Universitat de les
Illes Balears

Departamento de Biología

The All-Species Living Tree Project

TESIS DOCTORAL

Pablo Yarza Gómez-Galarza

Palma de Mallorca, 2011

Universitat de les
Illes Balears

The All-Species Living Tree Project

Tesis doctoral presentada por Pablo Yarza Gómez-Galarza para optar al grado de Doctor en Biología por la Universitat de les Illes Balears, bajo la dirección del Dr. Ramon Rosselló-Móra. y el Dr. Frank Oliver Glöckner.

PROGRAMA DE DOCTORADO:
Microbiología ambiental y Biotecnología

Director de la tesis	Director de la tesis	Ponente
Dr. Ramon Rosselló-Móra Investigador Científico del CSIC en el Instituto Mediterráneo de Estudios Avanzados	Dr. Frank Oliver Glöckner Head of Microbial Genomics and Bioinformatics group (Max Planck Institute for Marine Microbiology)	Dr. Rafael Bosch Zaragoza Profesor titular de Microbiología en la Universitat de les Illes Balears

a Adhara

a mis padres

Gracias Ramon, por tu gran generosidad, tu sabiduría, por dejarme la puerta abierta, por quererme y por creer en mí.

To the LTP team, Rudi, Karl, Wolfgang, Jean, Frank Oliver, Ramon, Jörg and Michael, thank you for your invaluable help, constant support and your proximity, for giving me the chance to work at the Max Planck Institute in Bremen, giving me a place in SAM, thank you for believing on the LTP and take it as part of your own scientific careers. I learned so much working with you.

A Ana, Jocelyn, Arantxa, Quico, Mercedes, Michael, Cifu, Raul, ha sido un placer trabajar con vosotros, os debo mucho... Muchas gracias por todo vuestro apoyo en Palma, por el buen rollo tan grande que tenéis y provocáis, sois muy grandes!

To the people at the MPI, to Elmar, Chris, Renzo, Ivo, Pelin, Wolfgang, and specially Sven, thanks for guiding me in Bremen and make my stays unforgettable.

A Francis, Fernando, Arantxa, Cristina, María, Manu, Judith, Pedro, y en especial a Pepa, mil gracias por hacerme un hueco en el departamento, por tratarme como a uno más de la familia, por creer en mí y ponerme en contacto con Ramon.

A todos mis compañeros de piso, Beni, Edu, Coti, Pablo, Inés (yoga), Cristina, Bertha, Inés y Mariana. Gracias por acogerme en la isla, por enseñarme los trucos de la vida independiente, por las risas, las historias de Argentina, Chile, y de México, el fuego en Esporles y en Palma, el arte, Sa Dragonera, y mil cosas más.

A Joseluis, Alicia, Gabriela, Jesús, Ángel, Alejandro, y toda la panda, a Pedro, Verónica Argentina, Alejo Colombia, y a todos los colegas que he ido recogiendo por el camino. A Sara, a los senderistas no madrugadores... A los que me dejo... Gracias a todos!

Anabel, Carlos, Roberto, Rubén, Horacio, gracias por El Pacto, por esperarme, por respetarme, por darme la música, por acompañarme en este viaje de rosas y espinas.

A Adhara, y a toda mi familia, Fernando, Marisa, Ignacio, Valle, Javi, Marisina, Víctor, Dani, Loren, Paula, Victoret, Cristina, Samuel, que sois mi punto de referencia, gracias por vuestra paciencia infinita, por el apoyo en los momentos más duros, por la motivación, la creatividad, por abrirme los ojos, os quiero.

INDEX

Introduction

Quick review about microbial taxonomy

Classification of life-forms has always been one of the most recurrent topics in science, as a need to obtain and share knowledge in biology. Aristotle was the first who dealt with the species concept around 400 years B.C, and our current perception of the order in nature is based since more than two centuries on Linnaeus' definitions. In contrast to botany or zoology, microbiology is a recent discipline that owes the technological advances (such as the microscope) its birth and development. Microorganisms were not apparent on fossil records, they were invisible to the naked eye and showed very low phenotypic polymorphism. Nevertheless, the first period in prokaryotic taxonomy history was completely influenced by the botanical and zoological systems, and it solely started with a morphology-based circumscription of new categories. In this regard, the advances on culturing techniques or the capability to obtain pure cultures were crucial to start to understand their physiological traits. Until the beginning of the 20th century several classifications had emerged in an independent way following different criteria but it was not until 1923 that a unifier and reference work was published: the *Bergey's Manual of Determinative Bacteriology*, which during the next editions would consolidate itself as the common framework for microbial taxonomists until today (now known as the *Bergey's Manual of Systematic Bacteriology*). In the middle of 20th century, microbiologists started to take profit of technological advances from other disciplines in order to improve old methods for bacteria characterization, hence approaching these organisms from a more biochemical point of view. With the increasing knowledge on the role that nucleic acids play in cellular life, researchers started for the first time to use genomic parameters as mol% G+C and DNA-DNA hybridization (DDH). DDH became a standard technique for new taxa circumscription and it mainly drove the prokaryotic taxonomic schema as we know it today. However, by the late 1970s, a last scientific contribution was the one that most significantly changed our view on prokaryotic classification. It was the use of “molecular clocks”, such as ribosomal genes or certain conserved proteins, to infer genealogical relationships from phylogenetic trees. The sequence of the small subunit of the ribosome (SSU) was the key to set up a more natural and reliable framework for classification of prokaryotes, the one we accept today. (Rosselló-Móra & Amann, 2001; Rosselló-Móra, 2005).

Nomenclature

However, at the early days of microbiology there was not a ruled nomenclature procedure. At the end of 19th century, the principles introduced by Carl von Linné motivated the generation of two independent codes of nomenclature. During more than 150 years, microbiologists used either the botanical code, the zoological code or just did not follow any kind of official regulations. During the first three *International Microbiological Congresses* (Paris 1930, London 1936, and New York 1939), the need for a code of nomenclature was recognized by the *International Society for Microbiology*, which authorized the creation of the "*International Committee on Bacterial Nomenclature*" in order to set it up. The first edition of the *Bacteriological Code of Nomenclature* was finally published in 1948, and edited for the first time in 1958 (International Committee on Bacterial Nomenclature, 1958). This committee was named later the *International Committee on Systematic Bacteriology* (ICSB) and now it is known as the *International Committee on Systematics of Prokaryotes* (ICSP). The ICSP is the international body within the *International Union of Microbiological Societies* (IUMS) that oversees the nomenclature of prokaryotes, and also determines the rules used to name them. Moreover its judicial commission is charged of other duties such as giving opinion about taxonomic matters, revising the Bacteriological Code, etc. After the Bacteriological Code was published in 1958, a new scientific journal for bacterial nomenclature started, the *International Bulletin of Bacterial Nomenclature and Taxonomy*, and it supposed the beginning of what we now know as the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) which is also governed by the ICSP.

