Available online at www.sciencedirect.com

ScienceDirect

Systematic and Applied Microbiology 31 (2008) 241–250

SYSTEMATIC AND APPLIED MICROBIOLOGY

www.elsevier.de/syapm

The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains

Pablo Yarza[a], Michael Richter[a], Jörg Peplies[b], Jean Euzeby[c], Rudolf Amann[d], Karl-Heinz Schleifer[e], Wolfgang Ludwig[e,**], Frank Oliver Glöckner[d,f,**], Ramon Rosselló-Móra[a,*]

[a]*Marine Microbiology Group, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), C/ Miquel Marqués 21, E-07190 Esporles, Illes Balears, Mallorca, Spain*
[b]*Ribocon GmbH, D-28359 Bremen, Germany*
[c]*Société de Bactériologie Systématique et Vétérinaire (SBSV) & École Nationale Vétérinaire de Toulouse (ENVT), F-31076 Toulouse Cedex 03, France*
[d]*Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany*
[e]*Lehrstuhl für Mikrobiologie, Technische Universität München, D-85350 Freising, Germany*
[f]*Jacobs University Bremen, D-28759 Bremen, Germany*

Abstract

The signing authors together with the journal Systematic and Applied Microbiology (SAM) have started an ambitious project that has been conceived to provide a useful tool especially for the scientific microbial taxonomist community. The aim of what we have called "The All-Species Living Tree" is to reconstruct a single 16S rRNA tree harboring all sequenced type strains of the hitherto classified species of *Archaea* and *Bacteria*. This tree is to be regularly updated by adding the species with validly published names that appear monthly in the Validation and Notification lists of the International Journal of Systematic and Evolutionary Microbiology. For this purpose, the SAM executive editors, together with the responsible teams of the ARB, SILVA, and LPSN projects (www.arb-home.de, www.arb-silva.de, and www.bacterio.cict.fr, respectively), have prepared a 16S rRNA database containing over 6700 sequences, each of which represents a single type strain of a classified species up to 31 December 2007. The selection of sequences had to be undertaken manually due to a high error rate in the names and information fields provided for the publicly deposited entries. In addition, from among the often occurring multiple entries for a single type strain, the best-quality sequence was selected for the project. The living tree database that SAM now provides contains corrected entries and the best-quality sequences with a manually checked alignment. The tree reconstruction has been performed by using the maximum likelihood algorithm RAxML. The tree provided in the first release is a result of the calculation of a single dataset containing 9975 single entries, 6728 corresponding to type strain gene sequences, as well as 3247 additional high-fquality sequences to give robustness to the reconstruction. Trees are dynamic structures that change on the basis of the quality and availability of the data used for their calculation.

**Also for correspondence. Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany.
Tel.: +49 421 2028 970; fax: +49 421 2028 580.

*Corresponding authors. Tel: +34 971 611 826; fax: +34 971 611 761 (R. Rosselló-Móra), Tel.: +49 8161 71 5451; fax: +49 8161 71 5475 (W. Ludwig).

E-mail addresses: ludwig@mikro.biologie.tu-muenchen.de (W. Ludwig), fog@mpi-bremen.de (F.O. Glöckner), rossello-mora@uib.es (R. Rosselló-Móra).

doi:10.1016/j.syapm.2008.07.001

Therefore, the addition of new type strain sequences in further subsequent releases may help to resolve certain branching orders that appear ambiguous in this first release.

On the web sites: www.elsevier.de/syapm and www.arb-silva.de/living-tree, the All-Species Living Tree team will release a regularly updated database compatible with the ARB software environment containing the whole 16S rRNA dataset used to reconstruct "The All-Species Living Tree". As a result, the latest reconstructed phylogeny will be provided. In addition to the ARB file, a readable multi-FASTA universal sequence editor file with the complete alignment will be provided for those not using ARB. There is also a complete set of supplementary tables and figures illustrating the selection procedure and its outcome. It is expected that the All-Species Living Tree will help to improve future classification efforts by simplifying the selection of the correct type strain sequences.

For queries, information updates, remarks on the dataset or tree reconstructions shown, a contact email address has been created (living-tree@arb-silva.de). This provides an entry point for anyone from the scientific community to provide additional input for the construction and improvement of the first tree compiling all sequenced type strains of all prokaryotic species for which names had been validly published.

Keywords: 16S rRNA; SSU; Phylogeny; Species; Living tree; ARB; Taxonomy; Classification; Taxa boundaries

