Metadata of the chapter that will be visualized online

Chapter Title	All-Species Living Tree Project	
Copyright Year	2013	
Copyright Holder	Springer Science+Business Media New York	
Corresponding Author	Family Name	**Yarza**
	Particle	
	Given Name	**Pablo**
	Suffix	
	Organization	Ribocon GmbH.
	Address	Bremen, Bremen, 28359, Germany
	Email	pyarza@ribocon.com
Author	Family Name	**Munoz**
	Particle	
	Given Name	**Raul**
	Suffix	
	Division	Department of Ecology and Marine Resources
	Organization	Institution mediterrani d'Estudis Avançats (CSIC-UIB), Marine Microbiology Group
	Address	E-07190 Esporles, Illes Balears, Spain
	Email	raul@imedea.uib-csic.es
Author	Family Name	**Euzéby**
	Particle	
	Given Name	**Jean**
	Suffix	
	Organization	Society of Systematic Bacteriology and Veterinary (SBSV) & National Veterinary School de Toulouse (ENVT)
	Address	Cedex 03, F-31076, Toulouse, France
	Email	jean.euzeby@gmail.com
Author	Family Name	**Ludwig**
	Particle	
	Given Name	**Wolfgang**
	Suffix	
	Division	Lehrstuhl FÜR Mikrobiologie
	Organization	Technische UNIVERSITÄT München
	Address	Freising, D-85350, Germany

	Email	ludwig@mikro.biologie.tu-muenchen.de
Author	Family Name	**Schleifer**
	Particle	
	Given Name	**Karl-Heinz**
	Suffix	
	Division	Lehrstuhl FÜR Mikrobiologie
	Organization	Technische UNIVERSITÄT München
	Address	Freising, D-85350, Germany
	Email	schleife@mikro.biologie.tu-muenchen.de
Author	Family Name	**Amann**
	Particle	
	Given Name	**Rudolf**
	Suffix	
	Organization	Max Plank Institute for Marine Microbiology
	Address	Bremen, D-28359, Germany
	Email	ramann@mpi-bremen.de
Author	Family Name	**Glockner**
	Particle	
	Given Name	**Frank Oliver**
	Suffix	
	Division	Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology
	Organization	Jacobs University
	Address	Celsiusstrasse 1, Bremen, 28359, Germany
	Email	fog@mpi-bremen.de
Author	Family Name	**Rosselló-Mora**
	Particle	
	Given Name	**Ramon**
	Suffix	
	Organization	IMEDEA (CSIC-UIB)
	Address	Esporles, Mallorca, Balearic Islands, Spain
	Email	rossello-mora@uib.es

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

All-Species Living Tree Project

Pablo Yarza[a]*, Raul Munoz[b], Jean Euzéby[c], Wolfgang Ludwig[d], Karl-Heinz Schleifer[d], Rudolf Amann[e], Frank Oliver Glockner[f] and Ramon Rosselló-Mora[g]
[a]Ribocon GmbH., Bremen, Bremen, Germany
[b]Department of Ecology and Marine Resources, Institution mediterrani d'Estudis Avançats (CSIC-UIB), Marine Microbiology Group, Illes Balears, Spain
[c]Society of Systematic Bacteriology and Veterinary (SBSV) & National Veterinary School de Toulouse (ENVT), Toulouse, France
[d]Lehrstuhl FÜR Mikrobiologie, Technische UNIVERSITÄT München, Freising, Germany
[e]Max Plank Institute for Marine Microbiology, Bremen, Germany
[f]Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Jacobs University, Bremen, Germany
[g]IMEDEA (CSIC-UIB), Esporles, Mallorca, Balearic Islands, Spain

Synonyms

16SrRNA(SSU) and 23SrRNA(LSU) gene sequence databases; Alignments; LTP project; Manual curation; "Orphan" species; Taxa boundaries; Taxonomy/classification/phylogeny of Bacteria and Archaea; Type strains

Definition

The All-Species Living Tree Project (LTP) is an international initiative for the creation and maintenance of highly curated 16SrRNA and 23SrRNA gene sequence databases, alignments, and phylogenetic trees for all the type strains of *Bacteria* and *Archaea*.