The rule number 10 in the 1958 edition of the bacteriological code determined the moment when prokaryotic names started to be considered validly published. It was 1st May, 1753, the publication date of Linné's *Species Plantarum.* In order to review the status of all classified taxa before the first Code's publication, a complete revision between 1950-1960 was carried out to check whether they were adequately classified, accounted with available type strains, neotypes or adequate descriptions. This initiative was directed by V. B. D. Skerman on behalf of the ICSB (now ICSP), and led to a completely new start point for nomenclature of bacteria on 1 January 1980 by replacing the old date. Lists were made of names that could be satisfactorily associated with known bacteria, and these formed the foundation document, the *Approved Lists of Bacterial Names*, 1980 (Skerman *et al*., 1980). From the tens of thousands of names in the past literature, only 2,500 could be retained in the *Approved lists*. For the rest it was impossible to find out what specific bacteria they referred to. Names not being on these lists lost standing in nomenclature (though provision was made to revive old names subject to certain safeguards). From this moment on, all new names had to be published in the IJSB (now IJSEM) either by being described there, or, if described elsewhere, by placing them there in Validation Lists. This unification effort in bacterial nomenclature was one of the major advances for microbial taxonomy. The most recent edition of the Code is the *International Code of Nomenclature of Bacteria* (hereafter, Bacteriological Code), 1990 Revision, published in 1992 by the American Society for Microbiology (Lapage *et al*., 1992).

Microbial taxonomy has been always in a constant change as new advances in science and technology have allowed either increasing the number of species' descriptions as reviewing and emending existing taxa circumscriptions. Accordingly, nomenclature of taxa especially after the publication of the Bacteriological Code, has been a very active field in microbiology. Since January 2000, around 750 validly published names per year are entering into the IJSEM records. Whereas the number of novel species described and classified per year is lower, i.e. around 300-500 between years 2000 and 2005, and nearly 650 in last five years (Yarza *et al*., 2010). Therefore, the current number of ~ 10,500 validly published species names exceeds the total number of ~ 8,900 distinct species (updated, March 2011) due to the existence of homotypic synonyms, heterotypic synonyms and new combinations. In general, keeping up to date with changes in prokaryotic nomenclature has always been problematic. To facilitate the daily work of all microbiologists, specially taxonomists, all information widely distributed along IJSEM issues was condensed into a single web platform called List of Prokaryotic Names with Standing in Nomenclature (LPSN; Euzéby, 1997) conceived by Prof. Dr. Jean Euzéby, a member of the ICSP judicial commission and associated member of the Bergey's Manual Trust. It provides the most updated information regarding nomenclature of all taxa, provides links to the most relevant bibliographic contributions on each taxon, information about type strains and their current availability in culture collections, sequence entries, current opinion, didactic material and much more. In microbiology everybody knows that the valid name of a prokaryote can always be found in the LPSN.

Characterization and classification

Contrarily to nomenclature, there is neither a ruled procedure for taxa characterization nor an official classification of prokaryotes. Characterization of an organism consists on a detailed study of its properties and traits in order to: (i) be able to discriminate it from the previously characterized ones giving it a new place in classification and, (ii) give it a name. Whereas at the beginning of microbiology just morphological traits were used, later on, physiology, biochemistry and finally genetics were added in combination to what was finally called polyphasic approach (Vandamme *et al*., 1996) for taxa characterization. Additionally, dedicated subcommittees on taxonomy have been created as well to recommend the ICSP about the minimum standards for the description of certain taxa (e.g. subcommittee on the taxonomy of *Mollicutes*, subcommittee on the taxonomy of methanoarchaea). (Lapage *et al*., 1992). Overall, phenotypic and genomic methods are constantly being renewed in the light of the new technologies, but a general consensus exists today for all microbiologists who want to characterize strains with taxonomic purposes (Tindall *et al*., 2010). On the side of classification, the reference today is the Taxonomic Outline of the Prokaryotes covered by the Bergey's Manual of Systematic Bacteriology (Garrity, 2001).

One of the most important aspects to consider in a taxonomic study is the use of type strains. In last place, isolated strains are the subject of characterization and classification in microbiology. Ideally, the description of a new species takes into account the properties shared out among a series of distinct studied strains, but, since 1989, there is a tendency to describe new species and genera based on the characterization of a single isolate (Rosselló-Móra & Amann 2001; Christensen *et al*., 2001) probably motivated by the great importance attached to phylogenetic criteria. Nevertheless, Bacteriological Code just demands a description based on a nomenclatural type, that even not necessarily needs to be the most typical or representative in the taxon. The type for species and subspecies is, whenever possible, a strain that must be called the type strain. When cultures are not possible to maintain, a description, preserved specimen, or illustration may serve as the type. For genera, the nomenclatural type is a type species that is, in turn, is represented by a type strain. For families, a type genus (represented by a type species and a type strain) must be designated, and so on. Therefore, a type strain is the isolate that officially represents a taxon, and must be used to test the uniqueness of new taxa. For example, in order to prove the assignment at a certain genus, the type species of that genus is the most important reference organism to be used for comparison. Since 2002, the ICSP stated that a subculture of the original nomenclatural type must be deposited in, at least, two recognized culture collections from different countries to guarantee its preservation. The web service straininfo (www.straininfo.net) maintains an updated database of type strains available in biological resource centres (Dawyndt *et al*., 2005). Additionally for each strain, straininfo gathers together all entries from nucleotide databases, the latest valid nomenclature inherited from LPSN, and a historical record of the cultures' exchange between collections since the first author's submission.

By the half of last century, the ability to compare the structure of macromolecules brought a new era for microbiology. The use of cell walls, membrane lipids, and genetic methods such as mol% G+C or DNA-DNA (DDH) hybridization were proven to be much more robust techniques than those used before. Chemotaxonomy was especially useful but at the same time it was only applicable to certain taxonomic groups. DDH was accepted as the standard to genomically circumscribe species but it lacked resolution when comparing more distant organisms (i.e. higher taxonomic ranks). In summary, around the 1950s microbiologists accepted that the study of macromolecules had to be fundamental for taxonomy.