The need for a curated all-species tree

Thirty years ago, the systematics of prokaryotes experienced an important breakthrough when attempts were made to establish the first genealogical relationships by using comparative cataloguing of the primary sequence of the small subunit (SSU) of the ribosome [8]. At that time, systematicists were already aware that the new tool for inferring genealogies would have an important impact on the way the taxonomy of prokaryotes developed [9]. However, the establishment of a phylogenetic backbone for the classification of prokaryotes has required the important task of validation for the tree topologies in comparison with other molecular clocks [19]. Nevertheless, nowadays, it is clear that the 16S rRNA gene sequence analysis applied to bacterial systematics is of paramount relevance. Nearly all descriptions of taxa are accompanied by relevant sequence information and reconstruction of their relationships based on the sequence of the SSU of the ribosome. Furthermore, it has been recommended that the inclusion of a high-quality sequence should be mandatory in the future [30]. Actually, the current overview of the classification of prokaryotes is mainly based on genealogical affiliations [11], and the circumscription of any new taxon with a higher hierarchy than species (i.e. genus and above categories) is based on genealogical relationships. The single category for which SSU sequence divergences cannot provide a sharp resolution is species [26]. In this respect, identical or nearly identical SSU sequences cannot guarantee that two organisms belong to the same species following the criteria traditionally used to define and circumscribe this category [10]. Despite the fuzziness of the resolution power of the SSU at the species level, it has been observed that, in general, two organisms with sequence divergence above a 3% nucleotide identity may not belong to the same species [1,31], and, for the same reason, lower divergences may be tested by DNA–DNA hybridization analysis. Currently, it is recommended that the hybridization is to be done when identity values are below 98.7–99% [29]. Nevertheless, SSU analysis is important for inferring monophyly [30], and this is one of the most important premises for circumscribing a prokaryotic species.

One of the main controversial issues concerning the validity of SSU gene analysis is whether this single gene really represents the genealogy of the organism that harbors it. Phenomena such as genetic crossover of ribosomal genes [27] or horizontal gene transfer (HGT, [6]) have been referred to as being responsible for blurring the validity of SSU to represent organismal genealogy. Today, whole genome comparisons provide unprecedented insights. On the one hand, and in the light of the current knowledge of the genetic content of prokaryotes, a large HGT occurrence has been hypothesized [14], whereas, on the other hand, there are severe criticisms of how data are interpreted [15]. In any case, it has been hypothesized that an organism's genome may contain a certain set of genes which would be largely excluded from HGT, and would be responsible for what an organism is and thus for its identification [16]. In general, large phylogenetic studies with different sets of housekeeping genes based on comparative genomics provide strong support for the genealogies based on SSU analysis [4,28]. Altogether, the comparisons indicate that, for classification purposes, SSU tree reconstructions may be the most parsimonious and accurate way to establish genealogical relationships.

Despite the criticisms, comparative sequence analysis of the SSU rRNA has been established as the gold standard for reconstructing phylogenetic relationships among prokaryotes for classification purposes [18]. As a consequence, the number of SSU sequences deposited in public databases has increased exponentially by about three orders of magnitude in approximately 15 years

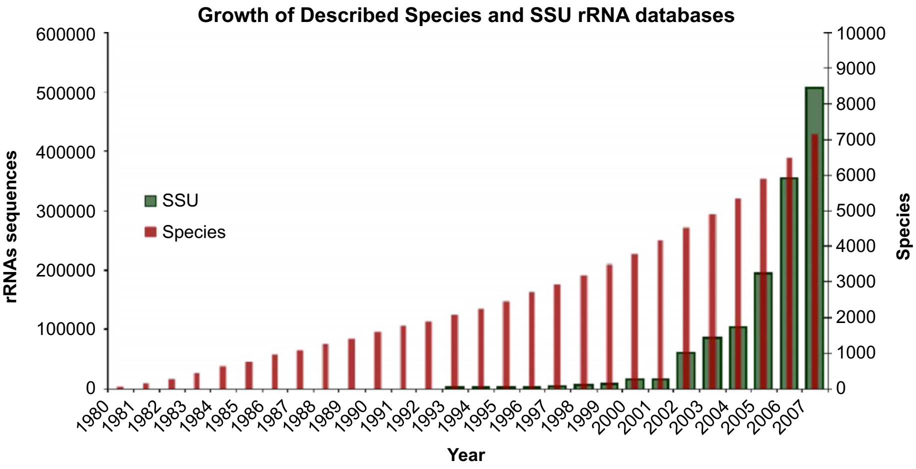

Fig. 1. Increase in the number of validated species from 1980 to 2007, and the SSU sequence submissions to public databases until SILVA release 93 (updated to 566,047).

(Fig. 1), as shown on the SILVA website (www.arb-silva.de). Most of the sequences deposited correspond to uncultured organisms, since the SSU has also become the tool for cultivation-independent analysis of the diversity of complex microbial communities [2,22]. Consequently, only the minority of sequences corresponds to cultured prokaryotes (Fig. 1). This enormous amount of information undoubtedly represents a useful tool for understanding the extent of microbial diversity. However, in order to achieve optimal and comparable reconstructions, it is necessary that all phylogenies are reconstructed following a similar approach. For this purpose, a universal SSU alignment has been devised taking into account not only the primary gene sequence, but also the secondary structure based on nucleotide pairing that represents the main SSU functional helices [20,24]. This alignment is implemented in the SILVA databases and is compatible with the ARB program package available online at www.arb-silva.de and www.arb-home.de, respectively. The ARB-SILVA team maintains the enormous dataset of publicly available SSU genes [24], and the SILVA website offers comprehensive databases of the aligned SSU and large subunit of the ribosome genes to the scientific community.

The novelty of a taxon is confirmed by discarding its assignment to a pre-existing species. The current list of species with validated names can be retrieved from the List of Prokaryotic names with Standing in Nomenclature (LPSN) public website www.bacterio.cict.fr. The culture collection numbers of the type strains of each species can also be identified on this website. It is a general approach to identify the uniqueness of a new species by checking that no previous publicly available sequence from an existing type strain exists. Due to this reason, most of the descriptions of new species and genera are generally accompanied by the SSU gene sequence of their type strains. One of the most important steps in order to recognize the uniqueness of new taxa is the identification of the available type strain sequences in the public databases. Unfortunately, this step is currently hampered by the inaccurate information submitted to the International Nucleotide Sequence Database Collaboration (INSDC; www.insdc.org), which comprises EMBL, GenBank and DDBJ. Common mistakes are related to incorrect species names, misassigned accession numbers or wrong biological resource collection identifiers. Furthermore, the respective sequence information deposited can be of low quality, thus rendering phylogenetic reconstructions difficult or even impossible.