Introduction

Classification and identification of *Bacteria* and *Archaea* came across to a turning point around 35 years ago. It was the time when Carl Woese and co-workers demonstrated that ribosomal markers were appropriate to infer genealogical relationships by means of phylogenetic reconstructions (Fox et al. 1977). Rapidly, comparative analysis of rRNA gene sequences became a standard procedure with mature implications in microbial ecology and taxonomy: culture-independent exploration of ecosystems' diversity (Amann et al. 1995) and settlement of the phylogenetic backbone (i.e., our current accepted classification of *Bacteria* and *Archaea*; Garrity 2001). As a result, the total amount of ribosomal RNA entries in the public DNA databases has grown exponentially since early 1990s, currently comprising at least 3,500,000 small (SSU) and 300,000 large (LSU) ribosomal subunit gene sequence entries. On the other hand, the number of bacterial and archaeal species with validly published names has followed arithmetic trends with a ratio of around 500–700 annual descriptions during the last 7 years (Fig. 1), currently (December 2012) exceeding

*Email: pyarza@ribocon.com

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

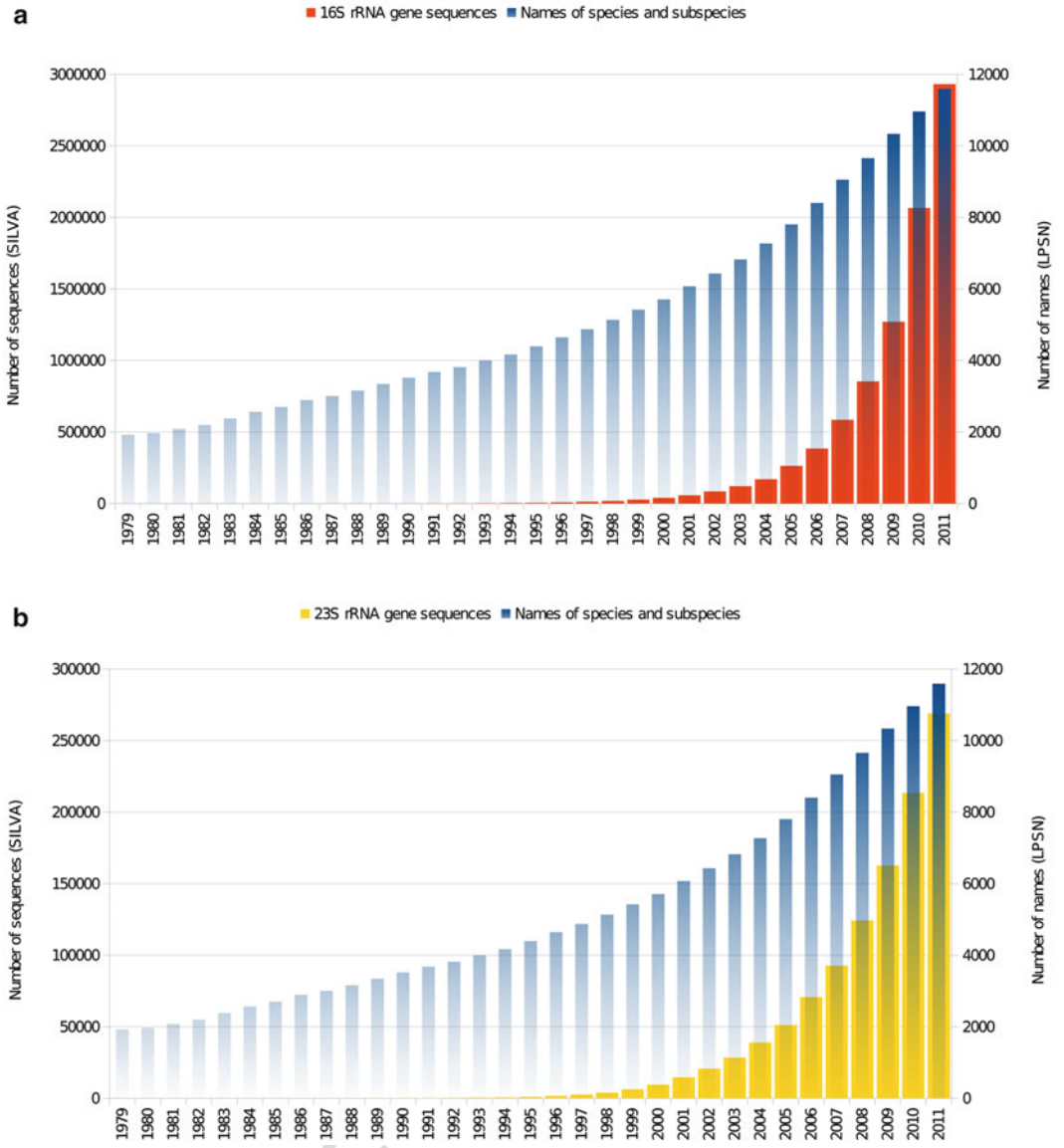

Fig. 1 Annual growth of ribosomal 16S rRNA (**a**) and 23S rRNA (**b**) gene sequence databases and species and subspecies names with standing in nomenclature until December 2011. SILVA SSU-Parc111 and LSU-Parc111 databases (http://www.arb-silva.de/documentation/release-111/) were filtered by submission date until December 2011 and its cumulative annual growth was plotted in red (SSU, 1A) and yellow bars (LSU, 1B). The cumulative growth of published species and subspecies names (according to LPSN; http://www.bacterio.cict.fr/number.html) since 1980 until December 2011 is plotted in blue. Note that the total number of names is around 2,000 above the total number of distinct type strains due to homotypic synonyms, new combinations, nomina nova, later heterotypic synonyms, or illegitimate names

the total number of 10,300 species and subspecies. A comparative overview of these trends until December 2011 is shown in Fig. 1.

As from early 1990s, the 16S rRNA has been, by orders of magnitude, the most often sequenced gene, there is no alternative phylogenetic marker with such a high coverage in public repositories. However, abundance is not the single requisite for a proper phylogenetic inference and other single molecules (e.g., 23S rRNA) or combinations of them might perform better at reflecting genealogies of certain groups given the higher information content (Ludwig and Klenk 2001). Although far from

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

reaching 16S rRNA levels, submission of alternative markers is growing fast, mostly because (i) the number of meta-genomes and complete genomes is growing exponentially due to the reduction on sequencing and analysis costs and (ii) the recent initiative to complete the genome sequence of all type strains (GEBA initiative). Undoubtedly, comparative genomics will involve a new breakthrough for microbial taxonomy and the current phylogenetic backbone based on ribosomal sequences will be carefully reviewed (Coenye et al. 2005). Nevertheless, at this point, the number of sequenced genomes of type strains is still low and therefore the current possibilities for an in-depth taxonomic study are sparse.

