feel significantly and could be used as a starting point for more in depth work on the ergonomic properties of the ARB user interfaces. It would also be extremely useful to change the way ARB launches and controls external tools such that these can continue to run after the ARB software is

closed. At the same time, an interface could be integrated that allows submitting external computations for offline computation on a cluster or in the cloud. Furthermore, an active synchronization feature that updates the ARB database to the most recent SILVA release or even integrates with the data stored by SILVA as part of a user account could drastically improve user experience. At a smaller scope, simply allowing the synchronization between selected sequences in ARB and on the SILVA website would be a powerful feature. This will, however, necessitate stable and compatible sequence identifiers in SILVA and ARB that can be assigned even before sequence submission to the nucleotide archives.

Part V

Appendix

APPENDIX A

SINA manual

ABSTRACT

SINA is a tool for aligning sequences with an existing multiple sequence alignment (MSA) at high accuracy. It can also execute a homology search based on the computed alignment and generate a per sequence classifications from the search results.

This manual documents the command line usage of sina. Please see `http://www.arb-silva.de/aligner` for a reference to the scientific description of the employed algorithms.

A.1 Synopsis

`sina` `-i` *sequences*`.fasta|arb` `-o` *output*`.fasta|arb`
{`--prealigned` | `--ptdb` *aligndb*`.arb`}
[`--search` `--search-db` *searchdb*`.arb`]
[*options*]

A.2 Description

You can view SINA as a one-command pipeline composed of the following stages:

1. Read sequences from FASTA or ARB file.
2. Align sequences with reference MSA.
3. Search for most similar sequences in search MSA.

4. Classify sequences using search result.

5. Write sequences to FASTA or ARB file.

You can enable or disable the middle three stages as required. By default, only the alignment stage is enabled. Briefly, this is what those stages do (see section Options for details on the configuration options accepted by each of the stages).

Read: Reads sequences from a multi-FASTA file or an ARB database. If reading from ARB, additional meta-data can be read as key-value pairs. These key-value pairs are carried with each sequence throughout the pipeline and will be exported at the end.

Align: Sequences are aligned using the POA algorithm with a set of reference sequences drawn from the reference MSA. Reference sequence selection is based on a kmer search.

Search: Sequences are compared after alignment with the aligned sequences in the configured search database. Comparison is done either against all sequences or against the best matches from a kmer search. Identity is computed as the number of identical column/base pairs divided by the length of the query sequence.

Classify: Sequence classification uses least-common-ancestor (LCA) to derive a classification from the classifications of the sequences found during the search stage.

Write: Writes sequences and meta-data to a multi-FASTA file or an ARB database. See section Options for possible format options to export meta data when writing to multi-FASTA.

The default parameters are a pretty good starting point. They were optimized using a large SSU rRNA gene reference MSA. If you want to use SINA for other gene sequences, see section Examples on how to do some simple accuracy benchmarks on them. To improve the results, the parameters you will want to start with are `-fs-full-len` (set to the typical size of a full-length sequence) and `-fs-kmer-len` (setting this to 8 may help with more variable or shorter sequences).

A.3 Options

Options beginning with a single "-" must be separated from arguments with a space character. Options beginning with "--" can also be separated from arguments with an equal sign.

A.3.1 General Options

`-h`, `--help` Print a summary of the available options and exit.

`--version` Print the version information and exit.

`--show-conf` Print a summary of all configuration settings before processing the input sequences.

`-i` *filename*, `--in=`*filename* Specify the source file containing the sequences to be aligned. The special filename ":" can be used to access an open ARB database when starting SINA from a shell spawned from within ARB. The sequence data may already be aligned (and should be, if you supply *--prealigned*).

`--intype` {*fasta*|*arb*} Specify the format of the source file. If the filename ends with "arb" or "fasta", the type is automatically deduced.

`-o` *filename*, `--out=`*filename* Specify the destination file for the aligned sequences. The special filename ":" can be used to access an open ARB database when starting SINA from a shell spawned from within ARB. If you want to discard the aligned sequences, you can set *filename* to */dev/null* and `--outtype` to *fasta*.

`--outtype` *fasta*|*arb* Specify the format of the destination file. If the filename ends with "arb" or "fasta", the type is automatically deduced.

`--prealigned` If set, the alignment stage is disabled. Sequences are passed to search (if enabled) and output stage unmodified. Mandatory alignment parameters (`--ptdb`) are not required in this case. The input file should contain correctly aligned sequences.

`--search` If set, the search stage is enabled.

A.3.2 Logging Options

`--show-diff` This flag enables visualization of alignment differences. This feature allows you to quickly assess where your alignment differs from the one SINA computed. By also showing you the alignment of the reference sequences used for aligning the sequence, you can get an idea of why SINA came to its conclusions. Many cases of "suboptimal" alignment can be attributed to inconsistent alignment of the reference sequences. To fix such problems, you could either correct the alignment of the reference sequences or add your corrected sequence to the reference alignment.

Alignment difference visualization requires the input sequences to have been previously aligned in a way compatible with the used reference alignment. For positions at which the original alignment and the alignment computed by SINA differ, output as shown below will be printed to the log:

```
Dumping pos 1121 through 1141:
---------  4 14 16-17 21 24
G-C-AGUC-  40 <---(%% ORIG %%)
GCA--GUC-  41 <---(## NEW ##)
GCA-AGUC-  0-3 5-13 15 18-20 22-23 25-27 29-39
GCAA-GUC-  28
```

In this case, the bases 'C' and 'A' where placed in other columns than as per the original alignment. The original alignment is marked with `<---(%% ORIG %%)`. The new alignment is marked with `<---(## NEW ##)`. The numbers to the right of the alignment excerpt indicate the indices of the sequences in the alignment reference (field *align_family_slv*) which the respective row represents. All-gap columns are not shown. The first line indicates the range of alignment columns displayed.

`--show-dist` This flag enables computing the values *sps*, *error*, *matches*, *mismatches*, *bps*, *cpm*, *idty* and *achieved_idty*. See section "Generated Meta Data Values" for an explanation of the individual values. All values except *bps* are computed by comparing the newly computed alignment with the original alignment of the sequences. If a database is configured using `--orig-db`, the original alignment is obtained from that database. Otherwise, the alignment of the input sequences is used.

`--orig-db` ***arb database*** The database *arb database* is used to retrieve a aligned sequences to be used as a reference for comparison by `--show-diff` and `--show-dist`. Sequences are retrieved based on the contents of the ARB field *name*. If FASTA is used as input format, the first word of the FASTA header will be used for matching.

`--colors` Enable color in the output of `--show-diff`.

`--log-file` ***filename*** Redirect the log output to *filename*.

A.3.3 Reading from ARB

`--select-file` ***filename*** If using an ARB database as sequence input file, only sequences with a *name* matching a line contained within *filename*

will be passed into alignment and search stages.

`--select-step` ***n*** If using an ARB database as sequence input file, only every *n*th sequence will be passed into alignment and search stages. This may be combined with *--select-file*. In combination with `--select-skip` this option can be used transparently distribute processing of a single ARB database to multiple instances of SINA.

`--select-skip` ***n*** If using an ARB database as sequence input file, the first *n* sequences will be skipped. Combination with *--select-file* is possible. In combination with `--select-step` this option can be used transparently distribute processing of a single ARB database to multiple instances of SINA

`--extra-fields` ***fieldnames*** Passing a colon separated list of field names will load the meta data contained within these database fields from ARB. The contents will be passed as key-value pairs through the internal SINA pipeline. They will be treated like meta data generated by SINA itself. That is, they will be printed to the log file and written to the output file.

A.3.4 Writing to ARB

`--prot-level` ***n*** Set the protection level used to write to the ARB database to *n*. If a field was set to have a protection level above *n*, SINA will (silently) fail to write to these fields. If your sequences have a protection level of for example 4 and you set *n* to 0, your sequence data will not be modified. If you use the same ARB database for input and output, this may be used in combination with `--show-diff` to inspect the effect of varying the alignment parameters without modifying the alignment.

A.3.5 Writing to FASTA

`--meta-fmt` ***{none|header|comment|csv}*** This option configures the format in which meta data will be exported if the output format is FASTA. *none* will discard all meta data (it will still be written to the log, however). *header* will export meta data values as bracket enclosed key value pairs on the FASTA header line. *comment* will export meta data values as key value pairs on FASTA comment lines, that is lines beginning with a semi-colon between the header and the sequence data. *csv* will export meta data values to a separate file in RFC4180 compatible comma separated value format. The filename will be generated from the output filename by appending "`.csv`".

`--line-length` ***n*** If *n* is different from 0, sequence data will be line wrapped after *n* characters.

A.3.6 Alignment Options

`--ptdb` ***filename*** Specifies the ARB database to be used as alignment reference. This is a mandatory parameter. The file must be in ARB format. See section Examples below for an explanation how to generate such a database from a FASTA file using only SINA. The name of this parameter is historical and refers to the fact that a ARB PT server will be started using the configured database to search for the sequences having the least kmer distance to the input sequences.

`--ptport` ***socket*** Configures the socket which will be used for communication with the ARB PT server. SINA will attempt to contact a running PT server via this port. If no PT server can be contacted, SINA will attempt to start one itself.

socket may either be of the format *hostname:port*, specifying a TCP socket, or of the format *:filename*, specifying a Unix socket. If no running PT server could be contacted and a Unix socket is specified or hostname is "`localhost`", a PT server will be started locally. If *hostname* is "`__SGE__`" SINA will start and contact a PT server on a cluster node using `qrsh1`. Otherwise, `ssh1` will be used to start a PT server on the configured host.

The default is to use port "`localhost:4040`".

CAUTION: If a PT server is already running on the configured socket, but its database does not match the database configured with `--ptdb` the results will be undefined. The search result retrieved from the PT server identifies sequences using the *name* field. For completely different databases, this will usually result in SINA being unable to find reference sequences. It may, however, also result in SINA retrieving the wrong sequences.

`--turn` ***{none|revcomp|all}*** Using this option, SINA can be configured to automatically reorient input sequences. If set to *none*, automatic reorientation is disabled. If set to *revcomp* only the reversed and complemented orientation of the input sequences is considered. If set to *all* all four combinations of reversing and complementing the sequence are considered. The default is *all*. Turning this feature off or reducing its scope will improve performance.

To determine which orientation is most likely, SINA uses the PT server to search for the sequence in the configured orientations. If an orientation

different to the original yields a higher scoring best match, the sequence is modified accordingly.

`--realign` Configures SINA not to copy alignment information from identical reference sequences or reference sequence of which the input sequence is a substring.

Normally, SINA will compare the input sequence with all reference sequences found via the PT server search. If the input sequence is a substring of any of the reference sequences, the alignment of the reference sequence of which the input sequence is a substring will be directly transferred to the input sequence.

If the input sequence is found to be an exact match to a reference sequence, this will be noted in the field *align_log_slv* with the string "`copied alignment from identical template sequence`". If the input sequence is found to be a substring of a reference sequence, this will be noted with the string "`copied alignment from (longer) template sequence`". In both cases, the contents of the fields *acc* and *start* will also be logged to identify the reference sequence.

If suitable sequences for alignment copying are found, but *--realign* is set, the sequences will be removed from the alignment reference. This will be noted in the log with the message "`sequences <list of accessio numbers> containing exact candidate removed from family;`".

`--overhang` **{*attach*|*remove*|*edge*}** If the reference sequences used for alignment do not cover the input sequence completely, e.g. because it contains bases beyond the gene boundary, these bases cannot be aligned. This option configures how SINA handles such unaligned bases at the end of the input sequences. If set to *attach*, the bases will be placed in consecutive columns outwards from the last aligned base, i.e. they will be "attached" to the outer most aligned base. If set to *remove*, these bases will be omitted from the output. If set to *edge*, these bases will be placed in consecutive columns inwards from the first and last alignment column, i.e. "moved to the edge of the alignment". The default is *attach*.

`--lowercase` **{*none*|*original*|*unaligned*}** Use this option to configure which bases you wish to be in lower case in the output. The default setting is *none*, which will output all bases in upper case. If set to *original*, the original cases will not be modified. If set to *unaligned*, the case will be used to convey which bases of the input sequences remained unaligned by setting aligned bases to upper case and unaligned bases to lower case in the output. Unaligned bases are either overhang (see *--overhang* above) or result from insertions which could not be found in any of

the reference sequences. If large insertions required shifting aligned bases (see *--insertion* below), the shifted bases will also be considered unaligned and shown in lower case.

`--insertion` **{*shift*|*forbid*|*remove*}** Since SINA aligns sequences to match a given fixed column reference alignment, insertions in the input sequences may have occurred that cannot be accommodated by the reference alignment. While the only correct way of dealing with this is certainly inserting further columns into the reference alignment to create sufficient room, this may not always be feasible.

The default setting is to *shift* the bases surrounding such a large insertion aside as required. This is done by iteratively choosing the nearest free column to the left or right until sufficient columns have been found. Each time bases are encountered between the insertion and the free column, these bases are added to the insertion. The main benefit of this naive approach is that the position and size of insertions that could not be accommodated are known. The message "`shifting bases to fit in N bases at pos X to Y`" will be logged each time an insertion of length N is attempted between positions X and Y with $Y - X < N$. The affected bases can be marked as unaligned by exporting them in lower case letters using the *--lowercase* option described above. A summary giving the total number of shifted bases and the longest insertion is also logged for each sequence.

The option *forbid* configures SINA to instead disallow insertions that will not fit the reference alignment during the dynamic programming stage of sequence alignment. While this option constitutes a loss of optimality of the alignment algorithm if the gap extension penalty (see `--pen-gapext` below) is different from the gap open penalty (see `--pen-gap` below) it results in slightly less damage to the alignment accuracy.