In order to produce a useful tool for the scientific community, so that a species classification can be retrieved in the form of a phylogenetic tree, we have started the All-Species Living Tree Project. This is an initiative between the journal Systematic and Applied Microbiology and the group of scientists authoring this work. Our intention is to (i) provide a curated SSU database of all type strains for which sequences are available; (ii) maintain an optimized and universally usable alignment; and (iii) reconstruct a tree harboring reliable genealogies. It is intended that the databases and tree will undergo regular updates to include all forthcoming validly described new taxa. To our knowledge, this is the first attempt to produce a single tree harboring all validly described species of prokaryotes for which an adequate sequence has been deposited in the public databases.

Sequence selection

In order to proceed with the selection of sequences to reconstruct the all-species tree, the SILVA database (www.arb-silva.de) was supplemented with a manually extracted list of all validly published names provided by the LPSN (www.bacterio.cict.fr). Fig. 1 shows the differences in the growth tendency of both databases. From the 8264 validly published names until 31 December 2007, about 7367 corresponded to distinct species with standing in nomenclature. This set of species was the starting point for a detailed cross-check with already existing information on type strains in the SILVA database. The use of the 154 ‘‘candidatus’’ species (i.e. uncultured, but ecologically conspicuous organisms accepted as putative taxa; [21]) was avoided, since several distinct sequences could be found for many of them. Consequently, it was decided to concentrate on the validly published names for which a type strain was designated. Later heterotypic synonyms of existing species were not included, especially for this first release, in order to avoid nomenclatural confusion. In addition, about 226 species [7] were included for which the names could not appear in the validation lists due to a lack of accordance with the Bacteriological Code (www.bacterio.cict.fr). This list has now been reduced to 69 [12]. Among the *Cyanobacteria*, only the six species published under the Bacteriological Code rules [17] were considered.

The first step in selecting the sequences was an automated search for criteria fulfilling the project requirements. From the 109,626,755 sequences present in the EMBL nucleotide sequence database, 1,200,423 corresponded to potential SSU sequence candidates made publicly available in EMBL and SILVA release 93. Less than half of them (566,047) could be chosen as accomplishing the minimum standards required to be harbored in the SSUParc database. From among these more than a half a million sequences, only 224,967 were recognized as nearly full-length sequences (>1200 nt) and of an alignment quality appropriate for the reference SILVA database SSURef. To reduce further the dataset, all sequences that were not labeled as cultivated or type strains were removed, thereby leaving 13,816 candidates for manual cross-checking. The information concerning cultivated strains and type strains in SILVA has been mainly provided by straininfo.net [5]. Detailed information is available at www.arb-silva.de/background/.

Once the sequences with a putative assignment to a type strain of an existing species were collected they were compared to the list of validly published species. One by one, each sequence was assigned to a species by proving the strain collection number assignments. It was surprising that the data uploaded to the EMBL were very often incomplete or wrong. About 1500 sequences had wrong names (supplementary Table 1), lack of strain information in the EMBL entries, or both. This information has already been corrected for the SILVA all-species tree, thus, in the database provided, names and incorrect entries had been updated. After having checked the whole sequence list, there was still a large set of species for which a sequence could not be assigned. At this point the process was inverted by searching in the EMBL sequence databases for those sequences matching one of the synonym strain culture collection numbers. This second process gathered 1713 additional sequences, 209 of them not recognized in the first sift due to the lack of a type strain label.

The curation study finished with four sets of species. The first set consisted of 362 ‘‘orphan’’ species with no sequences (supplementary Table 2), since most of these species had never been sequenced because they were described before easy SSU sequence analysis was available. A second set of 276 species comprised those for which a sequence existed, but they did not meet the quality standards for our project. Among these, 177 were directly rejected by the initial quality checks of the SILVA project (supplementary Table 3), 45 were listed in the SILVA SSUParc database, but were too short to be included in the SSURef database (supplementary Table 4), and, finally, a set of 54 sequences listed in the SSURef database were manually removed due to insufficient quality (supplementary Table 5). In Fig. 2, the final distribution of species is shown with regard to their sequence quality and usability in the living tree.