The responsible teams of the ARB, SILVA, and LPSN projects (www.arb-home.de, www.arb-silva.de, and www.bacterio.net) together with the journal Systematic and Applied Microbiology (SAM) started the "All-Species Living Tree Project" (LTP; http://www.arb-silva.de/projects/living-tree), a project conceived to provide a tool especially designed for the microbial taxonomist scientific community (Yarza et al. 2008). The main objectives considered so far are (1) provide a curated 16S and 23S rRNA gene database for the type strains of all species with validly published names; (2) set up an optimized and universally usable alignment; (3) reconstruct reliable phylogenetic trees with all the type strains; (4) maintain the database, alignments, and trees through regular updates including the new validly published taxa and their respective 16S and 23S rRNA gene sequences; and (5) investigate, with the use of the database, fundamental aspects about taxonomy of *Bacteria* and *Archaea* such as phylogenetic thresholds in new taxa circumscriptions, coherence of current taxonomy by means of phylogenetic schemes, and relevance of the ribosomal RNA genes in taxonomic studies.

Creation and Maintenance of LTP Releases

LTP Datasets

First LTP datasets (release LTPs93 for SSU (Yarza et al. 2008), release LTPs102 for LSU (Yarza et al. 2010)) were prepared following six main steps:

1. Set up a list of candidate sequences. An initial sequence dataset consisted on a subsample of the SILVA database, filtering by "type" (T) or "cultured" (C) strains; this information mainly came from StrainInfo.
2. Set up a list of species names. In parallel we built a comprehensive, updated, and nonredundant (i.e., free of synonyms and according to latest valid nomenclature) list of validly published species and subspecies names from LPSN. When a species is divided into subspecies, we substituted the original species name by that of the subspecies (e.g., *Staphylococcus sciuri* subsp. *sciuri* instead of *Staphylococcus sciuri*). We avoided the "Candidatus" names (e.g., "*Candidatus* Aciduliprofundum boonei"), *Cyanobacteria* not validly published under the Bacteriological Code (e.g., *Anabaena oscillatorioides*), and later heterotypic synonyms (e.g., *Pseudomonas chloritidismutans*).
3. Manual cross-check. Then, each entry from our initial list of sequences was assigned to a species name by manually examining the companion contextual metadata. This process had to be done manually given the often outdated, mistaken, or absent taxonomic information such as the organism name or the strain numbers.
4. Quest for missing type strains. We realized that not all species names were represented in the list of sequences. Then, we inverted the process by searching in resources like EMBL, Bergey's

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

Outlines, issues of the International Journal of Systematic and Evolutionary Microbiology (IJSEM), etc. with the aim to find a good-quality sequence entry for each missing type strain.