Undoubtedly, the most important breakthrough in microbial taxonomy occurred when the validity of the small subunit of the ribosome (SSU) as a universal phylogenetic marker was proved by Carl Woese and collaborators (Fox *et al*., 1977). It meant that genealogical relationships among organisms could be inferred from the comparative analysis of the primary sequence of the ribosomal genes, hence opening the door to a more natural classification. To be considered a phylogenetic marker, a molecule needs to show the following properties: (i) functional constancy, (ii) ubiquitous distribution and, (iii) high information content. (Ludwig & Klenk, 2001). All these features are fulfilled by the three ribosomal genes (16S, 5S and 23S rRNA, in prokaryotes). Although the 23S is the single molecule that exceeds in informative content (due to its larger size) to that of the 16S (28S over 18S, in case of eukaryotes), technical and economical issues made the SSU the most widely studied one. The 16S rRNA gene contains a range of variability along its primary sequence, which is especially useful for reconstructing phylogenies for a broad range of relationships (i.e. from species to domain level). During the following years, the database of SSU sequence entries increased rapidly, allowing reliable reconstructions from large input datasets. Since around 1990, taxonomists started to give more weight to phylogenies and, accordingly new species publications started to appear accompanied by the complete sequence of the SSU. In parallel, microbial ecologists introduced the PCR-based studies of microbial communities without the need of cultivation (Amann *et al*., 1995), contributing to the exponential growth of the 16S rRNA gene databases. Overall, there was clear that the comparative analysis of the SSU constituted a tool of paramount relevance and provided the key for a systematic of prokaryotes based on natural relationships. As a logical consequence of these facts, the second edition of the Bergey's Manual of Systematic Bacteriology launched the phylogenetic *backbone* of the prokaryotes, consisting on an updated and emended framework for prokaryotic classification based on rRNA sequence data (Garrity, 2001). Additionally, it was recommended that the almost complete sequencing of the 16S rRNA gene should be mandatory for any new species description (Stackebrandt et al., 2002).

Sequence repositories and data handling

Nucleotide sequences have to be submitted to one of the three databases members of the *International Nucleotide Sequence Database Collaboration* (INSDC; www.insdc.org): Genbank (USA, www.ncbi.nlm.nih.gov/genbank), EMBL (Europe, www.ebi.ac.uk) and DDBJ (Asia, www.ddbj.nig.ac.jp). On a daily basis, new submissions are exchanged among the three partners so all sequence entries are present in the three databases. As a consequence of last 30 years of activity in microbiology, the number of SSU-sequence submission to public repositories has reached unprecedented levels. The number of submissions of this gene per year has been growing exponentially since early nineties and currently exceeds the number of 3,000,000. Within this enormity it exists a vast range of quality, both in the sequence itself (e.g. short length, high number of ambiguities) and in the associated information to it (e.g. wrong species names, lack of strain information, etc.), thus hampering the task of preparing reliable initial datasets for phylogenetic reconstructions in taxonomic studies. However it is not exactly the duty of INSDC's databases to perform exhaustive quality controls, which indeed might not satisfy most of its users.

Evolutionary changes at the primary structure of the ribosomal RNA can be used for phylogenetic inference by recognizing homologous positions and arranging them into columns. Before a phylogenetic backbone for prokaryotes could be developed, a huge task of preparing a common and reliable alignment for *Bacteria* and *Archaea* was carried out by Wolfgang Ludwig and co-workers (Ludwig & Schleifer, 1994; Ludwig & Klenk, 2001). The presence of highly conserved regions, where positional orthology could be recognized unambiguously, facilitated the task of positioning the more variable ones. The observation that secondary structure formations such as loops and helices occurred at the same relative positions along the molecule reflects the underlying functional pressure. Thus, variable stretches with low sequence similarities could be optimally positioned by recognizing functional homology (i.e. positioning residues according to the high probability of participation into a loop or helix). Further, functional stability of helices (i.e. more than a half of the residues in the 16S rRNA may participate in helix formations) as indicated by canonical Watson-Crick base-pairings (G-C, A-U) helped to refine the alignments.

Consequently, dedicated databases have been designed to sieve, curate and enrich (i.e. with sequence associated information) the INSDC repositories in order to fit the requirements of the users community of microbial taxonomists. Three independent databases of curated ribosomal SSU and LSU were developed in parallel: RDP (USA, http://rdp.cme.msu.edu), Greengenes (Australia, http://greengenes.lbl.gov) and SILVA (Europe, www.arb-silva.de; Pruesse *et al*., 2007). In summary, their common objectives are: (I) provide updated universal alignments in order to achieve optimal and comparable phylogenetic reconstructions and (ii) produce and maintain curated datasets of nearly full length rRNA sequences to be used for in depth phylogenetic analyses.

At the beginning of the sequencing era (i.e. when just few sequence entries were available), Wolfang Ludwig and collaborators had the innovative idea of creating a database-driven software package for sequence data handling. It took more than 10 years to develop the ARB software package, and today is one of the most relevant tools for phylogenetics (Ludwig *et al*., 2004). The foundations of the ARB concept are: (i) a database of primary sequences that integrates any type of additional data (e.g. user-defined contextual data, phylogenetic trees, alignments, etc.) and, (ii) a comprehensive set of bioinformatic tools, that can interact with each other as well as the central database, which are controlled via a common graphical interface. One of the most important features was the innovation of a sequence editor that took into account the secondary structure of the 16S rRNA. A consensus for the secondary structure of the SSU was created using previously reported models and taking into account a refined dataset of archaeal, bacterial and eukaryotic sequences. Many gaps had to be created in order to keep the relative positions of the helices and loops of the distinct taxa, and to accommodate sequencing errors or just highly variable areas. The seed alignment was then imposed as a grid on the editor allowing to manually correct misplaced bases according to its potential belonging to a loop or a helix.

The SILVA project (www.arb-silva.de) started at the Max Planck Institute for Marine Microbiology in Bremen by complementing of the ARB project and both teams have been collaborating since years. Whereas the preparation and public release of curated ribosomal databases is basically done by SILVA, the ARB software development is mainly centralized at the Technical University in Munich.

Objectives

In order to produce a useful tool for the scientific community in which prokaryotic species classification can be retrieved in form of a phylogenetic tree, a project called "The All-Species Living Tree Project (LTP)" was initiated. It is an international collaboration between the scientific journal Systematic and Applied Microbiology (ELSEVIER) and the group of scientists responsible of the LPSN (www.bacterio.cict.fr), ARB (www.arb-home.de) and SILVA (www.arb-silva.de) projects. The job done on creation, maintenance, and management of the LTP during three years has been been summarized in the present thesis manuscript. The main objectives considered were:

1 Provide a curated SSU and LSU database of all the type strains of all species with validly published names, for which sequence entries of adequate quality exist.