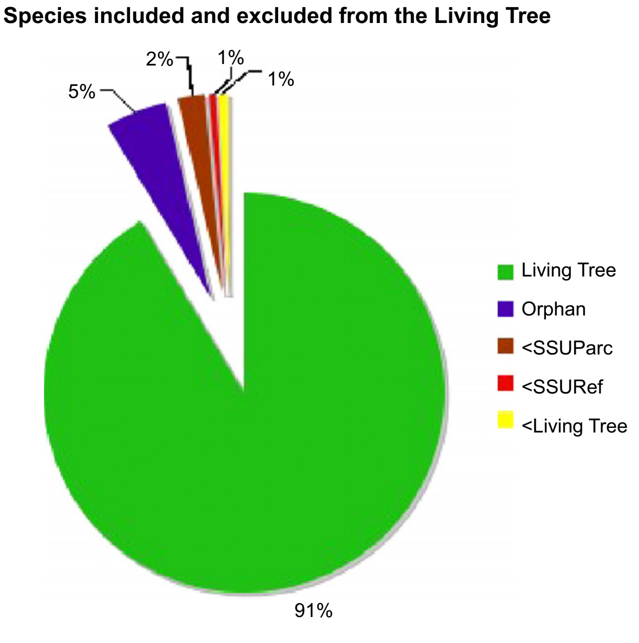

Fig. 2. Percentage distribution of (i) species with an adequate sequence for inclusion in the LTP_ARB tree (green); (ii) species (orphan) for which no sequence entry was found (blue); (iii) species with a sequence quality below the thresholds of the SSUParc database (brown); (iv) species with a moderate quality, but not adequate enough to be included in the SSURef database (red); and, (v) species with an adequate quality to be included in the SSURef database, but discarded due to alignment problems that made the identity dubious (yellow).

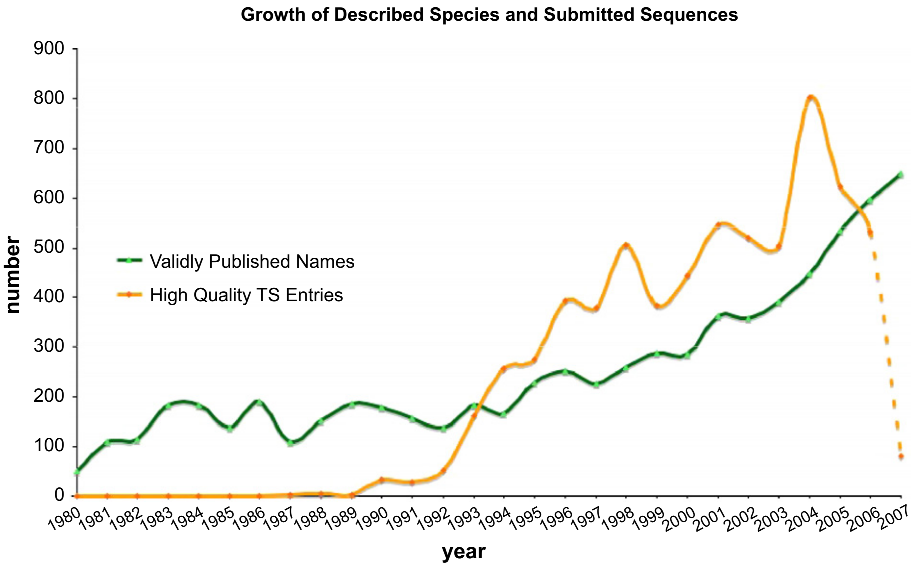

Fig. 3. Number of type strain sequences (orange) and validly published names (green) per year entering the public databases from 1980–2007.

This sieving process selected a sequence database covering 6782 species for which the type strain had an entry in the EMBL database. The final set of type strain sequences comprised 9682 entries. The increase of type strain sequences in the database as a cumulative or yearly absolute number is summarized in Fig. 3. Fig. 3 shows an approximate constant rate of descriptions from the early 1980s to the late 1990s. Subsequently, about 10 years ago, an arithmetic increase of new descriptions started. SSU sequence data of type strains underwent a period of synchronization, and now any new species description is accompanied by its SSU gene data. It is expected that in the near future, within 6 months to 1 year, the rate of type strain sequences will be the same as the new species descriptions.

A total of 4982 species were recorded with a single sequence entry, whereas 1800 species had more than one entry. These together gave 4700 sequence entries and, of these, 45 species contained one or more paralogs. Many had identical EMBL accession numbers as they corresponded to whole genome sequences of microorganisms with multiple rrn operons [13]. The remaining 1755 species were represented by multiple independent submissions. Our aim was to reduce the dataset to one sequence for each single type strain of the validated species. For this, the rationale for removing duplicates was to take the best-quality sequence from among the different entries. The criteria used were the following: (i) for a couple of sequences with the same quality, priority was given to the one submitted first; (ii) only one of the several operon sequences with 100% identity belonging to completely sequenced genomes (generally that with the first entry) was chosen; (iii) between duplicates with a distinct length and a SILVA quality mark, the longest sequence was chosen, unless the quality (ambiguities, homopolymers or sequence anomaly) was clearly worse than the shorter sequence. However, in all cases, a manual check of the alignment quality was also included as a final determinant in the selection.

The final sequence dataset used to construct the first all-species tree contained 6728 entries representing modest- to good-quality sequences of the 7367 distinct species classified up to the end of December 2007 (supplementary Table 6). This final set is equivalent to about 91% of the complete catalogue of classified prokaryotic species (Fig. 2).

Finally, a selection of 3247 additional sequences not belonging to any type strain was taken to complement the whole dataset, which gave a final number of 9975 bacterial and archaeal SSU sequences. The addition of the non-type strain sequences increased the presence of groups that were underrepresented with respect to the number of sequences, resulting in unstable branching topology (e.g. *Cyanobacteria*, *Lentisphaerae*, *Deferribacteres*). In general, the preliminary analyses of the tree topology obtained by just using type strains was, for a few groups (e.g. *Cyanobacteria*, *Thermomicrobia*, *Chrysiogenetes, Fusobacteria*), incongruent with the current knowledge of the tree branching order. This additional dataset is included in the LTP_ARB database, but has been removed from the tree to avoid confusion.