The option *remove* configures SINA to omit bases from insertions as necessary to fit these insertions into the alignment without moving surrounding aligned bases. This option should be handled with care as the original sequence is altered. If the alignment is subjected to column masking or column sampling (such as during tree reconstruction with bootstrapping), omitting bases is safe, as these methods interpret the resulting MSA from a column perspective.

Which option is the most suitable should be carefully considered for each use case. Whenever possible, circumventing the necessity to handle insertions that do not fit into the alignment by simply adding gap columns into the reference alignment is the preferred solution.

`--filter` ***filtername*** Using this option it is possible to configure using statistical information on positional variability during the alignment. "Filter" is a colloquial term used for "sequence associated information" or SAIs as used by ARB. Filters/SAIs of the type "positional variability by parsimony" (PVP) are eligible for use via this parameter. Please consult the SINA publication and the ARB documentation for more information.

`--auto-filter-field` ***fieldname*** This option allows automatically selecting a PVP filter based on strings contained in an ARB database field. If the configured field contains a shared prefix over all selected reference sequences, the ARB database configured with `--ptdb` is searched for a matching filter. A filter is considered matching if the part of its name following the first colon is a itself a prefix of the shared prefix described above. As an example, if *tax_slv* is chosen and all reference sequences share the prefix "Bacteria;Proteobacteria;" then the filter "silva_108:Bacteria" will match. If `--filter` is also provided, the *filtername* must match the part of the filter name before the first colon.

`--auto-filter-threshold` ***value*** The term "shared prefix of all reference sequence" can be relaxed to "longest prefix shared by *value* of the reference sequence" using this parameter. (Default: 0.8)

`--fs-min` ***value*** The minimum number of reference sequences that should be used. If less matches are returned by the kmer search, less sequences will be used. (Default: 40)

`--fs-max` ***value*** The maximum number of reference sequences that should be used. (Default: 40)

`--fs-msc` ***value*** The minimal kmer score reference sequences should have with the input sequence. At least as many sequences as configured by *--fs-min* will be used. Up to *--fs-max* sequences will be used **if** they have a kmer score higher than configured by *--fs-msc*. (Default: 0.7)

`--fs-msc-max` ***value*** Limits sequence selection to sequences having kmer score no higher than *value*. (Default: 2, that is, disabled)

`--fs-leave-query-out` Setting this option will remove the query sequence from the reference sequences based on its *name*. This is sensible for evaluation in comparison to other tools where leave-query-out style evaluation can only be done by excluding the exact query sequence from the reference. If the alignment must not be directly derived from any reference sequence, even if the reference dataset contains redundant data, `--realign` should be used.

--fs-req *value* The minimum number of reference sequences that must be used. If less matches are returned by the kmer search, alignment is refused. The sequence will not be contained in the output. (Default: 1)

--fs-req-full *value* The minimum number of full length sequences that should be included in the reference. The matches from the kmer search are parsed until, beyond the limits given by *--fs-max* and *--fs-msc*, at least `value` such sequences have been found and added to the reference sequences.

--fs-full-len *value* The minimum number of bases constituting a full-length sequence.

--fs-kmer-no-fast Disable the PT server fast search. The fast kmer search considers only kmers beginning with 'A'.

--fs-kmer-len *k* Configures the length k of the kmers used for the kmer similarity search.

--fs-kmer-mm *value* Configures the number of mismatching bases a kmer may have to be considered matching.

--fs-kmer-norel Computes the kmer score using the length of the query sequence only. If not set, the kmer score is computed as the number of shared kmeres between query and match candidate divided by the length of the shorter.

--fs-min-len *value* Minimal length sequences found via the kmer search must have to be considered for inclusion into the reference sequences.

--fs-weight *value* Factor with which the frequency at which a base occurs within the reference sequences will be used to weight match and mismatch scores between the base and bases from the input sequence.

--gene-start *value* Position within the alignment corresponding to the first base of the aligned gene.

--gene-stop *value* Position within the alignment corresponding to the last base of the aligned gene.

--fs-cover-gene *value* Minimum number of times the gene-start and gene-stop positions are at least touched by one of the reference sequences. If the above rules did not result in sufficient such sequences, further sequences covering the respective position are added until the condition is met.

`--match-score` ***value*** The match score used during the dynamic programming stage of partial order alignment (POA).

`--mismatch-score` ***value*** The mismatch score used during the dynamic programming stage of partial order alignment (POA).

`--pen-gap` ***value*** The gap open penalty used during the dynamic programming stage of partial order alignment (POA).

`--pen-gapext` ***value*** The gap extension used during the dynamic programming stage of partial order alignment (POA).

`--debug-graph` Enables dumping of graph data in graphviz format suitable for processing with e.g. `dot1`. For each aligned sequence, the DAG used as alignment template is dumped. Subsections of the dynamic programming graph/mesh, each covering the same fractions as shown with *--show-diff*, are also dumped. Please be aware that the output will be huge.

`--use-subst-matrix` Experimental. Do not use!

A.3.7 Search and Classification Options

`--search-db` ***filename*** Configures the name of the ARB database which will be used for sequence search. Unless *--search-all* is also set, a PT server will be started for this database. The same rules as for *--ptdb* apply. It is permissible to use the same file as in *--ptdb*. In this case, the database will be loaded only once.

`--search-port` ***socket*** Configures the port on which SINA should communicate with the PT server used for kmer searching. The same rules as for *--ptport* apply. If *--search-all* is set, no PT server will be used and this setting will be ignored.

`--search-all` Configures SINA to compare the aligned input sequence with **all** sequences contained in the database given by *--search-db*. No PT server will be used.

`--search-no-fast` Disable the PT server fast search. The fast kmer search considers only kmers beginning with 'A'.

`--search-kmer-candidates` ***n*** Configures the number of best matching results from the kmer search that should be compared with the input sequences based on the alignment.

`--search-kmer-len` ***arg*** Configures the length k of the kmers used for the kmer similarity search.

`--search-kmer-mm` ***arg*** Configures the number of mismatching bases a kmer may have to be considered matching.

`--search-kmer-norel` Computes the kmer score using the length of the query sequence only. If not set, the kmer score is computed as the number of shared kmeres between query and match candidate divided by the length of the shorter.

`--search-min-sim` ***value*** Minimal identity a sequence must have with the input sequence to be included in the search result.

`--search-ignore-super` Exclude sequences of which the input sequence is a substring from the search result.

`--search-max-result` ***value*** Limit the maximum number of search results per input sequence.

`--search-copy-fields` ***fieldnames*** Configures a colon separated list of ARB fields which will be copied into the input sequence. The field name will be prepended with "copy_<accession>_" in the output to indicate from which search result the data came.

`--lca-fields` ***fieldnames*** Derives a LCA classification of the input sequence from the classifications of the sequences found in the search. This feature requires the reference database to contain a field specifying the sequence classifications in materialized path format (i.e. "Bacteria;Proteobacteria;..."). The "least common ancestor" is the shared prefix of these strings. Prefixes must always end with a semicolon. Depending on the desired rank up to which the sequences should be classified, appropriate sequence similarity cutoffs should be configured with *--search-min-sim*. It is possible to specify multiple source taxonomies as `fieldnames` by passing colon separated list. Derived LCA classification will be stored in fields named "lca_<fieldname>".

`--lca-quorum` ***value*** Relaxes LCA classification from "shared by **all** search results" to a fraction `value` of the search results.

A.4 Generated Meta Data Values

align_bp_score_slv This is a score calculated from the aligned sequence and the HELIX SAI. If the reference database contains no HELIX SAI the score

will be NaN. Otherwise, the score is computed as follows. For each pair of columns covered by the aligned sequence a score of 1 is awarded if the pair is AU, GU or GC; a score of 0 is awarded if the pair is AG or GG; a score of -1 is awarded if the pair is AA, AC, CC, CU or UU or if one of the columns contains a gap character; the sum of these scores is divided by the number of considered columns. The value is scaled to match the range between 0 and 100.

This value is likely to change or disappear in future versions.

align_cutoff_head_slv This is the number of bases at the beginning of the sequences that remained unaligned.

align_cutoff_tail_slv This is the number of bases at the end of the sequences that remained unaligned.

align_family_slv This is a list of the sequences that were used to build to align the input sequence.

align_filter_slv If a PVP filter was applied, the name of that filter will be stored in this field.

align_log_slv Messages generated during the alignment process will be logged here.

align_startpos_slv This is the alignment position (column number) of the first aligned base.

align_stoppos_slv This is the alignment position (column number) of the last aligned base.

aligned_slv This is the current date.

full_name If FASTA is chosen as input format, this field will contain the part of the FASTA header lines after the first space character.

nearest_slv This field contains a space separated list of the results from the homology search stage. Each search result is given in the following form: "`<accession>.<version>:<start>:<stop> <identity>`"

nuc The number of nucleotides in the input sequence.

nuc_gene_slv The number of nucleotides in the sequence aligned to be within the gene borders.

turn_slv Documents actions taken by the automatic reorientation of sequences. Possible values are "disabled", "none", "reversed", "complemented" and "reversed and complemented".

sps This field contains the fractional identity of the aligned input sequence with the input sequence in its original alignment. The number of identical base/column pairs is divided by the number of nucleotides.

error The number of differing base/column pairs divided by the number of nucleotides. The sum of *error* and *sps* may be larger than 1 because of gap characters. If in the new alignment, a base ends up in what should be a gap position and a gap is placed where the base was in the original alignment, two misaligned positions are found.

matches Number of identical base/column pairs in SINA aligned sequence and input alignment.

mismatches Number of differing base/column pairs in SINA aligned sequence and input alignment.

bps The same as *slv_bp_score* but unscaled and not rounded to integer.

cpm Correctly placed mutations, or rather, an attempt at calculating such a measure intended to be used as a measure of alignment accuracy independent of the identity an input sequence as with its closest reference sequences. The value is the number of base/column pairs the aligned sequence shares with its original alignment **more** than the sequence in its original alignment shares with the closest found reference sequence divided by the number of base/column pairs in original alignment that are not matched by the closest reference.

This value is likely disappear or change in future versions.

idty The highest fractional identity of the input sequence with any of the selected reference sequences calculated as the number of matches (see above) divided by the length of the input sequence.

achieved_idty Identical to *idty* but using the SINA alignment rather than the original alignment.

lca_* These fields contain the classifications derived via LCA.

copy_* These fields contain the data copied from the search results.

A.5 Examples

Aligning some sequences To align sequences, you need to get a suitable reference alignment in ARB format. If you have LSU or SSU sequences

to align, the Ref or RefNR datasets from `www.arb-silva.de` work well. Otherwise, check below for an example on how to convert your own multi-fasta reference alignment to ARB format.

```
./sina -i mysequences.fasta -o alignedsequences.fasta \
       --ptdb reference.arb
```

The first time you run this, a PT server will be started and will begin building its index. The index is stored in `reference.arb.pt` and will only be computed again if reference.arb changes (the decision is made based on file timestamps only). The PT server will also continue to run once it has been started. Subsequent sina runs will be much faster therefore. Nonetheless, start-up time may be long if `reference.arb` is large.

Classifying some sequences If you are using a reference database that has a field containing classifications, you can use SINA to classify your sequences. The SILVA Ref database contain several taxonomies in fields beginning with "`tax_`". To classify sequences based on the SILVA taxonomy, you can use this command line:

```
./sina -i mysequences.fasta -o aligned.fasta \
       --meta-fmt CSV \
       --ptdb reference.arb \
       --search --searchdb reference.arb \
       --lca-fields tax_slv
```

The classifications will be (among the other values) written to the column labeled "`lca_tax_slv`" in the file "aligned.fasta.csv".

Converting FASTA to ARB By disabling all stages, SINA can be used to convert between ARB and FASTA format (in a limited fashion, use ARB if you want to do more fancy stuff):

```
./sina -i mysequences.fasta -o mysequences.arb \
       --prealigned
```

This will generate an ARB file from your aligned sequences suitable for use as a reference MSA. The first word of each FASTA header will be written to the ARB field "`name`". Make sure they are unique for each sequence. ARB uses this field to identify sequences, duplicates will overwrite the previous sequence with the same name. The remainder of the fasta header will be written to the field "`full_name`".

Running a leave-query-out accuracy benchmark You can run a quick check on the accuracy achieved by SINA with your reference MSA by having it align each of those sequences (ignoring the same sequence in the process) and log the accuracy with which it could reproduce the original alignment.

```
./sina -i myreference.arb --ptdb myreference.arb \
       -o /dev/null --outtype fasta \
       --fs-leave-query-out --show-dist
```

The average accuracy will be printed at the end of the SINA run.

Converting FASTA output from RNA to DNA SINA encodes bases in its FASTA format output as RNA using IUPAC characters. Conversion to DNA can be achieved by simply replacing all occurences of U and u with T and t using `sed`:

```
sed '/^[^>]/ y/uU/tT/' rna.fasta > dna.fasta
```

A.6 See Also

ARB, `http://www.arb-home.de`
SILVA, `http://www.arb-silva.de`

A.7 Version

Version: 1.2.9 of Oct. 10th 2010.

A.8 License and Copyright

Personal Use and Evaluation License (PUEL) for SINA Stand-Alone Software

This license applies if you download the SINA Stand-Alone Software Package (the "Product") from `www.arb-silva.de`. In summary, the license allows you to use the Product free of charge for academic Personal Use or, alternatively, for non-academic, time-limited Evaluation.

Overview: Personal Use (academic) is when you install the Product yourself and you make use of it. You can use the Product within an academic study to process as much data as you like and publish the processed data as long as you follow the terms below. If you deploy the Product to a single or multiple computers for colleagues within your institution, e.g. in the capacity as a system administrator, this would no longer qualify as Personal Use.

Personal Use does NOT include (1) any redistribution of the Product, (2) any kind of Product-based data analysis service for third parties, or (3) integration of the Product into another software.

License Agreement: You should have received a copy of the license agreement with this software in the file LICENSE.txt. If you did not, please visit `http://www.arb-silva.de/aligner/sina`.