2 Set up an optimized and universally usable alignment.

3 Reconstruct a single phylogenetic tree harbouring reliable topologies.

4 Provide regular updates of the database, alignments and trees with the new validly published taxa.

5 Create a web page for the project, where to host the complete set of materials and all data can be freely downloaded.

6 Investigate, with the use of the database, fundamental aspects about taxonomy of prokaryotes such as: phylogenetic thresholds in new taxa circumscriptions, coherence of current taxonomy by means of phylogenetic schemas and relevance of the 16S rRNA gene in taxonomic studies.

7 Achieve for the first time the complete catalogue of SSU sequences with all the hitherto classified species, consisting on the sequencing of the more than 500 species that still remain missing.

All chapters of the present manuscript have been originally written in English for publication in scientific journals of international scope. Thus, each one of the chapters has been presented in the way that they would be submitted for publishing. Details of publication are cited below.

Chapter 1 **Yarza, P., Richter, M., Peplies, J., Euzéby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., 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, 241-250.

Chapter 2 **Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.H., Glöckner, F.O., Rosselló-Móra, R.** (2010) Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst. Appl. Microbiol., 33, 291-299.

Chapter 3 **Yarza, P., Euzéby, J., Spröer, C., Mrotzek, N., Swiderski, J., Tindall, B.J., Pukall, R., Spring, S., Lang, E., Gronow, S., Verbarg, S., Klenk, H.-P., Crouch, A., Beck, B., Unosson, C., Moore, E.R.B., Nakagawa, Y., Clermont, D., Janssens, D., Sakamoto, M., Iida, T., Kudo, T., Kosako, Y., Oshida, Y., Ohkuma, M., Arahal, D.R., Spieck, E., Pommerening-Roeser, A., Figge, M., Park, D., Buchanan, P., Nicholson, A., Cifuentes, A., Schleifer, K.-H., Amann, R., Glöckner, F.O., Rosselló-Móra, R.** Taxonomic note: SOS, Sequencing Orphan Species: filling the gaps in the 16S rRNA gene sequence database for all classified species with validly published names. In preparation.

Chapter 4 **Yarza, P., Euzéby, J., Ludwig, W., Amann, R., Glöckner, F.O., Schleifer, K.-H., Rosselló-Móra, R.** Empirical circumscription of prokaryotic higher taxa based on comparative analyses of the 16S rRNA gene. In preparation.

Chapter 1

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

Published in Systematic and Applied Microbiology (2008).

DOI 10.1016/j.syapm.2008.07.001

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 9,975 single entries, 6,728 corresponding to type strain gene sequences, as well as 3,247 additional high-quality 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. 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.

1.1 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 (Fox *et al*., 1977). 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 (Fox *et al*., 1980). 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 (Ludwig & Schleifer, 2005). 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 (Stackebrandt *et al*., 2002). Actually, the current overview of the classification of prokaryotes is mainly based on genealogical affiliations (Garrity, 2001), 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 (Rosselló-Móra & Amann, 2001). 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 (Fox *et al*., 1992). 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 (Amann *et al*., 1992; Stackebrandt & Goebel, 1994), 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% (Stackebrandt & Ebers, 2006). Nevertheless, SSU analysis is important for inferring monophyly (Stackebrandt *et al*., 2002), 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 (Sneath, 1993) or horizontal gene transfer (HGT, (Doolittle, 1999)) 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 (Kunin *et al*., 2005), whereas, on the other hand, there are severe criticisms of how data are interpreted (Kurland, 2005). 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 (Lan & Reeves, 2000). 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 (Ciccarelli *et al*., 2006; Sória-Carrasco *et al.,* 2007). 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 (Ludwig & Klenk, 2001). 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. 1.1), 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 (Amann *et al*., 1995; Olsen *et al*., 1986). Consequently, only the minority of sequences corresponds to cultured prokaryotes (Fig. 1.1). 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 (Ludwig *et al*., 2004; Pruesse *et al*., 2007). 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 (Pruesse *et al*., 2007), and the SILVA website offers comprehensive databases of the aligned SSU and large subunit of the ribosome genes to the scientific community.

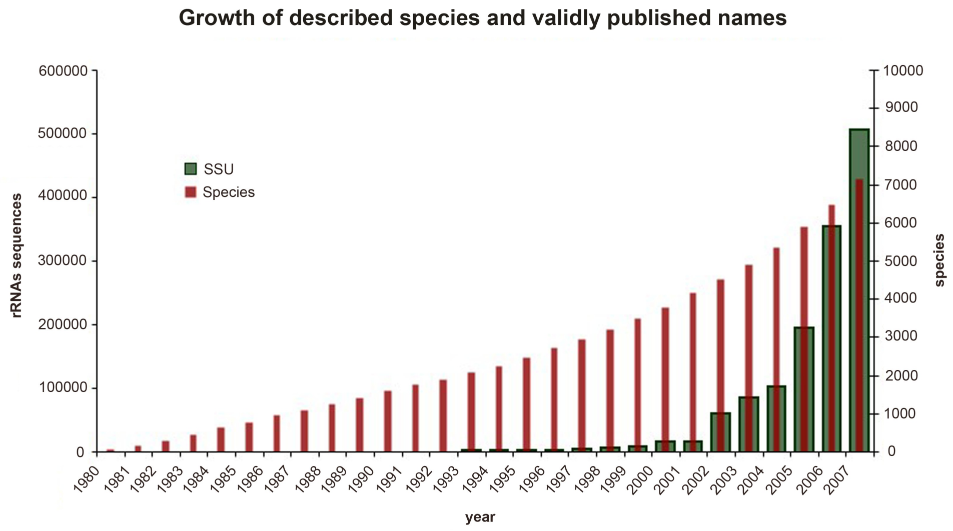


Fig. 1.1 Increase in the number of validated species from 1980 to 2007, and the SSU sequence submissions to public databases until SILVA release 93 (updated to 566,047).

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.