Alignment improvements

Sequences had been automatically aligned by SINA, as implemented by the SILVA database project [24].

Table 1. Statistics for the LTP_ARB database

1A

	Min		Max		Mean	SD
Length (bases)	1229	*Methanohalophilus portucalensis*	2210	*Pyrobaculum aerophilum*[a]	1465.38	50.65
Number of ambiguities	0	*Shewanella putrefaciens*	30	*Sebaldella termitidis*	1.45	3.82
Number of homopolymers	0	*Leuconostoc carnosum*	26	*Pyrobaculum aerophilum*[a]	4.05	2.19

1B

	No. of sequences
No ambiguities	4674
No homopolymers	27
No ambiguities and homopolymers	17
No ambiguities and homopolymers and full length[b]	1

[a]Contains a long insertion.

[b]*Leuconostoc mesenteroides* subsp. *mesenteroides*.

Table 2. Conservational filters of maximum frequency implemented in the LTP_ARB database

	Start position	Stop position	% Min[a]	% Max[a]	No. of positions[b]
LTP_ssu_30	0	50,000	30	100	1439
LTP_ssu_40	0	50,000	40	100	1400
LTP_ssu_50	0	50,000	50	100	1296

[a]Minimum and maximum identity. For tree reconstructions only columns are taken into account if they have a positional conservation above the respective minimum values.

[b]No of homologous positions (columns) taken into account for tree reconstructions.

Briefly, the system searches for the closest relatives in a set of 51,601 manually curated SSU sequences (Seed). Up to 40 related sequences are then used as references for the alignment of the sequence under investigation. Although the process is highly accurate, some of the bases usually escape optimal placement according to biological criteria. The complete dataset of 9975 sequences (type strains and non-type strains) was manually checked in order to improve inaccurately placed bases. For this, the secondary structure of the SSU was taken into account. The final alignment can be retrieved as an ARB database, as well as supplementary material in an aligned multi-FASTA file, and from www.arb-silva.de/living-tree.

The whole database contained sequences that had a range of quality. Among the type strain sequences, a total of 497 entries were detected that could be considered as full length (sequences larger than 1524 nucleotides). As shown in Table 1, the maximum length of a sequence corresponded to the 2210 nucleotide entry of *Pyrobaculum aerophilum*, which contains a large insertion of 712 nucleotides starting at *Escherichia coli* alignment position 373 [3]. The shortest sequence in the database corresponded to *Methanohalophilus portucalensis* with an entry of 1229 nucleotides. The average sequence length in the database was 1465, and a maximum of 30 ambiguities and/or 26 homopolymers in a single sequence was allowed.

Tree reconstruction

To exclude positions where positional orthology could not be guaranteed in the alignment, three filter sets were applied to remove positions where the highest occurring base was conserved at less than 30%, 40% and 50% (Table 2). This was designed to increase the signal to noise ratio and therefore improve the stability of the tree [23]. In this respect, by increasing the percentage conservation threshold, the number of homologous positions taken into account for reconstruction decreased, although prominence was given to conserved positions.

The complete dataset of 9975 sequences was submitted to different treeing approaches: neighbor-joining (using the Jukes–Cantor correction, as implemented in the ARB program package), maximum likelihood (using RAxML version 7.0 with the GTRGAMMA model; [32]), and ARB_PARSIMONY, as implemented in the ARB program. Each of the algorithms was tested by using the dataset treated with 30%, 40% and 50%

conservational filters. Furthermore, 100 bootstrap replicates were carried out for comparison using RAxML-MPI (Message Passing Interface) on a 5-node, 20-processor parallel environment (GTRGAMMA model). Congruence was checked between trees and with the previously established tree topologies for prokaryotes. A tree constructed using 40% positional homology filtering was regarded as optimal. Bootstrap support was generally high for nodes that could be unambiguously resolved by the different tree reconstruction algorithms and filters applied. Since no further information could be deduced from the bootstrap values, they are not shown in the final maximum likelihood tree available in the ARB living tree database.

Some features of the tree

The tree, based on the data gathered until 31 December 2007, contains 6728 type strain sequences. In this release, later heterotypic synonyms of existing species were not included in order to avoid confusion. However, it was constructed with the support of 3247 additional sequences that were removed after tree reconstruction. Among the type strain sequences, 1351 corresponded to type species of genera. These sequences have been highlighted in the LTP_ARB database, and are marked with a different color (ARB-color 10) than the non-type species sequences (ARB-color 12). Altogether, a total of 174 type species of genera are missing from the dataset, 112 of which were never sequenced, and 62 that did not accomplish the minimum standard set for the project (supplementary Table 7). It would be desirable to obtain a full-length sequence for these listed species in order to cover fully the sequence diversity of the hitherto described genera.

To our knowledge, this is the first reconstruction of an all-species tree based on carefully selected type strain SSU rRNA sequences of *Bacteria* and *Archaea*. The product provided has two major added values: (i) a curated dataset made from sequences representing type strains of hitherto described species, and (ii) the first maximum likelihood reconstruction based on a large set of sequences (9975 entries) representing the whole diversity of the cultured and validly described prokaryotic species.