APPENDIX B

SINA supplementary

B.1 Algorithm

B.1.1 Positional Variability by Parsimony (PVP)

The "positional variability by parsimony" (PVP) function of ARB computes a per-column conservation profile from a MSA and a phylogenetic tree. For each column, the number of transitions and transversions required to explain the tree given the aligned sequence data is computed. The sum of transitions and transversions divided by the number of observed bases, capped at 0.55 and corrected for not observed mutations using the Jukes Cantor formula [126]. From this rate, we compute the scoring weight as 0.5-log(rate). This weight is capped at 20. Columns containing gap characters in more than 80% of the sequences are assigned a weight of 1.

The PVP statistic applied may be chosen dynamically for each candidate sequence based on classification meta-data in the reference MSA database. This allows using for example domain specific statistics. The name of each PVP statistic stored in the reference database is compared with the configured classification attribute of each reference sequence. If a name is a prefix of the attribute for a majority of the sequences, the corresponding PVP statistic is used instead of a globally configured PVP statistic.

B.2 Results

All figures displaying mean accuracy (all except Fig. B.2) are shown twice, once using a linear (labeled A) and once using a logarithmic scale on the y-axis (labeled B).

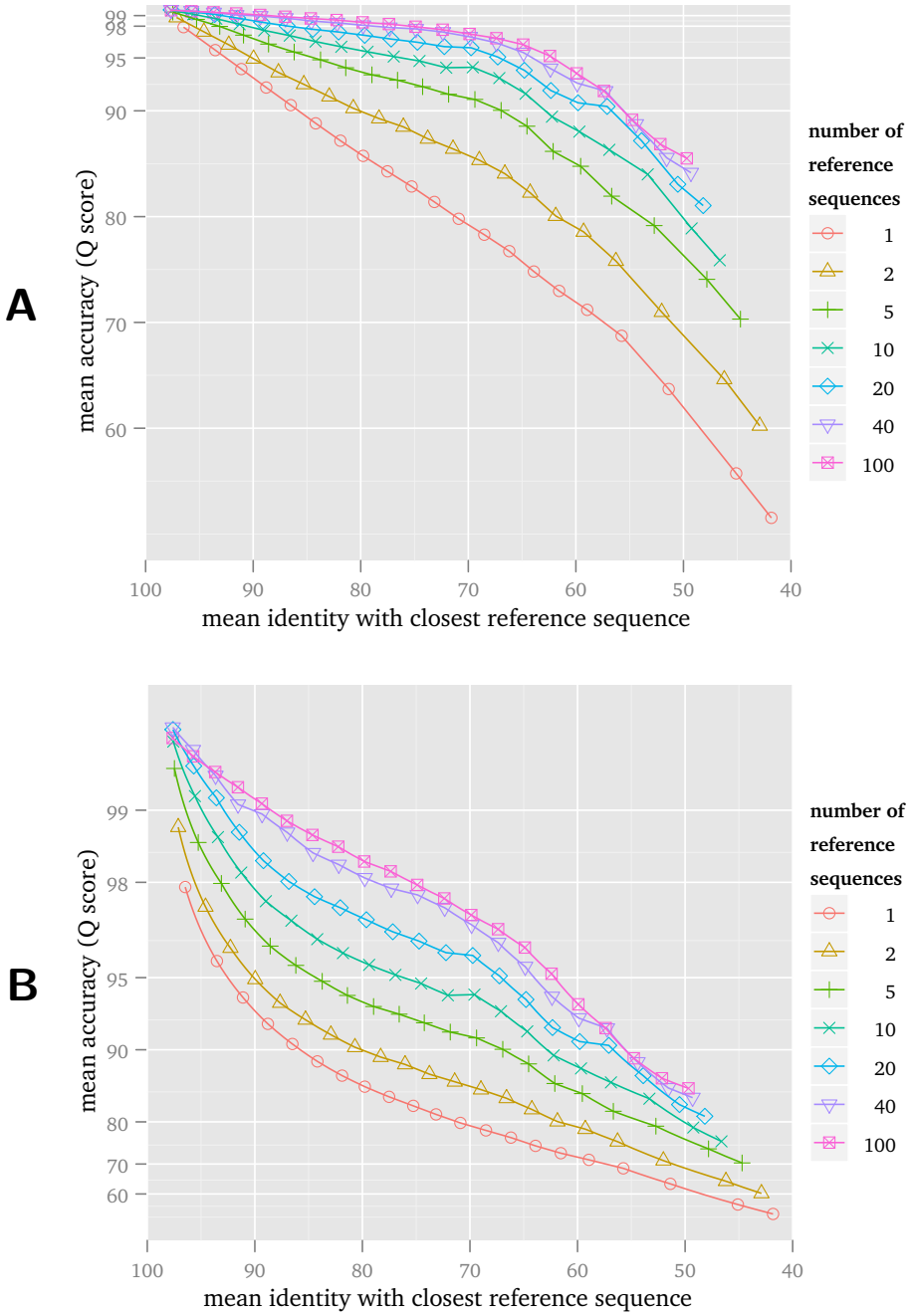


Figure B.1 *Effect of increasing the number of reference sequences on alignment accuracy at different levels of identity with the reference alignment.*

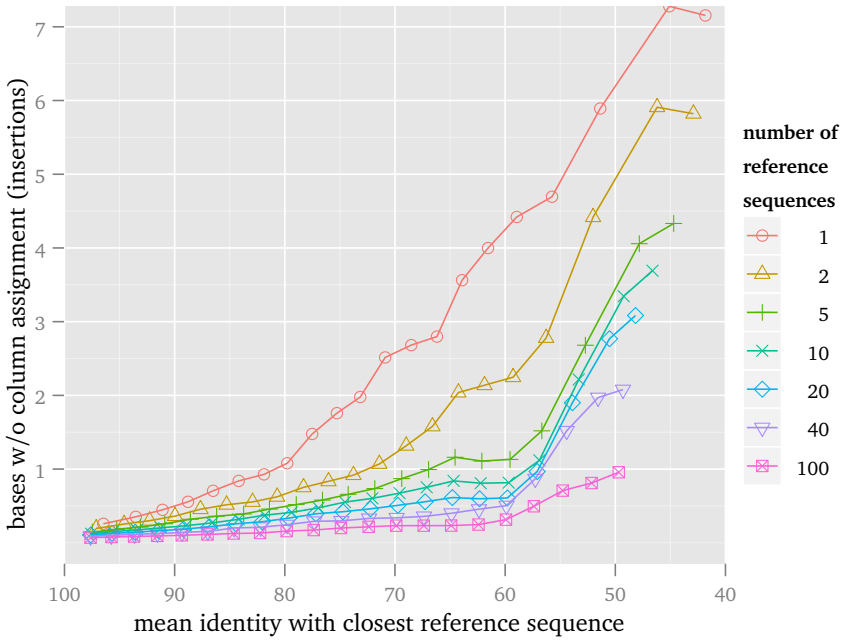


Figure B.2 *Effect of increasing the number of reference sequences on the fraction of "insertions" with respect to the alignment template.*

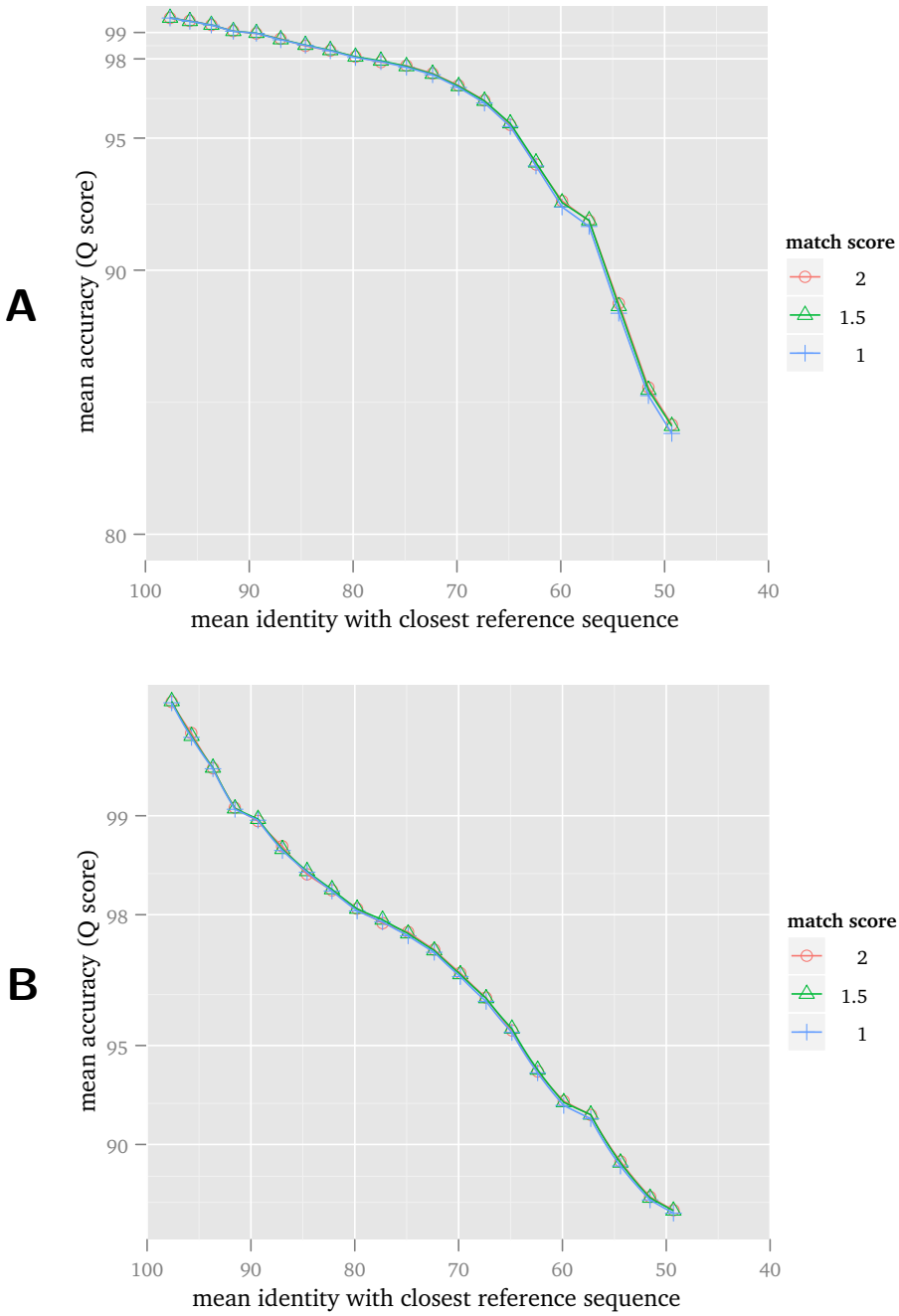


Figure B.3 *Using a mismatch score of -1 a match score of 2 is minimally better than a match score of 1 or 1.5. The difference, however, is almost beyond the resolution of this figure.*

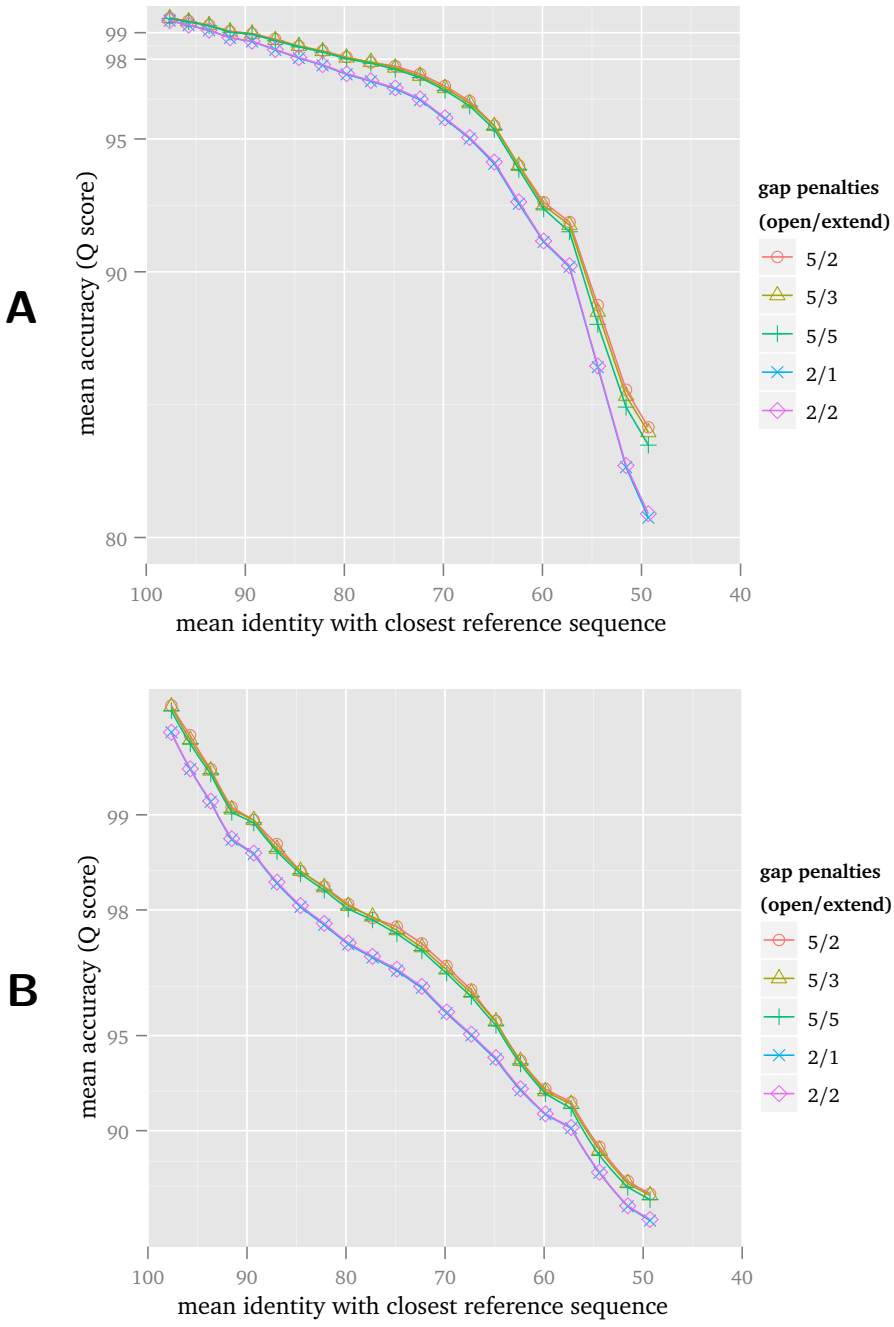


Figure B.4 *A gap open penalty of 5 works better than a gap open penalty of 2. Among the tested gap extend penalties using a gap open penalty of 5, a gap extension penalty of 2 is best by a small margin.*

