Journal of Biotechnology 261 (2017) 169–176

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

25 years of serving the community with ribosomal RNA gene reference databases and tools

Frank Oliver Glöckner[a,b,*], Pelin Yilmaz[b], Christian Quast[b], Jan Gerken[a], Alan Beccati[a], Andreea Ciuprina[a], Gerrit Bruns[a], Pablo Yarza[c], Jörg Peplies[c], Ralf Westram[c], Wolfgang Ludwig[d]

[a] Department of Life Sciences and Chemistry, Jacobs University gGmbH, Bremen, Germany
[b] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[c] Ribocon GmbH, D-28359 Bremen, Germany
[d] Department for Microbiology, Technical University Munich, D-85354 Freising, Germany

ARTICLE INFO

Keywords:
Ribosomal RNA
Alignments
Taxonomy
Phylogeny
Biodiversity
Databases

ABSTRACT

SILVA (lat. forest) is a comprehensive web resource, providing services around up to date, high-quality datasets of aligned ribosomal RNA gene (rDNA) sequences from the Bacteria, Archaea, and Eukaryota domains. SILVA dates back to the year 1991 when Dr. Wolfgang Ludwig from the Technical University Munich started the integrated software workbench ARB (lat. tree) to support high-quality phylogenetic inference and taxonomy based on the SSU and LSU rDNA marker genes. At that time, the ARB project maintained both, the sequence reference datasets and the software package for data analysis. In 2005, with the massive increase of DNA sequence data, the maintenance of the software system ARB and the corresponding rRNA databases SILVA was split between Munich and the Microbial Genomics and Bioinformatics Research Group in Bremen. ARB has been continuously developed to include new features and improve the usability of the workbench. Thousands of users worldwide appreciate the seamless integration of common analysis tools under a central graphical user interface, in combination with its versatility.

The first SILVA release was deployed in February 2007 based on the EMBL-EBI/ENA release 89. Since then, full SILVA releases offering the database content in various flavours are published at least annually, complemented by intermediate web-releases where only the SILVA web dataset is updated. SILVA is the only rDNA database project worldwide where special emphasis is given to the consistent naming of clades of uncultivated (environmental) sequences, where no validly described cultivated representatives are available. Also exclusive for SILVA is the maintenance of both comprehensive aligned 16S/18S rDNA and 23S/28S rDNA sequence datasets. Furthermore, the SILVA alignments and trees were designed to include Eukaryota, another unique feature among rDNA databases. With the termination of the European Ribosomal RNA Database Project in 2007, the SILVA database has become the authoritative rDNA database project for Europe. The application spectrum of ARB and SILVA ranges from biodiversity analysis, medical diagnostics, to biotechnology and quality control for academia and industry.

1. Introduction

Pioneered by Fox et al. (1977) more than 40 years ago and propagated by Giovannoni et al. (1988), Olsen et al. (1986), Pace et al. (1985) and Ward et al. (1990), the use of ribosomal RNA (rRNA) molecule has become the “gold-standard” for nucleic acid-based investigations of microbial diversity, their taxonomic assignment and phylogenetic reconstructions (Amann et al., 1995; Fuhrman et al., 2015; Pace, 1997). The increasing awareness to catalogue and protect biodiversity on Earth, in combination with the advent of next generation sequencing technologies has further fuelled the interest in using the rRNA gene as a ‘barcode’ of life. The resulting millions of small and large subunit (SSU and LSU) rRNA gene sequences in the public archives require specialised tools and databases for alignment, analysis, phylogenetic inference, and classification. 25 years ago, in anticipation of the impending deluge of rDNA data, the development of the ARB software workbench and the curation of its rDNA databases was initiated (Ludwig et al., 2004). ARB offers a broad spectrum of interacting software tools built around a central database and complemented with a time-tested graphical user interface. From the

[*] Corresponding author at: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany.
E-mail address: fog@mpi-bremen.de (F.O. Glöckner).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.1198
Received 10 February 2017; Received in revised form 17 June 2017; Accepted 21 June 2017
Available online 23 June 2017

beginning, the ARB project provided phylogenetically structured integrative knowledge databases for small and large subunit rDNAs for Bacteria, Archaea, and Eukaryota. Until 2004, these datasets were maintained and deployed by manually collecting sequences from the International Nucleotide Sequence Database Collaboration (INSDC) (Leinonen et al., 2011), the European Ribosomal RNA Database (Wuyts et al., 2001) as well as the Ribosomal Database Project (Cole et al., 2014). Using the ARB workbench, all sequences were aligned, quality checked and used for comprehensive phylogenetic tree reconstruction.