The significance of a curated dataset: It is expected that this curated database of the all-species tree project will facilitate the collection of sequences for the reconstruction of taxa genealogies. Nevertheless, despite the large set of sequences used in the project, it is highly probable that we have failed to select some of them. Consequently, any feedback from the scientific community regarding the improvement of the sequence selection would be welcomed and greatly appreciated. All requests should be referred to the project email address living-tree@arb-silva.de.

The significance of the first maximum likelihood tree: As stated above, we believe that this is the first rRNA genealogy created from such a large dataset, based on the maximum likelihood algorithm. The first attempts to reconstruct the all-species genealogy failed for several important groups due to the unbalanced numbers of the representative taxa. Whereas some branches contained large numbers of classified taxa (e.g. *Proteobacteria*, *Firmicutes*), others appeared underrepresented (e.g. *Chlorobi*, *Thermodesulfobiaceae*). Such differences in representative sequences for each branch may promote unstable topology [18]. For this reason, the dataset was enlarged with an additional 3247 sequences to provide a better balanced representation of phylogenetic branches. As a result, some of the incongruences in the tree topology were resolved. Nevertheless, with currently available computing power it is not possible to reconstruct a topology from a very large dataset to test further the influence of undersampling for some branches or phyla.

Most probably the tree topology shown in the LTP_ARB cannot reflect the correct reconstruction for all the represented taxa. Trees are dynamic structures that change on the basis of the quality and availability of the data used for their calculation. Therefore, the addition of new type strain sequences in further subsequent releases may help to resolve branching orders that appear ambiguous in this first release. However, the manual analysis of the tree topology indicated that, in most of the cases, the branching order was coherent with the hitherto accepted topologies based on data subsets. It is important to note here that for major new classification efforts the branch stability of the tree to be published needs to be reanalyzed based on multiple reconstructions from different datasets and using various algorithms [18].

Coherent and incoherent taxa: Taxa that may be susceptible to reclassification can be easily recognized simply by scrolling through the tree, whereas other taxa can be recognized as being coherent and thus adequately classified (e.g. *Geobacter*, *Desulfurella*, *Helicobacter*). Species susceptible to reclassification can be recognized quickly due to the fact that they do not coherently affiliate with the rest of the members of their genus (e.g. *Aeromonas sharmana*, *Pseudomonas mephitica*, *Pseudomonas cissicola*, *Pseudomonas boreopolis*), or they clearly affiliate with a different but coherent genus (e.g. *Weeksella virosa* affiliates within the genus *Bacteroides*; *Lawsonia intracellularis* affiliates within the genus *Desulfovibrio*, *Xylanibacter oryzae* affiliates within the genus *Prevotella*; *Streptomyces longisporoflavus* affiliates within the genus *Brevundimonas* of the *Betaproteobacteria*; *Streptomyces gardneri* affiliates with the genus *Nocardia*). Some taxa appear paraphyletic or polyphyletic (e.g. the genera *Eubacterium*, *Bacillus*, *Pseudomonas*, *Desulfotomaculum*), and thus a revision of their taxonomic status is suggested. In any case, and as stated above, the topology provided here needs to be further tested by complementary phylogenetic markers with higher resolution at the family, genus and species level in order to improve branching order stability.

New classifications and further living tree releases: In this first release of the project, we have provided the tree topology for all classified species up to 31 December 2007. However, during the curation of the dataset and reconstruction of the trees, several new species appeared in the literature. These and other new species may contribute to the local tree topology stability once they are added to the dataset. The aim is to provide updates for the datasets and trees at least twice a year. The new releases will not only contain the new classifications, but also all recommendations made by the scientific community that have been directly communicated via the feedback email address: living-tree@arb-silva.de.

Calculating taxa boundaries: Statistical analysis was undertaken in order to understand how the categories of genus, family and phylum could be circumscribed in terms of SSU similarities. For this purpose, the 451 genera harboring three or more species (Fig. 4), 28 families harboring three or more genera and 10 phyla harboring three or more families (Table 3) were studied. From the results, it was shown that a genus contains species that have an average identity to the corresponding type species of 96.4%, whereas the maximum identity between species within a genus is on average 98%. However, it has to be taken into account that there are genera (e.g. *Brucella*) that may contain species with 100% sequence identity. In general, the minimum identity value that guarantees the circumscription of a single genus is 94.9% ± 0.4 to the type species. In principle, lower values may lead to a new genus circumscription. In contrast to the genus calculations that were undertaken by using the whole database, the family boundaries were calculated by manually selecting 28 examples of clear-cut taxa. In this respect, the family boundaries may be set by a minimum identity of 87.5% ± 1.3 to the type species of the genus giving the name to the category. Values below this may lead to a circumscription of a new family. Finally, the results based on the 10 selected phyla indicated that 78.4% ± 2.0 may be a good threshold to recognize the members of a single phylum.