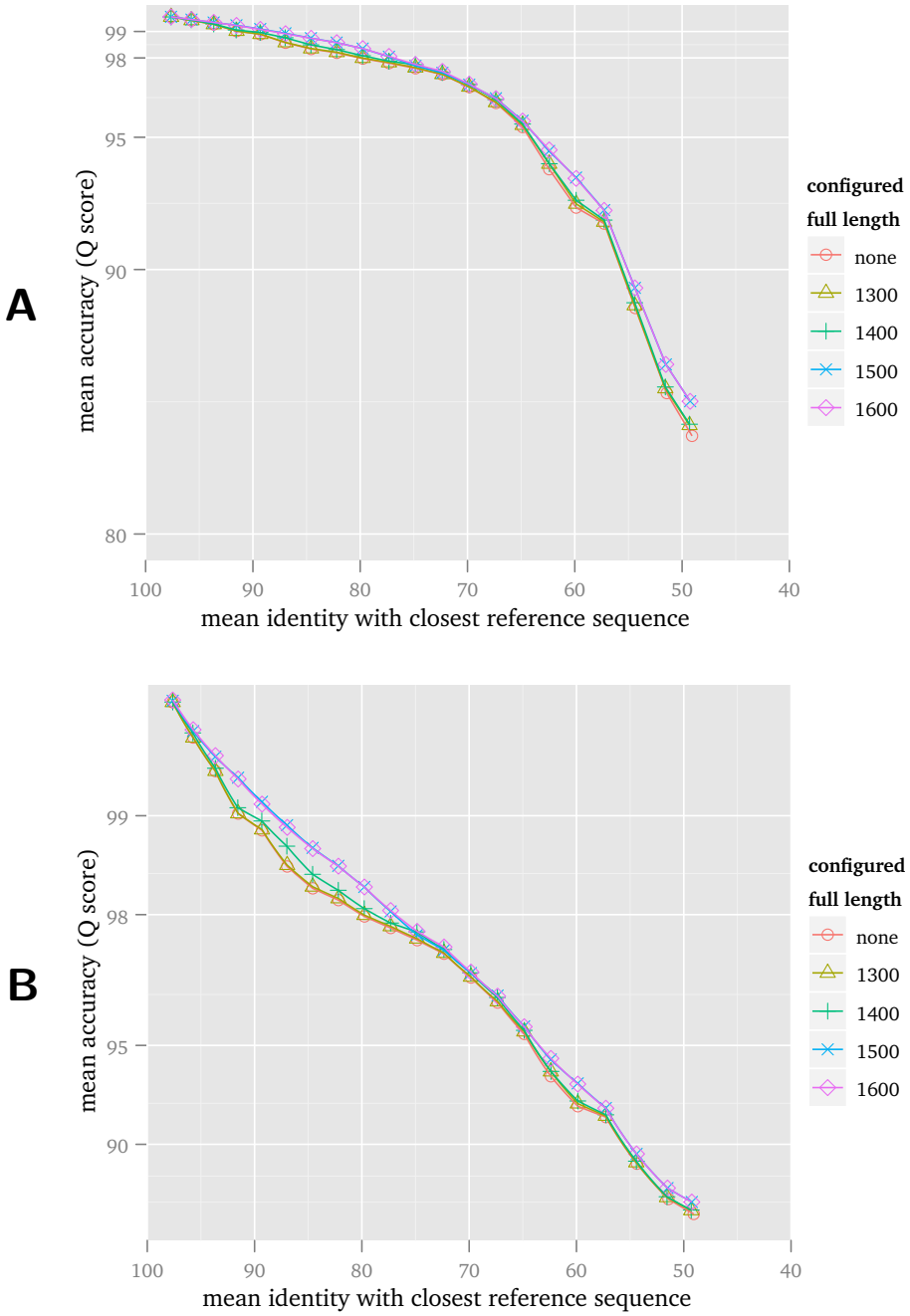


Figure B.5 *Requiring that at least one sequence of 1500 or 1600 bp length be included in the reference set improves averaged results.*

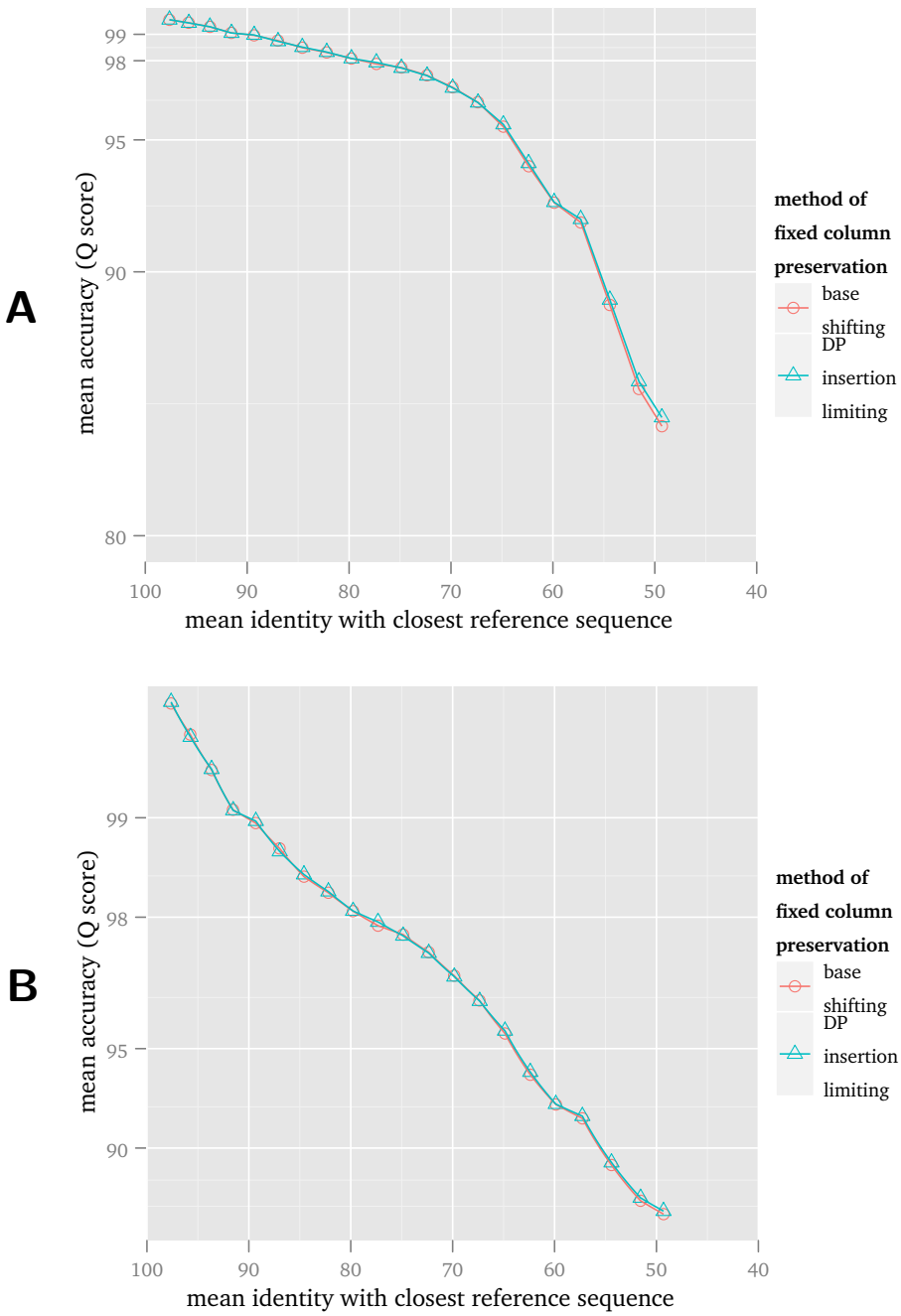


Figure B.6 *The method used for column preservation makes little difference to alignment accuracy using the SSU benchmark. Considering that the SSU alignment is more than 30 times wider than typical SSU sequences and that special care was taken to have sufficient alignment space between bases in the construction of the alignment, this is not unexpected.*

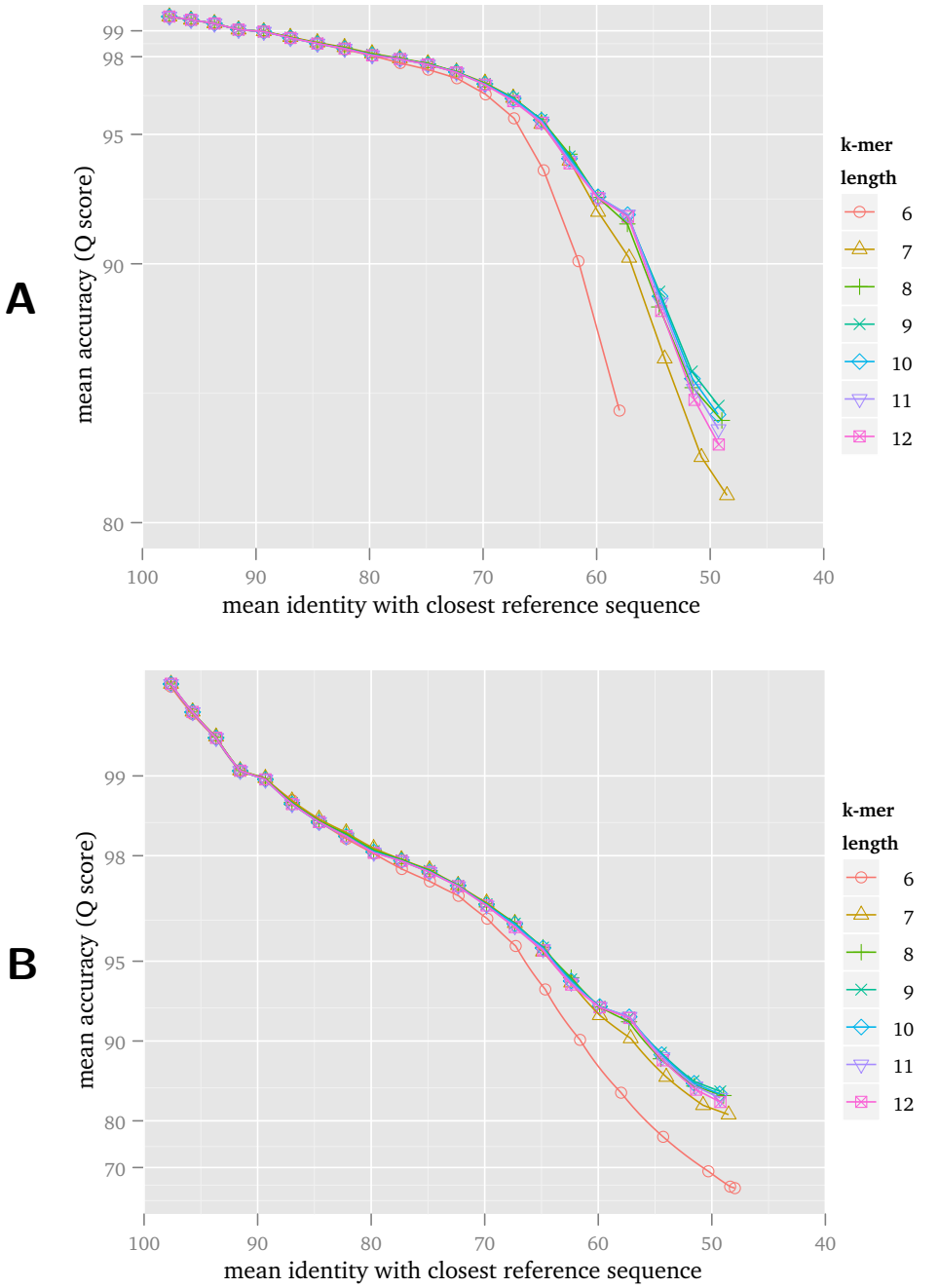


Figure B.7 *Varying the length of the kmeres used to find reference sequences has little impact, values between 8 and 12 work well for SSU sequences.*