1.2 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. 1.1 shows the differences in the growth tendency of both databases. From the 8,264 validly published names until 31 December 2007, about 7,367 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; (Murray & Schleifer, 1994)) 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 (Euzéby & Tindall, 2004) 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 (Judicial Commission of the International Committee on Systematics of Prokaryotes, 2008). Among the *Cyanobacteria*, only the six species published under the Bacteriological Code rules (Lapage *et al*., 1992) 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 (>1,200 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* (Dawyndt *et al*., 2005). 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 1,500 sequences had wrong names (Table S1.1), 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 1,713 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 (Table S1.2), 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 (Table S1.3), 45 were listed in the SILVA SSUParc database, but were too short to be included in the SSURef database (Table S1.4), and, finally, a set of 54 sequences listed in the SSURef database were manually removed due to insufficient quality (Table S1.5). In Fig. 1.2, the final distribution of species is shown with regard to their sequence quality and usability in the living tree.

This sieving process selected a sequence database covering 6,782 species for which the type strain had an entry in the EMBL database. The final set of type strain sequences comprised 9,682 entries. The increase of type strain sequences in the database as a cumulative or yearly absolute number is summarized in Fig. 1.3. Fig. 1.3 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.

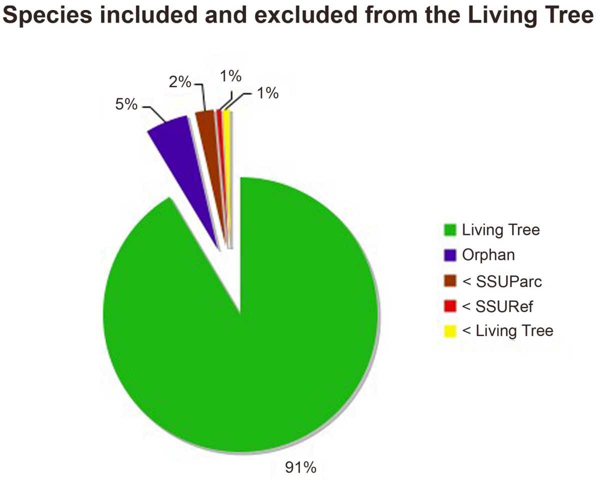


Fig. 1.2. Percentage distribution of (i) species with an adequate sequence for inclusion in the LTP_ARB tree (green); (ii) species (orphan) for which no sequence entry was found (blue); (iii) species with a sequence quality below the thresholds of the SSUParc database (brown); (iv) species with a moderate 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).

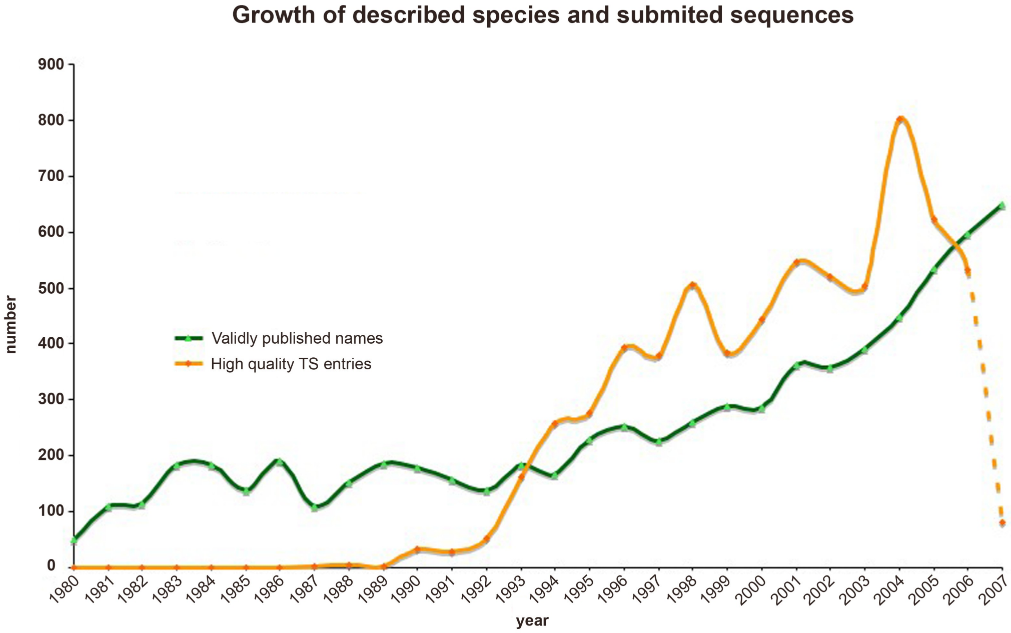


Fig. 1.3. Number of type strain sequences (orange) and validly published names (green) per year entering the public databases from 1980–2007.

A total of 4,982 species were recorded with a single sequence entry, whereas 1800 species had more than one entry. These together gave 4,700 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 (Klappenbach *et al*., 2001). The remaining 1,755 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 6,728 entries representing modest- to good-quality sequences of the 7,367 distinct species classified up to the end of December 2007 (Table S1.6). This final set is equivalent to about 91% of the complete catalogue of classified prokaryotic species (Fig. 1.2).

Finally, a selection of 3,247 additional sequences not belonging to any type strain was taken to complement the whole dataset, which gave a final number of 9,975 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.

1.3 Alignment improvements

Sequences had been automatically aligned by SINA, as implemented by the SILVA database project (Pruesse *et al*., 2007). 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 9,975 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 (Supplementary material -LTPs93), 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 1,524 nucleotides). As shown in Table 1.1, the maximum length of a sequence corresponded to the 2,210 nucleotide entry of *Pyrobaculum aerophilum*, which contains a large insertion of 712 nucleotides starting at *Escherichia coli* alignment position 373 (Brosius *et al*., 1981). The shortest sequence in the database corresponded to *Methanohalophilus portucalensis* with an entry of 1,229 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.

Table 1.1. Statistics for the LTP_ARB database

1A

	Min		Max		Mean	SD
Length	1,229	*Methanohalophilus portucalensis*	2,210	*Pyrobaculum aerophilum*[a]	1,465.38	50.65
nº ambiguities	0	*Shewanella putrefaciens*	30	*Sebaldella termitidis*	1.45	3.82
nº homopol.	0	*Leuconostoc carnosum*	26	*Pyrobaculum aerophilum*[a]	4.05	2.19
% vector	0	*Vibrio litoralis*	4.51	*Anaerobaculum mobile*	1.02	0.38

1B

	No. of sequences
No. ambiguities	4,674
No. homopolymers	27
No. ambiguities and homoplymers	17
No. ambiguities and homopolymers and full length [b]	1

[a] *Contains a long insertion*

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

1.4 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 1.2). This was designed to increase the signal to noise ratio and therefore improve the stability of the tree (Peplies *et al*., 2008). 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.