5. "Orphan" species recognition. Finally, we got a group of type strains whose 16S/23S rRNA genes had never been sequenced or that the existing sequences were of too low quality to be considered for the project (i.e., in terms of sequence length, number of ambiguities, etc.). We called them "orphan" species. The LTP project together with eleven international culture collections has driven the sequencing of these "orphan" species through the SOS initiative (Yarza et al. 2013).
6. Keep one sequence per species. On the other hand, the list of type-strain sequences was redundant in the sense that one single type strain could be represented by multiple sequence entries. This is the case of multiple independent sequencings and submissions, or the existence of several sequences due to multiple copies of the ribosomal operon. The aim of the LTP is, whenever possible, to keep one sequence per type strain in order to maintain simplicity, avoid confusion, and improve tree navigation and database usability. In general, the best quality available (including manual inspection of the alignment) was selected for the project and, in case of doubt, the earliest submission to an INSDC partner (www.insdc.org). From release LTPs102 (Yarza et al. 2010), when multiple paralogues exist due to rRNA operon copy number, several copies are kept if they show less than 98 % sequence identity (see below for further details).

LTP is maintained by a scrutiny of the new described species, nomenclatural changes, taxonomic notes, and opinions that are monthly published in the IJSEM journal. Their respective 16S and 23S rRNA gene sequence entries are acquired from the latest SILVA release and appended to the existing LTP database. Therefore, SILVA's Reference (Ref) ARB databases (http://www.springerreference.com/docs/html/chapterdbid/304116.html) serve as template for the new LTP-ARB databases. Until now (December 2012) one LSU-based and seven SSU-based LTP releases have been produced (Table 1). New species are incorporated into the database only if they account a good-quality sequence existing in the respective SILVA release. Certain entries can be deleted if their corresponding species names are seen to be later heterotypic synonyms, if they become rejected, or as a matter of taxonomic opinions. Sequence entries existing in an LTP database can also change by means of their metadata. Thus, for example, new combinations (i.e., a type strain which changes its name due to reclassification) or subdivision of a species into subspecies leads to an entry modification at its taxonomic information fields.

Inaccurate or Mistaken Metadata

Inaccurate sequence-associated metadata tend to happen in more than 50 % of the new added 16S rRNA entries (Table1). Often, these "mistakes" consist on a lack of entries' updating tasks at the time of their first appearance in a scientific publication. It mainly occurs in taxonomy-associated information fields. To prove the uniqueness of a new species and to name it take time and, in the meanwhile, sequences are quickly produced and easily submitted to nucleotide databases. Most often, these submissions only show genus specifications, for example, sequence entry GU808562 appears as "Hymenobacter sp. HMD1010" but its real name is *Hymenobacter yonginensis*. Indeed, a Bacteriological Code-compliant (Lapage et al. 1992) nomenclature may be somewhat tricky and is frequent to consider several Latin terms and derivations until one species name is finally accepted by authors and reviewers. Unavoidably, this bad-quality information is propagated from INSDC databases (primary sources) to other technological services like dedicated ribosomal databases (e.g., SILVA). Although extensive data curation is not a task of primary sources of information, it would be very beneficial that authors enhance their commitment with the correctness of the metadata provided (e.g., like the species name) or that authors are forced to update their INSDC entries prior to

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

t1.1 **Table 1** Summary of LTP releases. "Sync" fields correspond to IJSEM and EMBL release dates. "Net increase" of a release is the number of new entries minus the number of deleted entries. "% incorrect" refers to the percentage of new entries whose INSDC records carried incorrect information in the organism name field. Averages include standard deviation

t1.2 Release	Type	IJSEM sync	EMBL sync	Total entries	New entries	Deleted entries	Net increase	% incorrect[a]	Average length[a]	Average ambig.[b]
t1.3 LTPs93	SSU	Dec. 2007	Dec. 2007	6,728	6,728	0	6,728	22	1,465.0 ± 51.2	0.10 ± 0.26
t1.4 LTPs95	SSU	Jun. 2008	Jun. 2008	7,006	299	21	278	45	1,446.0 ± 46.3	0.04 ± 0.11
t1.5 LTPs100	SSU	Aug. 2009	Jun. 2009	7,710	750	46	704	50	1,448.0 ± 54.2	0.03 ± 0.11
t1.6 LTPs102	SSU	Feb. 2010	Nov. 2009	8,029	363	44	319	58	1,453.6 ± 52	0.33 ± 0.12
t1.7 LTPs102	LSU	Feb. 2010	Nov. 2009	792	792	0	792	6	2,866.1 ± 177.6	0.02 ± 0.11
t1.8 LTPs104	SSU	Dec. 2010	May 2010	8,545	545	29	516	74	1,444.6 ± 62	0.27 ± 0.11
t1.9 LTPs106	SSU	May 2011	Dec. 2010	8,815	279	9	270	77	1,445.9 ± 51.1	0.03 ± 0.12
t1.10 LTPs108	SSU	Dec. 2011	Jun. 2011	9,279	490	26	464	60	1,455.4 ± 51.9	0.02 ± 0.09

t1.11 [a]Average length for the "new entries"
[b]Average percentage of ambiguities for the "total entries"