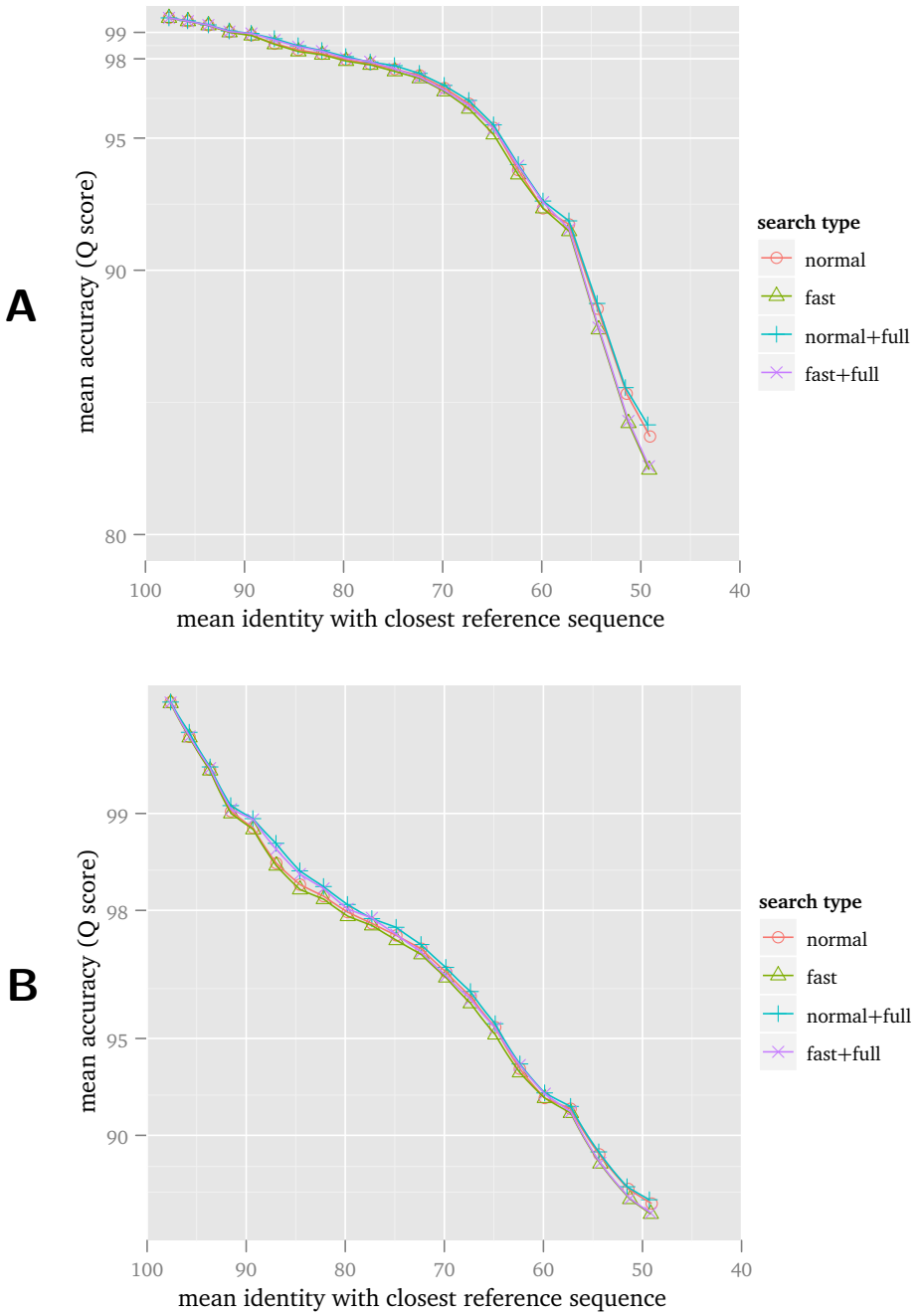


Figure B.8 *Using the "fast mode" of the kmer search provided by the ARB PT server ignores kmeres that do not begin with 'A'. The impact to alignment accuracy is visible. The graphs labled "+full" require that at least one sequence of at least 1400 bp be included in the reference.*

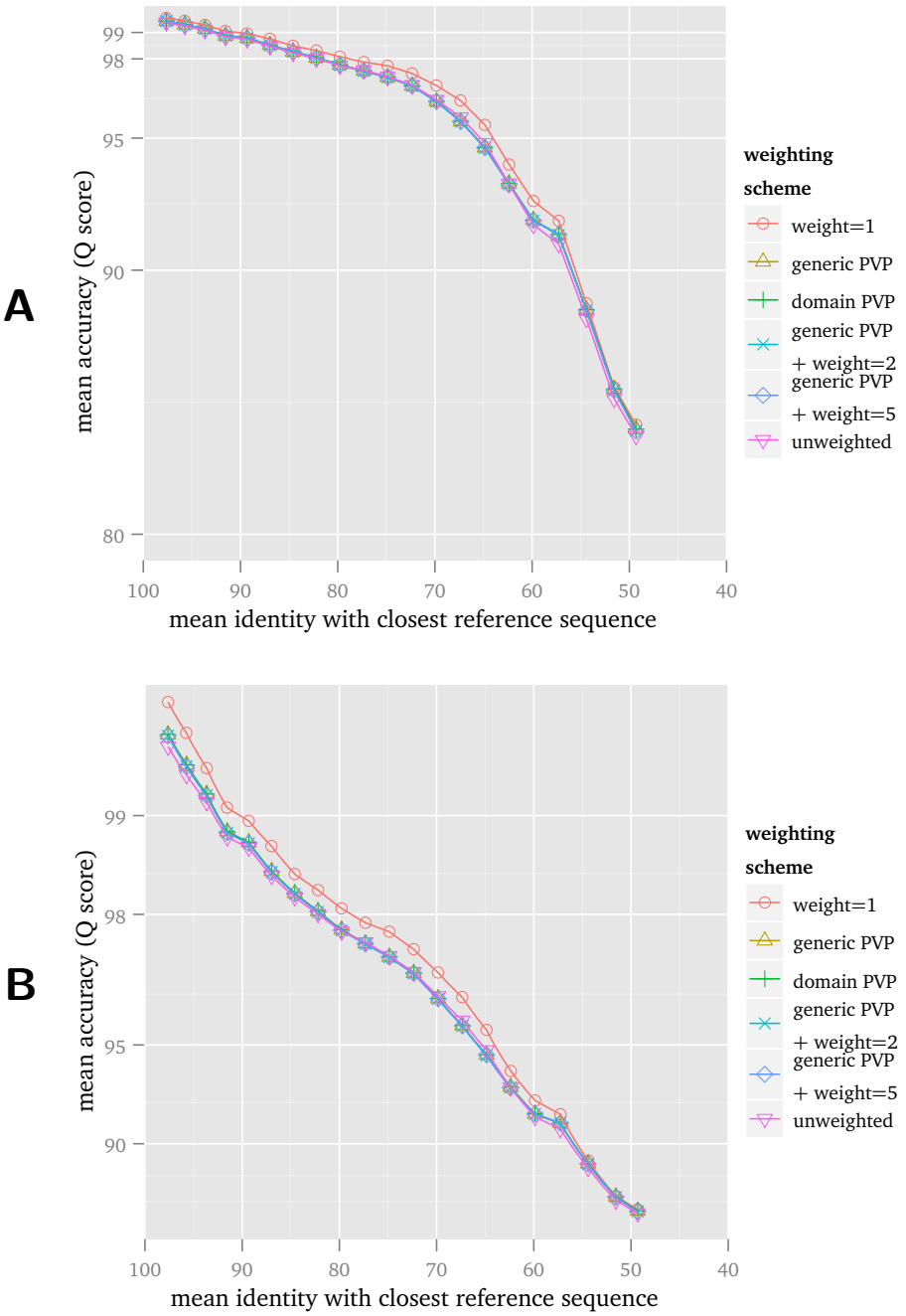


Figure B.9 *All weighting schemes and combinations thereof provide some improvement to unweighted alignment. Using only the base frequency among the selected reference sequences performs significantly better than all other tested methods.*

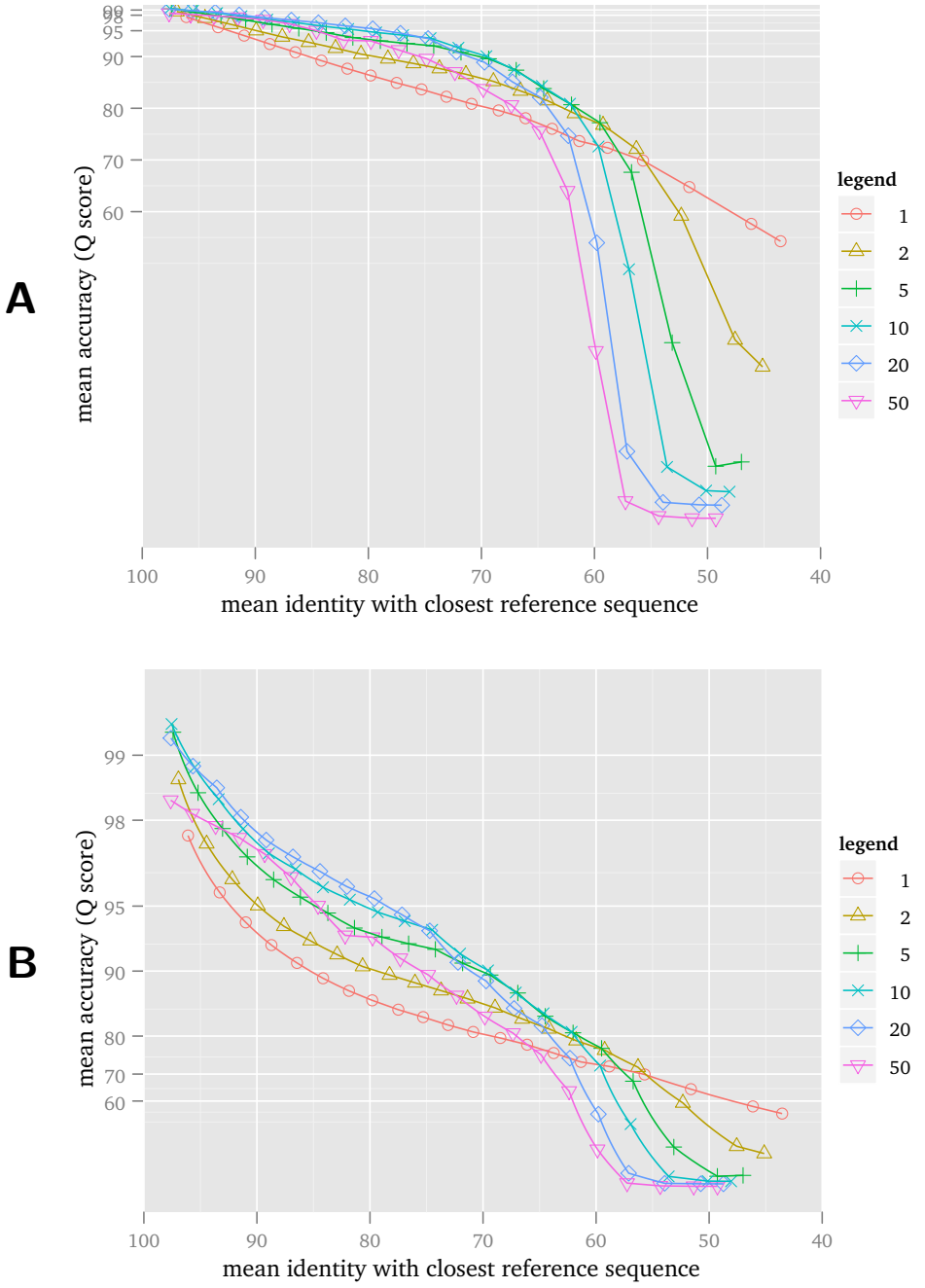


Figure B.10 *Effect of increasing the number of reference sequences when using a column profile and PSP_{xy} scoring function instead of the DAG based method.*

APPENDIX C

Supplementary Materials

The attached DVD contains further supplementary materials including test data and source code as well as a PDF version of this document.

Bibliography

[1] Adrian, M., Dubochet, J., Lepault, J., and McDowall, A. (1984). Cryo-electron microscopy of viruses. *Group*.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[3] Altschul, S. F., Madden, T. L., Schäffer, a. a., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

[4] Amann, R. and Fuchs, B. M. (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology*, 6(5):339–348.

[5] Amann, R. and Ludwig, W. (2000). Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS microbiology reviews*, 24(5):555–65.

[6] Amann, R., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, 59(1):143–169.

[7] Amaral-Zettler, L., Peplies, J., Ramette, A., Fuchs, B., Ludwig, W., and Glöckner, F. O. (2008). Proceedings of the international workshop on Ribosomal RNA technology, April 7-9, 2008, Bremen, Germany. *Systematic and applied microbiology*, 31(4):258–68.

[8] Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P. l., and Engstrand, L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PloS one*, 3(7):e2836.

[9] Andersson, A. F., Riemann, L., and Bertilsson, S. (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *The ISME journal*, 4(2):171–181.

[10] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32(90001):D115–119.

[11] Armougom, F. and Raoult, D. (2009). Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods. *Journal of Computer Science Systems Biology*, 02(01):74–92.

[12] Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol*, 71(12):7724–7736.

[13] Ausubel, J. H. (2009). A botanical macroscope. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31):12569–70.

[14] Bahr, A., Thompson, J. D., Thierry, J. C., and Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*, 29(1):323–326.

[15] Baker, G. C. and Cowan, D. A. (2004). 16 S rDNA primers and the unbiased assessment of thermophile diversity. *Biochemical Society Transactions*, 32(Pt 2):218–221.

[16] Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and reanalysis of domain-specific 16S primers. *Journal of Microbiological Methods*, 55(3):541–555.

[17] Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, 289(5481):905–920.

[18] Barber, R. C., Karol, P. J., Nakahara, H., Vardaci, E., and Vogt, E. W. (2011). Discovery of the elements with atomic numbers greater than or equal to 113 (IUPAC Technical Report). *Pure and Applied Chemistry*, 83(7):1485–1498.

[19] Barns, S. M., Fundyga, R. E., Jeffries, M. W., and Pace, N. R. (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proceedings of the National Academy of Sciences of the United States of America*, 91(5):1609–1613.

[20] Barr, J. J., Blackall, L. L., and Bond, P. (2010). Further limitations of phylogenetic group-specific probes used for detection of bacteria in environmental samples. *The ISME journal*, 4(8):959–961.