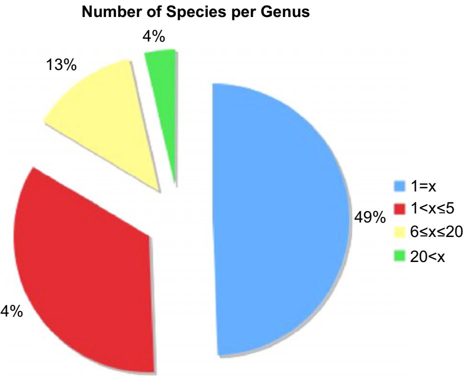

Fig. 4. Distribution of the number of species representing the pool of genera that could be identified within the framework of the all-species living tree. The LTP_ARB database contains 6728 species grouped into 1463 genera. A total of 710 genera harbored only one species, 492 contained between two and five species, 181 contained between five and 20 species, and only 53 genera harbored more than 20 species. The genus *Streptomyces* comprising 488 species is the largest genus in the database.

Table 3. Boundaries at different taxonomic levels

	Genus	Family	Phylum
Number of taxa	451	28	10
Number of species	4559	202	195
Maximum identity	98% ± 0.2	92.5% ± 1.2	84.7% ± 1.9
Average identity	96.4% ± 0.2	90.1% ± 1.1	81.7% ± 1.8
Minimum identity	94.9% ± 0.4	87.5% ± 1.3	78.4% ± 2.0

The table contains identity values calculated as the average observed within each individual group. 95% Confidence intervals are also displayed. For the genus calculations, about 63 species were not included as they were considered to be wrongly classified. Results were generated using only those taxa considered taxonomically well-defined. *Planctomycetes*, *Spirochaetes*, *Nitrospirae* and *Cyanobacteria* could not be included in the calculations of phyla boundaries due to the lack of a sequence for the type organism (i.e. type species in a genus, or the type species giving the name of the family, and/or the phylum).

It also has to be taken into account that the taxonomic schema, and especially the basal categories (family, genus and species), have been constructed by empirical observations of what may or may not belong to a given category, and that it is a product of belief that the whole microbial diversity can be explained by using universal criteria [25]. The species circumscription and the resolution power of the SSU for improving this category definition has already been largely discussed (e.g. [18]). In contrast, higher categories, especially genus and family, had been generally created after using exclusion criteria based on differences in phenotypic and genetic traits. This is different for the phylum level which is solely based on comparative sequence analysis of the SSU gene. A new phylum is defined by the segregation of a new branch in a tree reconstruction. The data shown in Table 3 are no more than the result of averaging the empirical decisions of the responsible scientists creating categories. However, and as can be deduced by the low variation in the averages calculated, the criteria generally used are homogeneous and do not lead to inconsistent circumscriptions. Although our values cannot be taken as tenets, they may help the further discrimination of taxa, and thus advance the construction of taxonomic schemes.

Important remarks concerning the project

First of all the all-species living tree team wants to state that this is not an attempt to reconstruct the currently described species genealogy with total fidelity, but to provide a curated taxonomic tool for the scientific community. The database presented contains all species with validly published names for which a sequence entry with adequate quality could be found. Poor or short sequences were not taken into account because of the reconstruction biases that can occur due to the phylogenetic noise they may generate. In addition, we have only considered species with a clear putative status in the taxonomic schema. For this first release, we have not included all such species considered to be later synonyms of already existing taxa, despite the existence of a designated type strain (supplementary Table 8). In this respect, we did not consider heterotypic synonym species as essential for the first release of the all-species tree, due to the fact that they may lead to confusion. Nevertheless, for completeness, they will most probably be included in future releases. Finally, we believe that although the project creates a curated database this may not prevent errors, and, therefore, we make a plea for understanding, as well as constructive feedback for improving further releases.

Acknowledgements

The authors want to acknowledge the Max Planck Society and the Project Consolider (CE-CSD2007-0005) of the CICYT for funding support. A. Orfila is acknowledge for computer support.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.syapm.2008.07.001.

References

[1] R.I. Amann, C. Lin, R. Key, L. Montgomery, D.A. Stahl, Diversity among *Fibrobacter* isolates: towards a phylogenetic and habitat-based classification, Syst. Appl. Microbiol. 15 (1992) 23–31.

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

[3] J. Brosius, T.L. Dull, D.D. Sleeter, H.F. Noller, Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*, J. Mol. Biol. 148 (1981) 107–127.

[4] F.D. Ciccarelli, T. Doerks, C. Von Mering, C.J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life, Science 311 (2006) 1283–1287.

[5] P. Dawyndt, M. Vancanneyt, H. De Meyer, J. Swings, Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources, IEEE Trans. Knowl. Data Eng. 17 (2005) 1111–1126.

[6] W.F. Doolittle, Phylogenetic classification and the universal tree, Science 284 (1999) 2124–2128.

[7] J.P. Euzéby, B.J. Tindall, Status of strains that contravene Rules 27(3) and 30 of the Bacteriological Code. Request for an opinion, Int. J. Syst. Evol. Microbiol. 54 (2004) 293–301.

[8] G.E. Fox, K.R. Pechman, C.R. Woese, Comparative cataloguing of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics, Int. J. Bacteriol. 27 (1977) 44–57.

[9] G.E. Fox, E. Stackebrandt, R.B. Hespell, J. Gibson, J. Maniloff, T.A. Dyer, R.S. Wolfe, W.E. Balch, R.S. Tanner, L.J. Magrum, L.B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B.J. Lewis, D.A. Stahl, K.R. Luehrsen, K.N. Chen, C.R. Woese, The phylogeny of Prokaryotes, Science 209 (1980) 457–463.

[10] G.E. Fox, J.D. Wisotzkey, P. Jurtshuk Jr., How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity, Int. J. Syst. Bacteriol. 42 (1992) 166–170.