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

manuscript acceptation (recommended action for scientific journals). Successively, this rough data arrives finally to resources like LTP, which have no choice but checking it carefully to provide new informational fields with corrected information; curated information can return back to other resources of information.

Multiple Copies of the Ribosomal Operon

In 2010, a comprehensive study was conducted to evaluate the intra-genomic variability on complete type-strain genomes (Yarza et al. 2010). We observed that in very unusual exceptions, the intra-genus (94.5 %; Yarza et al. 2008) or intraspecies (98.7 %; Stackebrandt and Ebers 2006) boundaries could be exceeded within a single genome. In such cases, the selection of one or another sequence might seriously affect the interpretation of a phylogenetic inference. However, despite the fact that the vast majority of strains contain multiple copies of the *rrn* operon, only 2 % of them reveal divergences beyond 2 % (30 nucleotides) sequence identity. Thus, most likely, the selection of one or another copy should not affect the phylogenetic reconstructions. Consequently, starting from release s104 (Munoz et al. 2011), the LTP database includes all paralogues with higher divergences than 2 %. By now, it is the case of three species: *Haloarcula marismortui* ATCC 43049^T, accession number AY596297, with 5.7 % of maximum inter-operonic divergence; *Thermoanaerobacter pseudethanolicus* ATCC 33223^T, accession number CP000924, with 3.66 % of maximum inter-operonic divergence; and *Desulfitobacterium hafniense* DCB-2^T, accession number CP001336, with 4.34 % of maximum inter-operonic divergence.

Sequence Quality in LTP Datasets

It has been suggested that sequences produced for taxonomic purposes should be equal or larger than 1,450 bases with less than 0.5 % ambiguities (Stackebrandt et al. 2002). Reason is that informative content of a molecular clock is linked to the total number of its variable positions (Ludwig and Klenk 2001). Statistics derived from LTP datasets indicate that in general, sequence quality is acceptable for in-depth phylogenetic studies (~1,455 bases and 0.02 % ambiguities for LTPs108; Table 1). Figure 2 shows annual variation of gene sequence length and percentage of ambiguities. Quality increase is mainly observed in terms of ambiguities reduction, probably related to amelioration of sequencing techniques. In any case, the completion of more full genome sequences of type strains will substantially increase the sequence quality (indicated by these two parameters) in the LTP database. Researchers should be encouraged to complete 5′ ends of 16S rRNA gene sequences, as first 250 bases contain hypervariable regions V1 and V2 which play an important role in comparisons between highly related organisms (Chakravorty et al. 2007).

Curated Metadata Introduced by the LTP

In addition to regular fields provided by the ARB-SILVA databases, sequence entries include now the following LTP-specific metadata fields:

1. *fullname_ltp*: corrected species name according to LPSN (http://www.bacterio.net).
2. *rel_ltp*: name of the LTP release where a sequence entry appeared for the first time.
3. *hi_tax_ltp*: name of the family where the taxon is classified. For unclassified genera, the name of the next available higher taxon above genus (e.g., "*Acidobacteria*" for *Bryobacter aggregatus*).
4. *type_ltp*: type species receive the label "type sp." in this field.
5. *riskgroup_ltp*: risk-group classification of microorganisms according to the German BG Chemie List TRBA 466 (http://www.baua.de/en/Topics-from-A-to-Z/Biological-Agents/TRBA/TRBA-466.html).

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

6. *tax_ltp*: taxonomic classification into higher taxonomic ranks according to LPSN (http://www.bacterio.cict.fr/classifphyla.html).
7. *url_lpsn_ltp*: it contains the variable part of the URL leading to the LPSN's species file (e.g., http://www.bacterio.cict.fr/b/bryobacter.html#aggregatus).