[21] Bell, G., Hey, T., and Szalay, A. (2009). Computer science. Beyond the data deluge. *Science (New York, N.Y.)*, 323(5919):1297–8.

[22] Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

[23] Benson, D. a., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). GenBank. *Nucleic acids research*, 39(Database issue):D32–7.

[24] Benson, D. a., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). GenBank. *Nucleic acids research*, 36(Database issue):D25–30.

[25] Biers, E. J., Sun, S., and Howard, E. C. (2009). Prokaryotic Genomes and Diversity in Surface Ocean Waters: Interrogating the Global Ocean Sampling Metagenome. *Applied and Environmental Microbiology*, 75(7):2221–2229.

[26] Blake, J. a. and Bult, C. J. (2006). Beyond the data deluge: data integration and bio-ontologies. *Journal of biomedical informatics*, 39(3):314–20.

[27] Booth, T., Gilbert, J., Neufeld, J., Ball, J., Thurston, M., Chipman, K., Joint, I., and Field, D. (2007). Handlebar: a flexible, web-based inventory manager for handling barcoded samples. *BioTechniques*, 42(3):300–302.

[28] Brosius, J., Dull, T. J., Sleeter, D. D., and Noller, H. F. (1981). Gene organization and primary structure of a ribosomal RNA operon from Escherichia coli. *Journal of Molecular Biology*, 148(2):107–127.

[29] Brunk, C. F. and Eis, N. (1998). Quantitative Measure of Small-Subunit rRNA Gene Sequences of the Kingdom Korarchaeota. *Applied and Environmental Microbiology*, 64(12):5064–5066.

[30] Brysse, K. (2008). From weird wonders to stem lineages: the second reclassification of the Burgess Shale fauna. *Studies in history and philosophy of biological and biomedical sciences*, 39(3):298–313.

[31] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10:421.

[32] Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., Pande, N., Shang, Z., Yu, N., and Gutell, R. R. (2002). The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(1):2.

[33] Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics (Oxford, England)*, 26(2):266–7.

[34] Carrillo, H. and Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, 48(5):1073–1082.

[35] Cavalier-Smith, T. (2006). Rooting the tree of life by transition analyses. *Biology Direct*, 1(1):19.

[36] Cavalier-Smith, T. (2010). Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biology letters*, 6(3):342–5.

[37] Chandler, D. P., Fredrickson, J. K., and Brockman, F. J. (1997). Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Molecular Ecology*, 6(5):475–482.

[38] Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, 311(5765):1283–7.

[39] Claesson, M. J., O'Sullivan, O., Wang, Q., Nikkilä, J., Marchesi, J. R., Smidt, H., De Vos, W. M., Ross, R. P., and O'Toole, P. W. (2009). Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. *PLoS ONE*, 4(8):15.

[40] Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., and Tiedje, J. M. (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, 33:D294–D296.

[41] Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Bandela, A. M., Cardenas, E., Garrity, G. M., and Tiedje, J. M. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*, 35(Database issue):D169—-D172.

[42] Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Tiedje, J. M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue):D141–D145.

[43] Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science (New York, N.Y.)*, 326(5960):1694–7.

[44] Crump, B. C., Hopkinson, C. S., Sogin, M. L., and Hobbie, J. E. (2004). Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. *Applied and Environmental Microbiology*, 70(3):1494–1505.

[45] Dagan, T., Roettger, M., Bryant, D., and Martin, W. (2010). Genome networks root the tree of life between prokaryotic domains. *Genome biology and evolution*, 2:379–92.

[46] Danchin, A., Fang, G., and Noria, S. (2007). The extant core bacterial proteome is an archive of the origin of life. *Proteomics*, 7(6):875–89.

[47] Darwin, C. (1859). *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life*. John Murray, London.

[48] Dawkins, R. (1976). *The Selfish Gene*, volume 8. Oxford University Press.

[49] Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In Dayhoff, M., editor, *Atlas of protein sequence and structure*, pages 345–358. National Biomedical Research Foundation, Washington D.C., vol. 5 edition.

[50] de Queiroz, K. and Donoghue, M. J. (1988). Phylogenetic Systematics and the Species Problem. *Cladistics*, 4(4):317–338.

[51] DeLong, E., Taylor, L., Marsh, T., and Preston, C. (1999). Visualization and enumeration of marine planktonic archaea and bacteria by using polyribonucleotide probes and fluorescent in situ hybridization. *Appl. Environ. Microbiol.*, 65(12):5554–5563.

[52] DeLong, E. F. (1992). Archaea in Coastal Marine Environments. *Proc. Natl. Acad. Sci. U.S.A.*, 89(12):5685–5689.

[53] DeSantis, T. Z., Dubosarskiy, I., Murray, S. R., and Andersen, G. L. (2003). Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics*, 19(12):1461–1468.

[54] DeSantis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., Phan, R., and Andersen, G. L. (2006a). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*, 34(Web Server issue):W394–W399.

[55] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006b). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 72(7):5069–5072.

[56] Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., Lakshmanan, A., and Wade, W. G. (2010). The Human Oral Microbiome. *Journal Of Bacteriology*, 192(19):5002–5017.

[57] Díez, B., Pedrós-Alió, C., and Massana, R. (2001). Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and Sequencing. *Applied and Environmental Microbiology*, 67(7):2932–2941.

[58] Do, C., Mahabhashyam, M., Brudno, M., and Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330.

[59] Droege, M. and Hill, B. (2008). The Genome Sequencer FLX System–longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology*, 136(1-2):3–10.

[60] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, second edition.

[61] Eddy, S. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Inform*, volume 23, pages 205–11.

[62] Eddy, S. R. (2001). HMMER: Profile hidden Markov models for biological sequence analysis. *Washington University School of Medicine St Louis MO httphmmer wustl edu*.

[63] Edgar, R. C. (2004a). Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res*, 32(1):380–385.

[64] Edgar, R. C. (2004b). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

[65] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, 26(19):2460–2461.

[66] Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol*, 16(3):368–373.

[67] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

[68] Elias, I. (2006). Settling the intractability of multiple alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 13(7):1323–39.

[69] Euzéby, J. P. (1997). List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *International Journal of Systematic Bacteriology*, 47(2):590–592.

[70] Fang, M., Kremer, R. J., Motavalli, P. P., and Davis, G. (2005). Bacterial Diversity in Rhizospheres of Nontransgenic and Transgenic Corn. *Applied and Environmental Microbiology*, 71(7):4132–4136.

[71] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–76.

[72] Felsenstein, J. (1989). PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, 5(1):164–166.

[73] Feng, D. F. and Doolittle, R. F. (1987). Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees. *Journal of Molecular Evolution*, 25(4):351–360.

[74] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glockner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., and Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541–547.

[75] Field, D., Garrity, G., Gray, T., Selengut, J., Sterk, P., Thomson, N., Tatusova, T., Cochrane, G., Glöckner, F. O., Kottmann, R., Lister, A. L., Tateno, Y., and Vaughan, R. (2007). eGenomics: Cataloguing our complete genome collection III. *Comp. Funct. Genomics*, 2007:1–7.

[76] Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., Robeson, M., Edwards, R. A., Felts, B., Rayhawk, S., Knight, R., Rohwer, F., and Jackson, R. B. (2007). Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. *Applied and Environmental Microbiology*, 73(21):7059–7066.

[77] Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39 Suppl 2(May):W29–37.

[78] Forterre, P. and Philippe, H. (1999). Where is the root of the universal tree of life? *BioEssays : news and reviews in molecular, cellular and developmental biology*, 21(10):871–9.

[79] Fox, G. E., Pechman, K. R., and Woese, C. R. (1977). Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. *International Journal of Systematic Bacteriology*, 27(1):44–57.

[80] Francis, C. A., Beman, J. M., and Kuypers, M. M. M. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *The ISME journal*, 1(1):19–27.

[81] Frank, D. N., ST Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13780–13785.

[82] Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America*, 105(10):3805–3810.

[83] Fuhrman, J. A. and Hagström, A. k. (2008). Bacterial and archaeal community structure and its patterns. In Kirchman, D. L., editor, *Microbial Ecology of the Oceans*, volume 2, pages 45–90. Wiley-Blackwell.

[84] Fuhrman, J. A., McCallum, K., and Davis, A. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, 356(6365):148–149.

[85] Garrity, G. M., Bell, J. A., and Lilburn, T. G. (2004). Taxonomic outline of the prokaryotes. In *Bergeys manual of systematic bacteriology*, number May. Springer-Verlag.

[86] Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S., and Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology*, 11(12):3132–3139.

[87] Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345(6270):60–63.

[88] Giovannoni, S. J., DeLong, E. F., Olsen, G. J., and Pace, N. R. (1988). Phylogenetic groupspecific oligodeoxynucleotide probes for identification of single microbial cells. *J. Bacteriol.*, 170:720–726.

[89] Giovannoni, S. J. and Stingl, U. (2005). Molecular diversity and ecology of microbial plankton. *Nature*, 437(7057):343–348.

[90] Glöckner, F. O., Babenzien, H. D., and Amann, R. (1998). Phylogeny and identification in situ of Nevskia ramosa. *Applied and environmental microbiology*, 64(5):1895–901.

[91] Glöckner, F. O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. (2000). Comparative 16S rRNA analysis

of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Applied and Environmental Microbiology*, 66(11):5053–5065.

[92] Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3):705–8.

[93] Grasso, C. and Lee, C. (2004). Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, 20(10):1546–1556.

[94] Grice, E., Kong, H., Renaud, G., Young, A., Bouffard, G., Blakesley, R., Wolfsberg, T., Turner, M., and Segre, J. (2008). A diversity profile of the human skin microbiota. *Genome Research*, 18(7):1043–1050.

[95] Grice, E. a., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., Turner, M. L., and Segre, J. a. (2009). Topographical and temporal diversity of the human skin microbiome. *Science (New York, N.Y.)*, 324(5931):1190–2.

[96] Grice, E. a., Snitkin, E. S., Yockey, L. J., Bermudez, D. M., Liechty, K. W., and Segre, J. a. (2010). Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33):14799–804.

[97] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

[98] Gupta, P. K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends in biotechnology*, 26(11):602–11.

[99] Gutell, R. R., Larsen, N., and Woese, C. R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.*, 58(1):10–26.

[100] Gutell, R. R., Schnare, M. N., and Gray, M. W. (1992). A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucleic Acids Research*, 20(Suppl):2095–2109.

[101] Haeckel, E. (1866). *Generelle Morphologie der Organismen : allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Decendenz-Theorie.* Berlin.

[102] Hamady, M. and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, 19(7):1141–52.

[103] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249.

[104] Hankeln, W., Buttigieg, P. L., Fink, D., Kottmann, R., Yilmaz, P., and Glöckner, F. O. (2010). MetaBar - a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinformatics*, 11(1):358.

[105] Hanner, R. (2009). Data Standards for BARCODE Records in INSDC (BRIs).

[106] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, a. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52.

[107] Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174.

[108] Hebert, P. D. N., Ratnasingham, S., and DeWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological sciences / The Royal Society*, 270 Suppl:S96–9.

[109] Heidelberg, K. B., Gilbert, J. A., and Joint, I. (2010). Marine genomics: at the interface of marine microbial ecology and biodiscovery. *Microbial biotechnology*, 3(5):531–543.

[110] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–9.

[111] Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME journal*, pages 1–9.

[112] Hernandez, T. and Kambhampati, S. (2004). Integration of biological sources: current systems and challenges ahead. *ACM SIgmod Record*, 33(3):51–60.

[113] Hewson, I. and Fuhrman, J. A. (2004). Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Applied and Environmental Microbiology*, 70(6):3425–3433.

[114] Hey, J. (2001). The mind of the species problem. *Trends in Ecology & Evolution*, 16(7):326–329.

[115] Hey, T. and Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In *Grid Computing*, number January 2003, pages 809–824. Wiley Online Library.

[116] Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, 18(6):341–343.

[117] Hong, S. H., Bunge, J., Jeon, S. O., and Epstein, S. S. (2006). Predicting microbial species richness. *Proc. Natl. Acad. Sci. U.S.A.*, 103(1):117–122.

[118] Horner-Devine, M. C., Carney, K. M., and Bohannan, B. J. M. (2004). An ecological perspective on bacterial biodiversity. *Proceedings of the Royal Society B Biological Sciences*, 271(1535):113–122.

[119] Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., and Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, 417(6884):63–67.

[120] Hunt, D. E., Klepac-Ceraj, V., Acinas, S. G., Gautier, C., Bertilsson, S., and Polz, M. F. (2006). Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Applied and Environmental Microbiology*, 72(3):2221–2225.

[121] Huse, S. M., Dethlefsen, L., Huber, J. A., Mark Welch, D., Relman, D. A., and Sogin, M. L. (2008). Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. *PLoS Genetics*, 4(11):10.

[122] Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386.

[123] Huws, S. a., Edwards, J. E., Kim, E. J., and Scollan, N. D. (2007). Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems. *Journal of microbiological methods*, 70(3):565–9.

[124] Jonasson, J., Olofsson, M., and Monstein, H.-J. (2002). Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16s rDNA fragments. *APMIS acta pathologica microbiologica et immunologica Scandinavica*, 115(5):668–677; discussion 678–679.

[125] Jonkers, H. M., Koh, I.-O., Behrend, P., Muyzer, G., and De Beer, D. (2005). Aerobic organic carbon mineralization by sulfate-reducing bacteria in the oxygen-saturated photic zone of a hypersaline microbial mat. *Microbial Ecology*, 49(2):291–300.

[126] Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *New York: Academic Press*, pages 21–132.

[127] Just, W. (2001). Computational complexity of multiple sequence alignment with SP-score. *Journal of computational biology*, 8(6):615–623.

[128] Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K., and Nakamura, Y. (2011). DDBJ progress report. *Nucleic acids research*, 39(Database issue):D22–7.

[129] Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856.

[130] Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, 33(2):511–8.

[131] Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–66.

[132] Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–98.