Table 1.2. 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	1,439
LTP_ssu_40	0	50,000	40	100	1,400
LTP_ssu_50	0	50,000	50	100	1,296

[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.

The complete dataset of 9,975 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; (Stamatakis, 2006)), 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. (Supplementary materials - LTPs93).

1.5 Some features of the tree

The tree, based on the data gathered until 31 December 2007, contains 6,728 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 3,247 additional sequences that were removed after tree reconstruction. Among the type strain sequences, 1,351 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 (Table S1.7). 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 (9,975 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, Thermodesulfobiacea*e). Such differences in representative sequences for each branch may promote unstable topology (Ludwig & Klenk, 2001). For this reason, the dataset was enlarged with an additional 3,247 sequences to provide a better balanced representation of phylogenetic branches. As a result, some of the incongruities 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 (Ludwig & Klenk, 2001).

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 intracellulari*s affiliates within the genus *Desulfovibrio, Xylanibacter oryzae* affiliates within the genus *Prevotella*; *Streptomyces longisporoflavus* affiliates within the genus *Brevundimona*s of the *Betaproteobacteria*; *Streptomyces gardneri* affiliates with the genus *Nocardia*). Some taxa appear paraphyletic or polyphyletic (e.g. the genera *Eubacterium, Bacillus, Pseudomonas, Desulfotomaculu*m), 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. 1.4), 28 families harboring three or more genera and 10 phyla harboring three or more families (Table 1.3) 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.

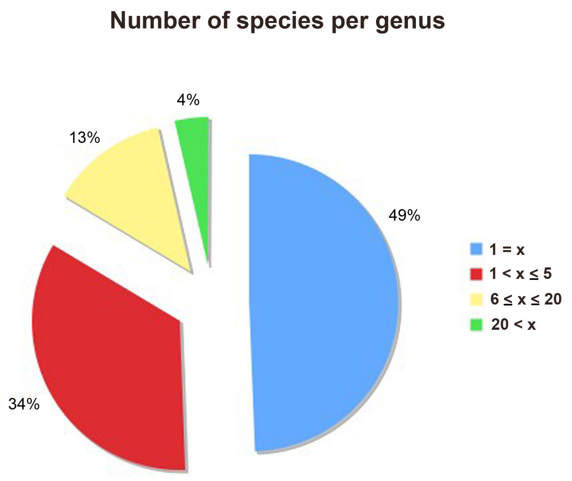


Fig. 1.4. Distribution of the number of species representing the pool of genera that could be identified within the framework of the all-species living tree. The LTP_ARB database contains 6728 species grouped into 1463 genera. A total of 710 genera harbored only one species, 492 contained between two 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 1.3. Boundaries at different taxonomic levels

	Genus	Family	Phylum
Number of taxa	451	28	10
Number of species	4,559	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 (Rosselló-Móra, 2005). The species circumscription and the resolution power of the SSU for improving this category definition has already been largely discussed (e.g. Ludwig & Klenk, 2001). 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 1.3 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.

1.6 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 (Table S1.8). 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 acknowledged for computer support.

Chapter 2

Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses

Published in Systematic and Applied Microbiology (2010).

DOI 10.1016/j.syapm.2010.08.001

The “All-Species Living Tree Project” (LTP) provides the scientific community with a useful taxonomic tool consisting of a curated database of type strain sequences, a universal and optimized alignment and a single phylogenetic tree harboring all the type strains of the hitherto classified species (Yarza *et al*., 2008). On the website http://www.arb-silva.de/projects/living-tree an update has been regularly maintained by including the 1301 new descriptions that have appeared in the validation and notification lists of the IJSEM journal. The topology of the 16S rRNA-based tree was validated with a detailed comparison against a collection of taxa-specific and broad-range trees made using different approaches, subsets of sequences and alignments. Seven percent of the classified species is still missing, as their type strains do not have a good quality SSU sequence. In addition, a new database of type strains for which adequate 23S rRNA entries existed in public repositories was built. Among the 8602 species with validly published names until February 2010, we were able to find good quality LSU representatives for 792 type strains, whereas around 91% of the complete catalogue still remains unsequenced. Despite the scarce representation of some groups in LSU databases, we have devised a highly optimized alignment and a reliable LSU tree in order to set up a stable phylogenetic starting point for taxonomic purposes. The current release corresponds to the fourth update of the project (LTPs102), and contains additional features which increase usability and compatibility. Use the contact address living-tree@arb-silva.de to provide additional input for the development of this taxonomic tool.

2.1 Introduction

More than ever, rRNA sequence analysis constitutes a sound and powerful approach in the daily routine of a microbial taxonomist. During the last three decades, the RNA gene sequences of the ribosomal subunits have revolutionized the way to infer genealogical relationships. These molecular markers became the basis for the phylogenetic backbone of microbial classification and today each new description of *Bacteria* and *Archaea* must be accompanied by the complete 16S rRNA sequence of the type strain (Fox *et al*., 1977; Ludwig & Klenk, 2001; Rosselló-Móra & Amann, 2001; Stackebrandt *et al*., 2002). The recent descriptions of three new phyla "*Gemmatimonadetes*" (Zhang *et al*., 2003), "*Caldiserica*" (Mori *et al*., 2009) and "*Elusimicrobia*" (Geissinger *et al*., 2009) endorse the high relevance of a SSU-based monophyly as a major premise for circumscribing new higher taxa based on a polyphasic approach (Vandamme *et al*., 1996; Rosselló-Móra & Amann, 2001). The increase of complete-genome sequencing efforts, such as the GEBA (Genomic Encyclopedia of Bacteria and Archaea) project, will open a new era for prokaryotic taxonomy, where novel procedures and methodologies will be needed for re-visiting existing gold standards (Wu *et al*., 2009; Richter & Rosselló-Móra, 2009). Nevertheless, we are still far from achieving ameaningful genomic overview of all validly published species due to the low coverage of the type strain collection by means of complete and draft genomes.

The number of informative positions in the sequence is one of the most important features that make a molecule a good phylogenetic marker. Whereas a reasonable number of conserved positions is needed to guarantee the identification of gene homology, the real informative portion of a molecule is given by the number of variable sites and the real number of possible allowed nucleotides per site (given functional and evolutionary constraints) (Ludwig & Klenk, 2001). Among the three rRNA genes, the 23S gene is considered to be the most informative chronometer as it doubles the size and information content of the 16S gene. However, technical and economical issues have facilitated the growth of 16S rRNA gene databases and it still remains the most suitable phylogenetic marker. In this regard, over 91% of the validly published species of *Bacteria* and *Archaea*, represented by their type strains, have a good quality SSU entry in public repositories (Yarza *et al*., 2008).