[11] G.M. Garrity, Bergey's Manual of Systematic Bacteriology, second ed, Springer, New York, 2001.

[12] Judicial Commission of the International Committee on Systematics of Prokaryotes, Status of strains that contravene Rules 27 (3) and 30 of the International Code of Nomenclature of Bacteria. Opinion 81. Int. J. Syst. Evol. Microbiol. 58 (2008) 1755–1763.

[13] J.A. Klappenbach, P.R. Saxman, J.R. Cole, T.M. Schmidt, Rrndb: the ribosomal RNA operon copy number database, Nucleic Acids Res. 29 (2001) 181–184.

[14] V. Kunin, L. Goldovsky, N. Darzentas, C.A. Ozounis, The net of life: reconstructing the microbial phylogenetic network, Genome Res. 15 (2005) 954–959.

[15] C.G. Kurland, The paradigm lost, in: J. Sapp (Ed.), Microbial Phylogeny and Evolution Concepts and Controversies, Oxford University Press, Oxford, 2005, pp. 207–223.

[16] R. Lan, P.R. Reeves, Intraspecies variation in bacterial genomes: the need for a species genome concept, Trends Microbiol. 8 (2000) 396–401.

[17] S.P. Lapage, P.H.A. Sneath, E.F. Lessel, V.B.D. Skerman, H.P.R. Seeliger, W.A. Clark, International Code of Nomenclature of Bacteria (1990 Revision). Bacteriological Code, American Society for Microbiology, Washington, DC, 1992.

[18] W. Ludwig, H.-P. Klenk, Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics, in: D.R. Boone, R.W. Castenholz, G.M. Garrity (Eds.), Bergey's Manual of Systematic Bacteriology, second ed, Springer, New York, 2001, pp. 49–65.

[19] W. Ludwig, K.-H. Schleifer, The molecular phylogeny of Bacteria based on Conserved genes, in: J. Sapp (Ed.), Microbial Phylogeny and Evolution Concepts and

Controversies, Oxford University Press, Oxford, 2005, pp. 70–98.

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

[21] R.G.E. Murray, K.-H Schleifer, Taxonomic note: a proposal for recording the properties of putative taxa of prokaryotes, Int. J. Syst. Bacteriol. 44 (1994) 174–176.

[22] G.J. Olsen, D.J. Lane, S.J. Giovannoni, N.R. Pace, D.A. Stahl, Microbial ecology and evolution: a ribosomal RNA approach, Annu. Rev. Microbiol. 40 (1986) 337–365.

[23] J. Peplies, R. Kottmann, W. Ludwig, F.O. Glöckner, A standard operation procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes, Syst. Appl. Microbiol., in press.

[24] E. Pruesse, C. Quast, K. Knittel, B.M. Fuchs, W. Ludwig, J. Peplies, F.O. Glockner, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB, Nucleic Acids Res. 35 (2007) 7188–7196.

[25] R. Rosselló-Mora, Updating prokaryotic taxonomy, J. Bacteriol. 187 (2005) 6255–6257.

[26] R. Rosselló-Móra, R. Amann, The species concept for prokaryotes, FEMS Microbiol. Rev. 25 (2001) 39–67.

[27] P.H. Sneath, Evidence from *Aeromonas* for genetic crossing-over in ribosomal sequences, Int. J. Syst. Bacteriol. 43 (1993) 626–629.

[28] V. Sória-Carrasco, M. Valens-Vadell, A. Peña, J. Antón, R. Amann, J. Castresana, R. Rosselló-Mora, Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments, Syst. Appl. Microbiol. 30 (2007) 171–179.

[29] E. Stackebrandt, J. Ebers, Taxonomic parameters revisited: tarnished gold standards, Microbiol. Today 33 (2006) 152–155.

[30] E. Stackebrandt, W. Frederiksen, G.M. Garrity, P.A. Grimont, P. Kampfer, M.C. Maiden, X. Nesme, R. Rosselló-Móra, J. Swings, H.G. Truper, L. Vauterin, A.C. Ward, W.B. Whitman, Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology, Int. J. Syst. Evol. Microbiol. 52 (2002) 1043–1047.

[31] E. Stackebrandt, B.M. Goebel, Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology, Int. J. Syst. Bacteriol. 44 (1994) 846–849.

[32] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690.

Glossary

Ambiguities: calculated percent ambiguities in the sequences
ARB: a software environment for sequence data
DDBJ: DNA Data Bank of Japan
DNA: deoxyribonucleic acid
EMBL: European Molecular Biology Laboratory
FASTA: FAST-All
HGT: horizontal gene transfer
Homopolymers: calculated percentages repetitive bases with more than four bases,
INSDC: Nucleotide Sequence Database Collaboration
LPSN: List of Prokaryotic names with Standing in Nomenclature
LSU: large subunit
LSUParc: comprehensive LSU rRNA database
LSURef: Reference database for LSU rRNA
MPI: Message Passing Interface
nt: nucleotides
RAxML: Randomized Axelerated Maximum Likelihood
RNA: ribonucleic acid
rRNA: ribosomal RNA
SAM: Systematic and Applied Microbiology
SILVA: Comprehensive ribosomal RNA databases
SSU: small subunit
SSUParc: comprehensive SSU rRNA database
SSURef: reference database for SSU rRNA