Alignments and Phylogenetic Trees

Setting up universal alignments is a key step in order to achieve optimal and comparable phylogenetic reconstructions. It has been one of the constant motivations of Wolfgang Ludwig and co-workers who dealt with the huge task of preparing common and reliable alignment of ribosomal SSU and LSU sequences of *Bacteria*, *Archaea*, and *Eukarya* (Ludwig and Schleifer 1994). They found out that secondary structure formations such as loops and helices occurred at the same relative positions along the molecule. This helped to refine the alignments because variable stretches, with low sequence similarities, could be optimally positioned by recognizing functional homology (due to evolutionary pressure) and functional stability of helices (due to chemical stability of base pairs' bounds, G-C/A-U). A core dataset of sequences with highly curated alignments was incorporated into the SILVA system so new added sequences can be automatically aligned using this "seed alignment" as a reference (Ludwig et al. 2004; Pruesse et al. 2007). Periodically more and more manually curated sequences are added into the seed which improves its quality over time.

Although all new sequences incorporated into the LTP come from an ARB-SILVA database, they are again manually revised to further correct misplaced bases and to check highly variable regions. Before tree calculation, the complete alignment is shifted using maximum frequency filters (Table 2) that remove dubious orthologous positions caused by sequencing errors and hypervariability. Typically, LTP phylogenetic trees are calculated using the 40 % maximum frequency filter.

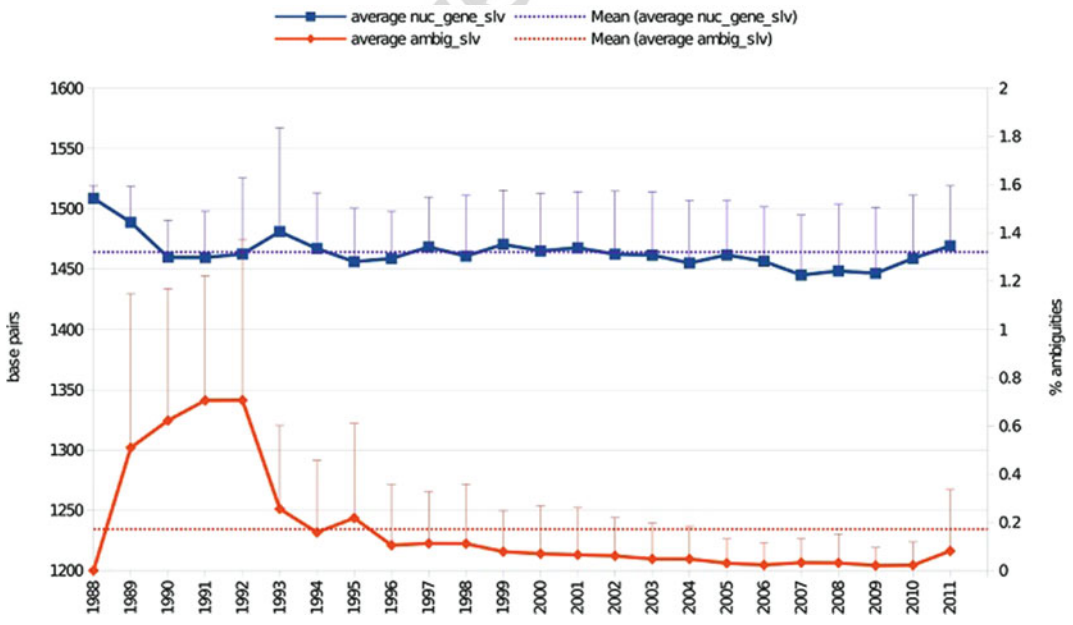

Fig. 2 Annual distribution of the 16S rRNA gene sequence length and % of ambiguities in the 9,279 type-strain sequences corresponding to LTP release s108. Gene sequence length is given by the SILVA parameter "nuc_gene_slv" which cuts off the bases at the extremes when beyond the *E.coli*'s16S rRNA gene limits. Percentage of ambiguities is given by the SILVA descriptor "ambig_slv"

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

t2.1 **Table 2** Maximum frequency filters implemented into the LTPs 108ARB database

t2.2 Filter name	Start position	Stop position	%min[a]	%max[a]	No. of positions[b]
t2.3 LTPs108_ssu_10	0	50,000	10	100	1,433
t2.4 LTPs108_ssu_20	0	50,000	20	100	1,433
t2.5 LTPs108_ssu_30	0	50,000	30	100	1,432
t2.6 LTPs108_ssu_40	0	50,000	40	100	1,390
t2.7 LTPs108_ssu_50	0	50,000	50	100	1,288

t2.8 [a]Minimum and maximum sequence identity. For tree reconstructions, only columns are taken into account if they have a positional conservation above the respective minimum values
[b]Number of homologous positions (columns) taken into account for tree reconstructions