In the early 1980s, ribosomal gene sequence deposition started to increase exponentially in the public repositories. There are currently 10^6 SSU entries, and this is about one order of magnitude higher than the large subunit (LSU) entries in the SILVA-Parc (comprehensive) database (Pruesse *et al*., 2007). Nearly 500,000 SSU sequences and 17,000 LSU entries build the high quality SILVA-Ref datasets which are of very high quality for undertaking phylogenetic studies in depth (http://www.arb-silva. de/documentation/background/release -102/). Due to the fact that SSU has also become the tool for cultivation-independent analyses of microbial communities (Olsen *et al*., 1986; Amann *et al*., 1995) the rRNA databases are overloaded with clone sequences of uncultured organisms. Only a minority of the ribosomal sequences in the public databases corresponds to a cultured fraction of prokaryotes, including those of the type strains of the hitherto validly published names.

The type strain of a species is an isolate deposited in several international culture collections, and that is referenced in the protologue that describes the new taxon (Tindall *et al*., 2010). This strain is actually the taxonomic reference for any taxonomic study, as it guarantees the correct identification of new putative members of the same taxon. For this reason, it is necessary to use the sequences of the type strains for the phylogenetic reconstructions with identification or classification purposes. Actually, the uniqueness of a new taxon is firstly identified by means of the phylogenetic distance after genealogic reconstruction. Similarly to the species category, higher taxa are also represented by their type material (e.g. type species of a genus or type genus of a family), which is always represented in the last place by a type strain (Tindall *et al*., 2010).

One of our goals in the Living Tree Project (LTP) (Yarza *et al*., 2008) was the creation and periodic actualization of a curated dataset comprising high quality SSU sequences for each one of the classified species with hitherto validly published names. In the initial sieving process, the correct identification and labeling of type strain sequences was hampered in part because of the frequently inaccurate information submitted to the International

Nucleotide Sequence Database (INSDC; http://www.insdc.org). Incorrect species names, misassigned accession numbers, or wrong biological resource collection numbers were commonmistakes found, which made the sequence selection and interpretation of phylogenetic reconstructions difficult. In order to give the scientific community a reliable starting point for phylogenetic reconstructions with taxonomic purposes, we now maintain and correct an SSU database that also has an optimized and universally usable alignment of *Bacteria* and *Archaea* which takes into account the primary and secondary structure of the SSU. In addition, we provide a “state-of-the-art” tree reconstructed with the maximum likelihood algorithm RAxML (Stamatakis, 2006). This tree had been calculated with more than 10,000 sequences and was carefully checked against the currently accepted classification of prokaryotes (http://www.bergeys.org/outlines.html, http://www.bacterio.cict.fr/classifphyla.html). This, and additional material, has been integrated into the regular All-Species Living Tree (LTP) release, hosted by the SILVA project (http://www.arbsilva.de) and is compatible with the ARB software package (http://www.arb-home.de).

The LTP project has now been upgraded by extending its scope with the curation and analysis of the 23S rRNA gene sequences. Although we have basically followed the same procedure as for the SSU-LTP construction, the low number of prokaryotic LSU entries in the databases implied a challenge for phylogenetic reconstruction. Our improved LSU alignment and a 23S rRNA-based phylogenetic reconstruction represent the second contribution of the LTP project (Yarza *et al*., 2008). The curated dataset of type strain sequences and a large collection of supplementary material is also provided. By coupling the LSU release into the periodic LTP releases we intend to improve the usefulness of this tool for taxonomic purposes.

2.2 Updating the SSU-LTP

The curated all-species tree

The All-Species Living Tree and all its supplementary materials were published in August 2008. Briefly, the process consisted of: (i) a list of all classified bacterial and archaeal species up to December 2007, according to the LPSN (http://www.bacterio.cict.fr), excluding later heterotypic synonyms, *Cyanobacteria* whose names were not validly published under the terms of the bacteriological code rules, and the *Candidatus* category. Subsequently, the SILVA-SSURef 93 database was automatically checked to obtain a first round of candidate sequences. Each one of those labeled as type (T) or cultivable (C) organisms, which was information mainly originating from the http://www.straininfo.net bioportal (Dawyndt *et al*., 2005), was manually assigned to a type strain by proving equivalence of culture collection numbers. This crosscheck between species and sequences had to be done manually due to the high rate of mistakes in the EMBL fields. In a second round, we inverted the process by searching in EMBL for those species missing in the first subset of sequences. About 91% of the complete catalogue of classified prokaryotic species was finally represented in the phylogenetic tree. Among the remaining 9% missing type strains, nearly 4% were discarded due to the low quality of the available SSU sequences, and 5% because no entry existed in public DNA databases (we called them “orphan” species); (ii) since multiple copies of the ribosomal operon, and multiple independent submissions of the SSU rRNA gene sequence of the same organism, caused redundancy in our sequence selection, we applied quality criteria to select the best sequence available for a given species. Parameters such as sequence length, number of ambiguities, number of homopolymers, together with a manual inspection of the alignment were some of the criteria used to reduce the data. When multiple and identical copies of the SSU rRNA gene existed for a given genome, one was chosen randomly. These procedures yielded a final number of 6728 entries representing each one of the type strains of the 7367 classified bacterial and archaeal species up to December 2007; (iii) corrected organism name and additional meta-data were imported into new fields of a pre-existing SILVA template to build the LTP-ARB database. The total collection of 6,728 sequences had an average length of 1465 nucleotides and an average number of 1.45 ambiguities per sequence. A total of 4,674 sequences had no ambiguities, thus further corroborating the high quality of the selection. It was not possible to be more restrictive as we wanted to keep a meaningful number of type strains having a suitable SSU sequence for phylogenetic studies; (iv) the sequences were automatically aligned by SINA, as implemented in the SILVA project (Pruesse *et al*., 2007), using a manually curated seed of approximately 60,000 sequences. In addition, we manually revised the 10,000 (sum of type and non-type strains) selected sequences for our study in order to correct misplaced bases. The secondary structure of the SSU implemented in the ARB-editor was taken into account to refine the alignments. The improved alignment was afterwards introduced into the SILVA seed, hence improving the accuracy of SINA in forthcoming SILVA releases; (v) to exclude positions where positional orthology could not be guaranteed unambiguously, three filter sets were applied to increase the signal-to-noise ratio and therefore increase the stability of the tree (Peplies *et al*., 2008). Once filtered accordingly, the entire subset of sequences was submitted to different treeing approaches using the neighbor-joining, maximum likelihood and maximum parsimony algorithms. The final topology used in this release was produced by the maximum likelihood algorithm RAxML, calculated with 1,400 positions (40% filter), using GTRGAMMA correction, on a single run using a 5-node, 20-processor parallel environment; (vi) groups in the tree were recognized by the presence of the type species of the distinct taxa, and the topology was completely checked against previously established tree topologies for prokaryotes; (vii) the final product is a collection of materials intended to constitute a useful tool for microbial taxonomists.