In January 2004 Wolfgang Ludwig released the last SSU dataset comprising 59,609 sequences. Further releases were hampered by the rapid increase of sequence data, which could no longer be manually inspected and processed. In 2005, the decision was made to split the workload by separating the maintenance of the ARB software package from the production and dissemination of the rDNA datasets. This laid the foundation of the SILVA ribosomal RNA database project in the Microbial Genomics and Bioinformatics Group in Bremen. Within two years, a semi-automatic software pipeline was designed to mimic the manual curation process and amended by automated quality assessments. To compete with the growing amount of rDNA data that need to be aligned, the SILVA INcremental Aligner (SINA) was implemented and integrated (Pruesse et al., 2012). SINA is able to align millions of sequences within hours using a SEED alignment. In February 2007, SILVA released the first SSU (353,366 sequences) and LSU (46,979 sequences) datasets under the version number 89 and since then the SILVA release numbers follow the numbering of the EMBL-EBI/ENA releases.

The current SILVA database release 128 (September 2016) contains 5,616,941 SSU and 735,238 LSU rDNA sequences. All sequences are checked for anomalies and carry a rich set of sequence-associated contextual information, multiple taxonomic classifications (obtained from EMBL-EBI/ENA (INSDC) (Cochrane et al., 2012), RDP (Cole et al., 2014), Greengenes (DeSantis et al., 2006b), LTP (Yarza et al., 2008)) as well as the latest validly described nomenclature. SILVA maintains manually curated and non-public reference alignments of 75,000 16S/18S and 23S/28S ribosomal RNA genes (the SEED) in order to re-align all sequences for each SILVA release. With every full release, also a curated phylogenetic guide tree is provided that contains the latest taxonomy and nomenclature based on multiple references. The complete history of all releases is available on the SILVA website under ‘Documentation’ and ‘Archive’.

2. The ARB workbench

Powerful interoperable bioinformatics tools are inevitable for creating comprehensive multiple alignments and the reconstruction of phylogenetic trees, as well as the design of probes and primers for the in situ analysis of microbial communities. The two major objectives that were formulated at the beginning of the ARB project and were followed until today are: (1) the maintenance of a structured integrative secondary (knowledge) database of high-quality sequences, combining processed primary structures and any type of additional data assigned to the individual sequence entries, and (2) a comprehensive selection of directly interacting software tools, as well as a central database controlled via a common graphical user interface. Initially, the ARB package was designed for analysis of rDNA data only. Later, it was extended by developing and including features for managing protein sequences.

2.1. The ARB main window

The ARB main window provides the central workbench for accessing the various software tools and functions that are interacting with a local ARB sequence database. Users can select a tree representing the complete database or subsets thereof. To easily dive into the sequence space, the selected tree can be displayed in radial or dendrogram form. Any primary and metadata can be visualised at the terminal nodes.

2.2. The central database

The central component of ARB is a highly compressed hierarchical database. During operation it is loaded into the main memory (RAM) of the computer, ensuring rapid access and operation. The sequences representing genes (DNA) or gene products (rRNA or proteins) are stored in individual database fields.

2.3. Sequence editor

A powerful sequence editor facilitates user access to primary structure visualisation, arrangement, and modification (nucleotide or amino acid sequences). A special feature of the editor is the instant secondary structure check while rDNA alignments are visualised. Symbols indicating expected base pairings (or their absence) in the secondary structure of the native gene product are shown below each nucleotide symbol and constantly refreshed while editing the alignment. A three-domain consensus secondary structure mask, based on commonly accepted secondary structure models (Cannone et al., 2002), provides a guide for this tool.

2.4. Phylogenetic reconstruction

ARB hosts distance matrix, maximum likelihood, and maximum parsimony software tools for nucleotide and amino acid sequence based tree reconstruction. They directly cooperate with the respective ARB components and database elements such as alignment and filters.

ARB-parsimony has been specifically developed to handle several thousands of sequences (more than 600,000 in the current small subunit (SSU Ref NR 99) rDNA SILVA dataset (Quast et al., 2013)). New sequences are successively added to an existing tree according to the parsimony criterion. A special feature of ARB-parsimony allows adding sequences to an existing tree without altering the initial tree. This enables the user to include partial, low quality, or preliminarily aligned sequences, without disturbing the topology of an optimised tree constructed with high-quality data.

2.5. Probe design and evaluation

Gene or taxon-specific probes or primers (sequence signatures) are central for many molecular biological research and analysis projects. Examples are the delineation and identification of microorganisms in complex environmental samples (Amann et al., 1995) and the amplicon based analysis of microbial communities using next generation sequencing technologies. The ARB ‘Probe Design’ and ‘Probe Match’ tools use a suffix tree based search engine to identify short (10–100 k-mers) diagnostic sequence stretches, which are evaluated against the background of all sequences in the dataset.

2.6. Acceptance of ARB

The ARB workbench is used worldwide in academia and industry with an estimated user community of more than 10,000 users. The corresponding paper (Ludwig et al., 2004) has been cited 5431 times (Google Scholar, last assessed June 2017). The download statistics show an average combined download rate of 930 downloads for the full-text and the PDF file per year.

2.7. Future developments

Future developments of ARB are focused on the taxonomic curation of marker gene data sets. Various features are added to improve the topology comparison of phylogenetic trees and to integrate topology-based group detection and search functions. Additionally, the ARB-