194 The first 16S rRNA-based phylogenetic tree was calculated for the release LTPs93 (Yarza
195 et al. 2008). The sequence dataset consisted of 6,728 type-strain sequences plus 3,247 supporting
196 sequences belonging to non-type strains used to reinforce underrepresented groups and to stabilize
197 the topology. The multiple alignment of 9,975 16S rRNA gene sequences was submitted to different
198 treeing methodologies including neighbor-joining, maximum likelihood, and maximum parsimony,
199 all tested with several filters (30 %, 40 %, and 50 % maximum frequency filters) and all implemented
200 in the ARB software package (Ludwig et al. 2004). A high degree of congruence was observed
201 among them. The tree considered as optimal was a 40 %-filtered maximum likelihood reconstruction
202 calculated using the RAxML algorithm (Stamatakis 2006), with the GTRGAMMA correction, with
203 100 bootstrap replicates, in a 5-node and 20-processor parallel environment. The last de novo
204 phylogenetic reconstruction appears in the release LTPs108 and was similarly calculated; tree
205 calculation was run with a dataset of 12,166 16S rRNA gene sequences, plus 490 additional type-
206 strain SSU entries were added a posteriori using the parsimony tool implemented in the ARB
207 program.
208 The phylogenetic tree calculated using the 23S rRNA gene was particularly challenging due to
209 data shortage in many groups. In order to set up a reliable phylogeny based on 23S rRNA data, we
210 defined a core dataset made of high-quality sequences (type and non-type strains). The stringent
211 quality filtering approach ended with around 2,000 high-quality and nonredundant LSU sequences.
212 This dataset was submitted to a maximum likelihood reconstruction in combination with a 50 %
213 maximum frequency filter allowing 2,463 positions of the entire alignment. The missing partial or
214 lower-quality type-strain sequences were added to the tree using the ARB parsimony tool with the
215 option for keeping the initial topology while inserting additional data.
216 The groups shown in the trees are defined by recognizing the type members and according to the
217 taxonomic classification. The trees are carefully compared against previously reported topologies
218 and current taxonomic classifications (Yarza et al. 2010). All the additional supporting sequences
219 used to reconstruct the phylogeny are removed from the final tree by keeping its topology intact.
220 Within the ARB database, the type species are labeled with a distinct color for easy recognition and
221 tree handling.

222 Files Provided by the LTP

223 As a taxonomic tool, the LTP must be understood as a collection of reference materials, all publicly
224 available at the project's Web page (http://www.arb-silva.de/projects/living-tree), including:

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

1. Release documentation: (I) readme file with a release description and (II) PDF document describing the metadata fields introduced by the LTP
2. Tables: (I) new entries with outdated submission names and (II) list of changes in the dataset: added/deleted/modified entries
3. Export filter: ARB-export filter (*.eft* format) to extract data from LTP-ARB databases
4. Databases: (I) complete ARB databases including sequences, alignments, metadata, filters, and trees and (II) datasets in CSV format including LTP metadata
5. Alignments: (I) gapped exports in multi-FASTA format and (II) compressed exports in multi-FASTA format
6. Phylogenetic trees: (I) collapsed overviews in PDF format showing the distinct phyla, (II) full SSU (more than 80 pages long) and LSU trees in PDF format, and (III) full trees in NEWICK format, including group names and branch lengths

Side Research

Sequencing the Orphan Species Initiative (SOS)

The understanding that around 6 % of all classified species were missing from the ribosomal SSU sequence catalogues motivated us to start the "Sequencing the Orphan Species" (SOS) initiative (Yarza et al. 2013). During 3 years of work, the LTP team coordinated a network of 12 partner researchers and culture collections (ATCC, BZF, CECT, CIP, CCUG, DSMZ, JCM, ICMP, BCCM/LMG, MMG, NBRC, NCCB) in order to improve this situation by (re)sequencing the 16S rRNA gene of the "orphan" species. As a result, 351 type strains appear represented now by a good-quality SSU gene sequence in the databases. They comprise representatives of 14 bacterial and archaeal phyla, 76 type species, and 78 pathogenic species. However, 201 type strains could not be accessed as cultivable strains were not available at recognized culture collections. They represent 10 phyla and 17 type species.