[133] Kellenberger, E., Ryter, a., and Sechaud, J. (1958). Electron microscope study of DNA-containing plasms. II. Vegetative and mature phage DNA as compared with normal bacterial nucleoids in different physiological states. *The Journal of biophysical and biochemical cytology*, 4(6):671–8.

[134] Kemena, C. and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465.

[135] Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., and Phillips, D. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666.

[136] Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307.

[137] Kibler, M. R. (2006). From the Mendeleev periodic table to particle physics and back to the periodic table. *Foundations of Chemistry*, page 15.

[138] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120.

[139] Kirchman, D. L., Cottrell, M. T., and Lovejoy, C. (2010). The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environmental Microbiology*, 12(5):1132–1143.

[140] Koch, R. (1877). Die Atiologie der Milzbrand-Krankheit, begrundet auf die Entwicklungsgeschichte des Bacillus anthracis. Beitr. *Z. Biol. Pflanzen*, 2(2):277–310.

[141] Kong, Q.-P., Yao, Y.-G., Sun, C., Bandelt, H.-J., Zhu, C.-L., and Zhang, Y.-P. (2003). Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *American journal of human genetics*, 73(3):671–6.

[142] Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., and Glöckner, F. O. (2008). A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *Omics a journal of integrative biology*, 12(2):115–21.

[143] Kottmann, R., Kostadinov, I., Duhaime, M. B., Buttigieg, P. L., Yilmaz, P., Hankeln, W., Waldmann, J., and Glöckner, F. O. (2010). Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Research*, 38(Database issue):D391–D395.

[144] Kress, W. J., Wurdack, K. J., Zimmer, E. a., Weigt, L. a., and Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):8369–74.

[145] Kumar, Y., Westram, R., Behrens, S., Fuchs, B., Glöckner, F. O., Amann, R., Meier, H., and Ludwig, W. (2005). Graphical representation of ribosomal RNA probe accessibility data using ARB software package. *BMC Bioinformatics*, 6(1):61.

[146] Kumar, Y., Westram, R., Kipfer, P., Meier, H., and Ludwig, W. (2006). Evaluation of sequence alignments and oligonucleotide probes with respect

to three-dimensional structure of ribosomal RNA using ARB software package. *BMC Bioinformatics*, 7(1):240.

[147] Kutschera, U. and Niklas, K. J. (2004). The modern theory of biological evolution: an expanded synthesis. *Naturwissenschaften*, 91(6):255–276.

[148] Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H.-H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, 35(9):3100–3108.

[149] Lake, J. A., Skophammer, R. G., Herbold, C. W., and Servin, J. A. (2009). Genome beginnings: rooting the tree of life. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 364(1527):2177–2185.

[150] Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences of the United States of America*, 82(20):6955–6959.

[151] Lane, D. S. (1991). 16S and 23S rRNA sequencing. In Stackebrandt, E. and Goodfellow, M., editor, *Nucleic acid techniques in bacterial systematics*, pages 115–148. John Wiley & Sons, New York.

[152] Lassmann, T. and Sonnhammer, E. L. L. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, 6:298.

[153] Laurin, M. and Reisz, R. (1995). A reevaluation of early amniote phylogeny. *Zoological Journal of the Linnean Society*, 113(2):165–223.

[154] Lazarevic, V., Whiteson, K., Hernandez, D., François, P., and Schrenzel, J. (2010). Study of inter- and intra-individual variations in the salivary microbiota. *BMC Genomics*, 11(1):523.

[155] Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464.

[156] Lee, M. S. Y. (2003). Species concepts and species reality: salvaging a Linnaean rank. *Journal of Evolutionary Biology*, 16(2):179–188.

[157] Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., Cheng, Y., Demiralp, F., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Hunter, C., Jang, M., Leonard, S., Lin, Q., Lopez, R., Maguire, M., McWilliam, H., Plaister, S., Radhakrishnan, R., Sobhany, S., Slater, G., Ten Hoopen, P., Valentin, F., Vaughan, R., Zalunin, V., Zerbino, D., and Cochrane,

G. (2010). Improvements to services at the European Nucleotide Archive. *Nucleic Acids Research*, 38(Database issue):D39–D45.

[158] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., and Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic acids research*, 39(Database issue):D28–31.

[159] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8.

[160] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–9.

[161] Linnaeus, C. (1758). *Systema Naturae*, volume 1. Laurentii Salvii.

[162] Liolios, K., Chen, I.-M. A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., and Kyrpides, N. C. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 38(Database issue):D346–D354.

[163] Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science (New York, N.Y.)*, 227(4693):1435–41.

[164] Liu, Z., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, 36(18):e120.

[165] Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., and Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research*, 35(18):e120.

[166] Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C., and Glöckner, F. O. (2006). Megx.net–database resources for marine ecological genomics. *Nucleic Acids Res.*, 34:D390–393.

[167] López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, 409(6820):603–7.

[168] Loy, A., Maixner, F., Wagner, M., and Horn, M. (2007). probeBase–an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic acids research*, 35(Database issue):D800–4.

[169] Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (New York, N.Y.)*, 320(5883):1632–5.

[170] Ludwig, W., Klenk, H. P., and Garrity, G. M. (2001). A phylogenetic backbone and taxonomic framework for prokaryotic systematics. volume 1, pages 49–65. Springer-Verlag.

[171] Ludwig, W., Rossello-Mora, R., Aznar, R., Klugbauer, S., Spring, S., Reetz, K., Beimfohr, C., Brockmann, E., Kirchhof, G., Dorn, S., Bachleitner, M., Klugbauer, N., Springer, N., Lane, D., Nietupsky, R., Weizenegger, M., and Schleifer, K. H. (1995). Comparative sequence analysis of 23S rRNA from Proteobacteria. *Systematic and Applied Microbiology*, 18:164–188.

[172] Ludwig, W. and Schleifer, K. (1999). Phylogeny of bacteria beyond the 16S rRNA standard.

[173] Ludwig, W., Schleifer, K., and Whitman, W. (2009). Revised road map to the phylum Firmicutes. In Whitman, W., editor, *Bergey's Manual of Systematic Bacteriology*, pages 1–13. Springer, New York, 2nd edition.

[174] Ludwig, W. and Schleifer, K. H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol Rev*, 15(2-3):155–173.

[175] Ludwig, W. and Schleifer, K. H. (2005). Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes. In Sapp, J., editor, *Microbial phylogeny and evolution, concepts and controversies*, pages 70–98. Oxford university press, New York.

[176] Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumeier, J., Bachleitner, M., and Schleifer, K.-H. (1998). Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis*, 19:554–568.

[177] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis,

A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.-H., and Lüß mann, R. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research*, 32(4):1363–1371.

[178] Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science (New York, N.Y.)*, 302(5649):1401–4.

[179] Manz, W., Amann, R., Ludwig, W., Vancanneyt, M., and Schleifer, K. H. (1996). Application of a suite of 16S rRNA-specific oligonucleotide probes designed to investigate bacteria of the phylum cytophaga-flavobacter-bacteroides in the natural environment. *Microbiology (Reading, England)*, 142 (Pt 5(1 996):1097–106.

[180] Manz, W., Amann, R., Ludwig, W., Wagner, M., and Schleifer, K. H. (1992). Phylogenetic oligodeoxynucleotide probes for the major subclasses of proteobacteria: problems and solutions. *Systematic and Applied Microbiology*, 15(4):593–600.

[181] Marchesi, J. R., Sato, T., Weightman, A. J., Martin, T. A., Fry, J. C., Hiom, S. J., and Wade, W. G. (1998). Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Appl. Environ. Microbiol.*, 64(2):795–799.

[182] Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141.

[183] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80.

[184] Martinez, A., Tyson, G. W., and Delong, E. F. (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environmental microbiology*, 12(1):222–38.

[185] Massana, R., Murray, A. E., Preston, C. M., and DeLong, E. F. (1997). Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Applied and Environmental Microbiology*, 63(1):50–56.

[186] Mayden, R. L. (1997). A hierarchy of species concepts: the denouement in the saga of the species problem. In Claridge, M. F., editor, *Species the units of biodiversity*, volume 54, pages 381–424. Chapman & Hall.

[187] McLoughlin, N., Furnes, H., Banerjee, N., Muehlenbachs, K., and Staudigel, H. (2009). Ichnotaxonomy of microbial trace fossils in volcanic glass. *Journal of the Geological Society*, 166(1):159.

[188] Medini, D., Serruto, D., Parkhill, J., Relman, D. A., Donati, C., Moxon, R., Falkow, S., and Rappuoli, R. (2008). Microbiology in the post-genomic era. *Nature Reviews Microbiology*, 6(6):419–430.

[189] Mendel, G. (1865). Versuche über Planzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, IV:3–47.

[190] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46.

[191] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386.

[192] Minz, D., Flax, J. L., Green, S. J., Muyzer, G., Cohen, Y., Wagner, M., Rittmann, B. E., and Stahl, D. A. (1999). Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. *Applied and Environmental Microbiology*, 65(10):4666–71.

[193] Moon-Van Der Staay, S. Y., De Wachter, R., and Vaulot, D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, 409(6820):607–610.

[194] Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8):114ff.

[195] Morgenstern, B., Dress, a., and Werner, T. (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12098–103.

[196] Moritz, C. and Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biology*, 2(10).

[197] Morrison, D. A. and Ellis, J. T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Molecular Biology and Evolution*, 14(4):428–441.

[198] Muerta, M., Haseltine, F., Liu, Y., Downing, G., and Seto, B. (2000). NIH Working Definition of Bioinformatics and Computational Biology.

[199] Munoz, R., Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F. O., and Rosselló-Móra, R. (2011). Release LTPs104 of the All-Species Living Tree. *Systematic and applied microbiology*, 34(3):169–70.

[200] Muyzer, G., Brinkhoff, T., Nübel, U., Santegoeds, C., Schäfer, H., and Wawer, C. (1998). Denaturing gradient gel electrophoresis (DGGE) in microbial ecology. In Akkermans, A. D. L., Van Elsas, J. D., and De Bruijn, F. J., editors, *Molecular Microbial Ecology Manual*, volume 3, pages 1–27. Kluwer Academic Publishers, Dordrecht, The Netherlands.

[201] Muyzer, G., de Waal, E. C., and Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.*, 59(3):695–700.

[202] Muyzer, G., Teske, A., Wirsen, C. O., and Jannasch, H. W. (1995). Phylogenetic relationships of Thiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Archives of Microbiology*, 164(3):165–172.

[203] Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *Comput Appl Biosci*, 4(1):11–17.

[204] Nawrocki, E. P. and Eddy, S. R. (2008). Infernal 1 . 0 : RNA sequence analysis with covariance models. *BMC Bioinformatics*, pages 2008–2008.

[205] Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337.

[206] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–53.

[207] Neef, A. (1992). *Application of in situ identification of bacteria to population analysis in complex microbial communities*. PhD thesis, Technical University of Munich.

[208] Nossa, C. W., Oberdorf, W. E., Yang, L., Aas, J. r. A., Paster, B. J., DeSantis, T. Z., Brodie, E. L., Malamud, D., Poles, M. A., and Pei, Z. (2010). Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World Journal of Gastroenterology*, 16(33):4135–4144.

[209] Notredame, C. (2007). Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Computational Biology*, 3(8):4.

[210] Notredame, C. and Higgins, D. G. (1996). SAGA: sequence alignment by genetic algorithm. *Nucleic acids research*, 24(8):1515–24.

[211] Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17.

[212] Nübel, U., Engelen, B., Felske, A., Snaidr, J., Wieshuber, A., Amann, R. I., Ludwig, W., and Backhaus, H. (1996). Sequence heterogeneities of genes encoding 16S rRNAs in Paenibacillus polymyxa detected by temperature gradient gel electrophoresis. *Journal Of Bacteriology*, 178(19):5636–5643.

[213] Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., and Stahl, D. A. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.*, 40:337–365.

[214] Ovreas, L., Forney, L., Daae, F. L., Ø vreås, L., Forney, L., and Daae, F. L. (1997). Distribution of bacterioplankton in meromictic Lake Saelenvannet , as determined by denaturing gradient gel electrophoresis of PCR-amplified gene fragments coding for 16S rRNA Distribution of Bacterioplankton in Meromictic Lake Sælenvannet , as Determined. *Appl Environ Microbiol*, 63(9):3367–3373.

[215] Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276:734–740.

[216] Pace, N. R., Stahl, D. A., Olsen, G. J., and Lane, D. J. (1985). Analyzing natural microbial populations by rRNA sequences. *ASM News*, 51:4–12.

[217] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448.

[218] Pedros-Alio, C. (2006). Marine microbial diversity: can it be determined? *Trends Microbiol.*, 14(6):257–263.

[219] Pei, A., Nossa, C. W., Chokshi, P., Blaser, M. J., Yang, L., Rosmarin, D. M., and Pei, Z. (2009). Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. *PLoS ONE*, 4(5):9.

[220] Pei, J. (2008). Multiple protein sequence alignment. *Current Opinion in Structural Biology*, 18(3):382–386.

[221] Peplies, J., Glöckner, F. O., Amann, R., and Ludwig, W. (2004). Comparative sequence analysis and oligonucleotide probe design based on 23S rRNA genes of Alphaproteobacteria from North Sea bacterioplankton. *System. Appl. Microbiol.*, 27(5):573–580.

[222] Peplies, J., Kottmann, R., Ludwig, W., and Glöckner, F. O. (2008). A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. *Systematic and applied microbiology*, 31(4):251–257.

[223] Pernthaler, A., Pernthaler, J., and Amann, R. (2004). Sensitive multicolor fluorescence in situ hybridization for the identification of environmental microorganisms. *Molecular microbial ecology manual, 2nd ed. Kluwer Academic Publishers, Dordrecht, The Netherlands*, pages 711–726.

[224] Pommier, T., Canbäck, B., Riemann, L., Boström, K. H., Simu, K., Lundberg, P., Tunlid, A., and Hagström, A. (2007). Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology*, 16(4):867–880.