Taxonomic Boundaries

In order to understand how the higher taxonomic categories could be circumscribed by means of a sequence identity threshold, we performed a statistical procedure to get the lowest similarity found within the members of a certain taxon (Yarza et al. 2008, 2010). By taking into account all the taxa at a particular taxonomic rank, we obtained general lower cutoff values of sequence identity for genus, family, and phylum based on 16S rRNA and 23S rRNA. In general, minimum 16S rRNA gene sequence identities of 94.9 % $\pm$ 0.4, 87.5 % $\pm$ 1.3, and 78.4 % $\pm$ 2.0 lead to the circumscription of a new genus, family, and phylum, respectively. For 23S rRNA genes, these values are slightly different: 93.2 % $\pm$ 1.3 (genus), 87.7 % $\pm$ 2.5 (family), and 75.3 % (phylum). As shown by the low errors, historically used criteria for genera, families, and phyla are quite homogeneous and do not lead to unambiguous circumscriptions. These cutoffs should be used with caution and always as a complementary approach. They are especially recommended for prospective studies in clone libraries and as additional support for the circumscription of new taxa or emendation of existing ones.

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

Summary

SSU and LSU databases made by the All-Species Living Tree Project (LTP; http://www.arb-silva.de/projects/living-tree) provide high-quality nearly full-length sequences of the type strains of all *Archaea* and *Bacteria* with validly published names. Setting up a type-strain sequences database included the sieving of the public DNA databases whose sequence entries often appeared outdated or mistaken at their taxonomic metadata. It involved the initial manual cross-check of nearly 14,000 SSU and 6,000 LSU sequence entries against the catalogue of distinct species with validly published names retrieved from LPSN. Databases are complemented with manually curated metadata, manually curated alignments, and state-of-the-art phylogenetic reconstructions (in contrast to other similar resources like the EzTaxon (Santamaria et al. 2012)). The LTP team wants to remark that the aim of the project is not to reconstruct the currently described species genealogy with total fidelity but to provide a curated taxonomic tool for the scientific community. Our small but very representative SSU and LSU datasets may be used as a reference for identification and classification purposes in several fields of application, for example, facilitating the collection of sequences for the reconstruction of taxa genealogies (Cousin et al. 2012), enabling fast and reliable taxonomic affiliations in rRNA surveys (Santamaria et al. 2012), or serving as reference datasets for testing bioinformatic procedures (Mizrahi-Man et al. 2013).

Cross-References

- ARB
- Archaea, Definition
- Bacteria, Definition, Features and Classification Schemes
- Culture Collections in the Study of Microbial Diversity, Importance
- Phylogenetics, Overview
- SILVA Databases

References

Amann R, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59:143–69.

Chakravorty S, Helb D, Burday M, et al. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007;69:330–9.

Coenye T, Gevers D, Van de Peer Y, et al. Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev. 2005;29:147–67.

Cousin S, Gulat-Okalla ML, Motreff L, et al. Lactobacillus gigeriorum sp. nov., isolated from chicken crop. Int J Syst Evol Microbiol. 2012;62:330–4.

Fox GE, Pechman KR, Woese CR. Comparative cataloguing of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. Int J Bacteriol. 1977;27:44–57.

Garrity GM. Bergey's manual of systematic bacteriology. 2nd ed. New York: Springer; 2001.

Lapage SP, Sneath PHA, Lessel EF, et al. International code of nomenclature of bacteria (1990 revision). Washington, DC: American Society for Microbiology; 1992. p. 295.

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

Ludwig W, Klenk HP. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW, Garrity GM, editors. Bergey's manual of systematic bacteriology. 2nd ed. New York: Springer; 2001. p. 49–65.

Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev. 1994;15:155–73.

Ludwig W, Strunk O, Westram R, et al. ARB: a software environment for sequence data. Nucleic Acids Res. 2004;32:1363–71.

Mizrahi-Man O, Davenport ER, Gilad Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. PLoS One. 2013;8:e53608.

Munoz R, Yarza P, Ludwig W, et al. Release LTPs104 of the all-species living tree. Syst Appl Microbiol. 2011;34:169–70.

Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35:7188–96.

Santamaria M, Fosso B, Consiglio A, et al. Reference databases for taxonomic assignment in metagenomics. Brief Bioinform. 2012;13:682–95.

Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today. 2006;33:152–5.

Stackebrandt E, Frederiksen W, Garrity GM, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol. 2002;52:1043–7.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.

Yarza P, Richter M, Peplies J, et al. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol. 2008;31:241–50.

Yarza P, Ludwig W, Euzéby J, et al. Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. Syst Appl Microbiol. 2010;33:291–9.

Yarza P, Spröer C, Swiderski J, et al. Sequencing Orphan Species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst Appl Microbiol. 2013;36:69.

Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_740-8

Author Queries

Query Refs.	Details Required
Q1	Please check if affiliation details are okay.
Q2	Entry titles "ARB; Archaea, Definition; Bacteria, Definition, Features and Classification Schemes" do not match with TOC. Please check.
Q3	Please check if inserted volume ID and page range for Yarza et al. (2013) is okay.