[225] Pruesse, E. (2007). *Incremental Approach to Multiple Sequence Alignment using Directed Acyclical Graphs.* Diplomarbeit, Universität Bremen.

[226] Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196.

[227] Pruesse, E., Quast, C., Yilmaz, P., Ludwig, W., Peplies, J., and Glöckner, F. O. (2011). SILVA: comprehensive databases for quality checked and aligned ribosomal RNA sequence data compatible with ARB. In de Bruijn, F. J., editor, *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, pages 393–398. John Wiley & Sons.

[228] Queiroz, K. D. (1998). The general lineage concept of species, species criteria, and the process of speciation. In Howard, D. J. and Berlocher, S. H., editors, *Endless Forms Species and Speciation*, number 1331, chapter 5, pages 57–75. Oxford University Press.

[229] Raghava, G. P. S., Searle, S. M. J., Audley, P. C., Barber, J. D., and Barton, G. J. (2003). OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, 4:47.

[230] Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual review of microbiology*, 57:369–94.

[231] Reeder, J. and Knight, R. (2009). The 'rare biosphere': a reality check. *Nature Methods*, 6(9):636–637.

[232] Reynolds, E. (1963). The use of lead citrate at high pH as an electron-opaque stain in electron microscopy. *The Journal of Cell Biology*, 17(1):208.

[233] Reysenbach, A.-L. and Pace, N. (1995). Reliable Amplification of Hyperthermophilic Archaeal 16S rRNA Genes by the Polymerase Chain Reaction. In Robb, F. and Place, A., editors, *Archaea: a laboratory manual*, pages 101–107. Cold Spring Harbour Laboratory Press, New York.

[234] Rice, C. M., Fuchs, R., Higgins, D. G., Stoehr, P. J., and Cameron, G. N. (1993). The EMBL data library. *Nucleic acids research*, 21(13):2967–71.

[235] Rijk, P., Peer, Y., Broeck, I., and Wachter, R. (1995). Evolution according to large tribosomal subunit RNA. *Journal of Molecular Evolution*, 41(3):366–375.

[236] Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., and Sansone, S.-A. (2010). ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356.

[237] Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K. M., Kent, A. D., Daroub, S. H., Camargo, F. A. O., Farmerie, W. G., and Triplett, E. W. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME journal*, 1(4):283–290.

[238] Röntgen, W. (1898). Ueber eine neue Art von Strahlen. *Annalen der Physik*, 300(1):1–11.

[239] Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., and Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*, 5(3):34.

[240] Sabatini, D. D., Bensch, K., and Barrnett, R. J. (1963). Cytochemistry and electron microscopy. The preservation of cellular ultrastructure and enzymatic activity by aldehyde fixation. *The Journal of cell biology*, 17:19–58.

[241] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

[242] Sandler, I. and Sandler, L. (1986). On the Origin of Mendelian Genetics. *Amer. Zool.*, 26:753–768.

[243] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–54637.

[244] Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240.

[245] Schaller, R. (1997). Moore's law: past, present and future. *IEEE Spectrum*, 34(6):52–59.

[246] Scherzer, O. (1949). The Theoretical Resolution Limit of the Electron Microscope. *Journal of Applied Physics*, 20(1):20.

[247] Schloss, P. D. (2009). A high-throughput DNA sequence aligner for microbial ecology studies. *PloS one*, 4(12):e8230.

[248] Schloss, P. D. and Handelsman, J. (2006). Toward a Census of Bacteria in Soil. *PLoS Computational Biology*, 2(7):8.

[249] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent,

Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.

[250] Schmidt, H. A., Strimmer, K., Vingron, M., and Von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–504.

[251] Schweer, T. (2011). *Qualitätsmanagement ribosomaler RNA Sequencen in der SILVA Datenbank*. Diplomarbeit, Fachhochschule Bingen.

[252] Sellers, P. H. (1974). On the Theory and Computation of Evolutionary Distances. *SIAM Journal on Applied Mathematics*, 26(4):787–793.

[253] Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: A Community Resource for Metagenomics. *PLoS Biology*, 5(3):4.

[254] Shaw, K. L. (2002). Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16122–7.

[255] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45.

[256] Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *Fems Micriobiology Ecology*, 60(2):341–350.

[257] Sites Jr, J. W. and Marshall, J. C. (2003). Delimiting species: a Renaissance issue in systematic biology. *Trends in Ecology & Evolution*, 18(9):462–470.

[258] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

[259] Smith, T. F., Waterman, M. S., and Fitch, W. M. (1981). Comparative biosequence metrics. *J Mol Evol*, 18(1):38–46.

[260] Snigirev, a., Bjeoumikhov, a., Erko, a., Snigireva, I., Grigoriev, M., Yunkin, V., Erko, M., and Bjeoumikhova, S. (2007). Two-step hard X-ray focusing combining Fresnel zone plate and single-bounce ellipsoidal capillary. *Journal of synchrotron radiation*, 14(Pt 4):326–30.

[261] Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.*, 103(32):12115–12120.

[262] Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, 38(February):1409–1438.

[263] Stahl, D. A. and Amann, R. (1991). Development and application of nucleic acid probes. In Stackebrandt, E. and Goodfellow, M., editors, *Nucleic acid techniques in bacterial systematics*, number 8, pages 205–248. John Wiley and Sons.

[264] Stahl, D. A., Lane, D. J., Olsen, G. J., and Pace, N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science*, 224(4647):409–411.

[265] Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

[266] Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol*, 57(5):758–771.

[267] Stamatakis, A. and Ott, M. (2008). Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. *Philos Trans R Soc Lond B Biol Sci*, 363(1512):3977–3984.

[268] Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome biology*, 11(5):207.

[269] Stevens, H. and Ulloa, O. (2008). Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environmental Microbiology*, 10(5):1244–1259.

[270] Subramanian, A. R., Kaufmann, M., and Morgenstern, B. (2008). DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for molecular biology : AMB*, 3:6.

[271] Swofford, D. (1993). PAUP: Phylogenetic Analysis Using Parsimony (PAUP).

[272] Sze, S.-H., Lu, Y., and Yang, Q. (2006). A polynomial time solvable formulation of multiple sequence alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 13(2):309–19.

[273] Takai, K. and Horikoshi, K. (2000). Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Applied and Environmental Microbiology*, 66(11):5066–5072.

[274] Tanaka, M., Hadjantonakis, A. K., and Nagy, A. (2001). Aggregation chimeras. Combining ES cells, diploid and tetraploid embryos. *Methods in molecular biology (Clifton, N.J.)*, 158(20):135–54.

[275] Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., and Vogler, A. P. (2002). DNA points the way ahead of taxonomy - In assessing new approaches, it's time for DNA's unique contribution to take a central role. *Nature*, 418(6897):479.

[276] Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *questions in biology-DNA sequence analysis*.

[277] Taylor, C., Field, D., Sansone, S.-A., Apweiler, R., Ashburner, M., Ball, C., Binz, P.-A., Brazma, A., Brinkman, R., Deutsch, E., Fiehn, O., Fostel, J., Ghazal, P., Brimes, G., Hardy, N., Hermjakob, H., Julian, R., Kuiper, M., Le Novere, N., Leebens-Mack, J., Lewis, S., McNally, R., Morrison, N., Paton, N., Quackenbush, J., Robertson, D., Rocca-Serra, P., Smith, B., Snape, J., and Wiemann, S. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26(8):889–896.

[278] Teske, A., Hinrichs, K.-U., Edgcomb, V., De Vera Gomez, A., Kysela, D., Sylva, S. P., Sogin, M. L., and Jannasch, H. W. (2002). Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. *Applied and Environmental Microbiology*, 68(4):1994–2007.

[279] Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. *Nature*, 465(7295):219–22.

[280] Thompson, J. D., Higgins, D. G., and 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(22):4673–4680.

[281] Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–136.

[282] Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLoS ONE*, 6(3):14.

[283] Thompson, J. D., Plewniak, F., and Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–88.

[284] Thompson, S. and Parthasarathy, S. (2006). Moore's law: the future of Si microelectronics. *Materials Today*, 9(6):20–25.

[285] Tindall, B. J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W., and Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *International journal of systematic and evolutionary microbiology*, 60(Pt 1):249–66.

[286] Tringe, S. G. and Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 11(5):442–446.

[287] Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P., and Rubin, E. M. (2005). Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557.

[288] Tung, C.-S., Joseph, S., and Sanbonmatsu, K. Y. (2002). All-atom homology model of the Escherichia coli 30S ribosomal subunit. *Nature Structural Biology*, 9(10):750–755.

[289] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164):804–810.

[290] Van Camp, G., Chapelle, S., and De Wachter, R. (1993). Amplification and sequencing of variable regions in bacterial 23S ribosomal RNA genes with conserved primer sequences. *Current Microbiology*, 27(3):147–151.

[291] Van Walle, I., Lasters, I., and Wyns, L. (2004). Align-m–a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics (Oxford, England)*, 20(9):1428–35.

[292] Van Walle, I., Lasters, I., and Wyns, L. (2005). SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics (Oxford, England)*, 21(7):1267–8.

[293] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

[294] Verslyppe, B., Slabbinck, B., Smet, W. D., Vos, P. D., Baets, B. D., and Dawyndt, P. (2008). StrainInfo.net Web Services: Enabling Microbiologic Workflows Such as Phylogenetic Tree Building and Biomarker Comparison. In *ESCIENCE '08: Proceedings of the 2008 Fourth IEEE International Conference on eScience*, pages 603–607, Washington, DC, USA. IEEE Computer Society.

[295] Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., Elsas, V., Dirk, J., Bailey, M. J., Nalin, R., and Philippot, L. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology*, 7(4):252–252.

[296] Wallace, A. (1858). On the tendency of varieties to depart indefinitely from the original type. *Proceedings of the Linnean Society of London*, pages 1–9.

[297] Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic acids research*, 34(6):1692–9.

[298] Walters, W. A., Caporaso, J. G., Lauber, C. L., Berg-Lyons, D., Fierer, N., and Knight, R. (2011). PrimerProspector: de novo design and taxonomic analysis of barcoded PCR primers. *Bioinformatics*, 27(8):2–4.

[299] Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 1(4):337–48.

[300] Wang, Y. and Qian, P.-Y. (2009). Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies. *PLoS ONE*, 4(10):9.

[301] Ward, D. M., Weller, R., and Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270):63–65.

[302] Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[303] Werner, J. J., Koren, O., Hugenholtz, P., Desantis, T. Z., Walters, W. a., Caporaso, J. G., Angenent, L. T., Knight, R., and Ley, R. E. (2011). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *The ISME journal*, pages 1–10.

[304] Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–5090.

[305] Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4576–9.

[306] Wommack, K. E., Bhavsar, J., and Ravel, J. (2008). Metagenomics: Read Length Matters. *Applied and Environmental Microbiology*, 74(5):1453–1463.

[307] Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2):e1000667.

[308] Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., Hooper, S. D., Pati, A., Lykidis, A., Spring, S., Anderson, I. J., D'Haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J.-F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E. M., Kyrpides, N. C., Klenk, H.-P., and Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276):1056–1060.

[309] Wuyts, J., De Rijk, P., de Peer, Y., Winkelmans, T., and De Wachter, R. (2001). The European Large Subunit Ribosomal RNA Database. *Nucleic Acids Res.*, 29(1):175–177.

[310] Wuyts, J., Perriere, G., and de Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.*, 32(suppl_1):D101–103.

[311] Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F. O., and Rosselló-Móra, R. (2010). Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Systematic and applied microbiology*, 33(6):291–9.

[312] Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F. O., and Rosselló-Móra, R. (2008). The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol*, 31(4):241–250.

[313] Yeates, C., Saunders, A. M., Crocetti, G. R., and Blackall, L. L. (2003). Limitations of the widely used GAM42a and BET42a probes targeting bacteria in the Gammaproteobacteria radiation. *Microbiology*, 149(Pt 5):1239–1247.

[314] Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B. W., Blaser, M. J., Bonazzi, V., Booth, T., Bork, P., Bushman, F. D., Buttigieg, P. L., Chain, P. S. G., Charlson, E., Costello, E. K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J. A., Gallery, R. E., Gevers, D., Gibbs, R. A., Gil, I. S., Gonzalez, A., Gordon, J. I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A. L., Kelley, S. T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C. L., Legg, T., Ley, R. E., Lozupone, C. A., Ludwig, W., Lyons, D., Maguire, E., Methe, B. A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K. E., Nemergut, D., Neufeld, J. D., Newbold, L. K., Oliver, A. E., Pace, N. R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D. A., Assunta-Sansone, S., Schloss, P. D., Schriml, L., Sinha, R., Smith, M. I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J. M., Ward, D. V., Weinstock, G. M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J. R., Yatsunenko, T., and Glockner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotech*, 29(5):415–420.

[315] Yooseph, S., Nealson, K. H., Rusch, D. B., McCrow, J. P., Dupont, C. L., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., Beeson, K., Williamson, S. J., Tovchigrechko, A., Allen, A. E., Zeigler, L. A., Sutton, G., Eisenstadt, E., Rogers, Y.-H., Friedman, R., Frazier, M., and Venter, J. C. (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468(7320):60–66.

[316] Zehr, J. P., Mellon, M. T., and Zani, S. (1998). New Nitrogen-Fixing Microorganisms Detected in Oligotrophic Oceans by Amplification of Nitrogenase (nifH) Genes. *Applied and Environmental Microbiology*, 64(9):3444–3450.

[317] Zheng, D., Alm, E. W., Stahl, D. A., and Raskin, L. (1996). Characterization of universal small-subunit rRNA hybridization probes for quantitative molecular microbial ecology studies. *Applied and Environmental Microbiology*, 62(12):4504–4513.

[318] Zhou, J., Bruns, M. a., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Applied and environmental microbiology*, 62(2):316–22.

[319] Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Analysis*, 97:97–166.