The ARB database, alignments in fasta format, lists of EMBL entries with mistaken information, lists of “orphan” species, and the complete All-Species Living Tree in a pdf layout are some of the materials that are updated on a regular basis and publicly available on our web site: http://www.arbsilva.de/projects/living-tree. (Supplementary materials -LTPs95, -LTPs100, -LTPs102)

New SSU releases

Until the date of this publication, there have been four releases of the LTP (Table 2.1) accounting for newly classified species with validly published names (Euzéby, 1997). As nowadays new species descriptions are accompanied with the almost complete sequence of the 16S rRNA gene of their type strain, the updating process basically consisted of gathering the new information provided in the validation and notification lists of the IJSEM journal. However, an additional search in the public databases was always needed in order to find alternative SSU entries that might have been submitted independently of the species description, and their correspondence to the type strain was carefully checked. In case of new combinations, only the names were updated. In cases of heterotypic synonymy or rejection of given names, the sequences were removed from our database accordingly.

Table 2.1. Summary of LTP releases

Name of release	Date of release	SILVA release	IJSEM issue	No. sequences	Release information
LTP_s93	August 2008	SSURef-93	December 2007	6,728	Yarza *et al*., 2008
LTP_s95	October 2008	SSURef-95	June 2008	7,006	File S2.1
LTP_s100	September 2009	SSURef-100	August 2009	7,710	File S2.2
LTPs102_SSU	September 2010	SSURef-102	February 2010	8,029	www.arb-silva.de/projects/living-tree
LTPs102_LSU	September 2010	LSURef-102	February 2010	792	www.arb-silva.de/projects/living-tree

When multiple entries were found for a given type strain the removal of duplicates was carried out taking into account the alignment as the main indicator of sequence quality (Table 2.2). Up to 45%, 50% and 58% of the newly added sequences to the ‘LTP_s95’, ‘LTP_s100’ and ‘LTP_s102’ releases carried mistakes in important fields for taxonomy, such as the organism name. This increase in wrong submissions reflects a severe lack of updating effort in the submitted entries. The new sequences were added to the existing database, automatically aligned with SINA, and manually inspected in order to improve the positional orthology.

Table 2.2. Sieving the total number of sequences and species in LTPs102.

	Species[a]		Sequences[b]	
Total	8,602		167,493,839	
	SSU	LSU	SSU	LSU
Potential candidates	----	----	2,494,470	629,496
Above SILVA-Parc threshold[c]	----	----	1,246,462	180,344
Type strains in SILVA-Ref	8,029	792	11,154	1,854
Missing in LTP	573	7,810	----	----

[a] With validly published names up to the February 2010 issue of IJSEM.

[b] EMBL release 102 (December 2009)

[c] SILVA release 102. Comprehensive (Parc) and reference (Ref) datasets.

The new sequences in release ‘LTP_s95’ were added to the calculated and published tree of the ‘LTP_s93’ release (Yarza *et al*., 2008) with the ARB-parsimony tool. This special implementation of the parsimony embedded in ARB (Ludwig *et al*., 2004) allows the placement of the new sequences according to optimal criteria in a published tree without changing its topology. The large number of new sequences in release ‘LTP_s100’ made it necessary to calculate a new tree following the established procedures. Finally, the new sequences added to the current release ‘LTP_s102’ have again been added using ARB-parsimony, so that the current LTP-SSU topology (Fig. S2.1) is the one calculated for ‘LTP_s100’.

The ‘LTP_s102’ tree was compared in detail with a large collection of taxa-specific and broad-range trees made using different approaches, subsets of sequences and alignments. A subset of about 35,000 high quality sequences from SILVA databases was used to reconstruct the topology using the following algorithms and filters: a maximum parsimony (MP) tree optimized for each phylum with respective individual filters, maximum likelihood (ML) trees for each individual phylum and their neighboring groups with the respective individual filters, and a consensus tree extracted from all tree comparisons (data not shown). All trees were in strong agreement. The discrepancies found were within acceptable significance ranges. In the LTP tree, as in other “complete” trees, there was a low significance of the relative branching order of phyla, as indicated by relatively short branch lengths. Whereas some phyla may ‘jump’ in the topology (*Fusobacteria, Cyanobacteria, Acidobacteria*) depending on treeing procedures, other putative ‘super-phyla’ always appear as highly significant and stable clusters (*Bacteroidetes-Chlorobi, Chlamydia-Verrucomicrobia-Lentisphaerae*). Other analyses (Ludwig, 2010) do not clearly support a monophyletic structure of *Proteobacteria*, because there is a tendency to segregate *Deltaproteobacteria* and *Epsilonproteobacteria* from their counterparts. Additionally, for the phylum *Actinobacteria*, both in our tree as in others, there was a low significance of the relative branching order at intra-class and intra-order categories. For example, there is no support for a clear monophyletic structure of the orders *Kineosporiineae* and *Frankineae*. Moreover, special attention needs to be paid to the phylum *Fusobacteria*, where the comparison with other trees shows that this phylum has unstable rooting and maybe related to *Firmicute*s. In the case of *Firmicutes*, the remote position of *Thermoanaerobacterales* is well supported by other analyses.

The LTP has experienced an increase of 1301 sequences and species at a rate of about 650 descriptions each year. For the final number of 8602 validly published species, the update represents an increase of 15%. The growth of the phyla (Fig. 2.1) has been different. As expected, the higher increase in species was observed for *Proteobacteria* (533 taxa), *Actinobacteria* (344 taxa), *Firmicutes* (230 taxa), *Bacteroidetes* (165 taxa) and the archaeal group *Euryarchaeota* (29 taxa). These are also the prokaryotic phyla with the highest number of classified species. However, the increase of some groups (as indicated by the percentage of taxa classified in the period January 2008-February 2010) shows a differential trend that is not related to their representation in the taxonomic schema. For example, taking into account the growth of the largest phylum *Proteobacteria* (18%), *Bacteroidetes* exhibited an increase of 26% for new taxa and *Verrucomicrobia* 41% (from 19 to 32). A 100% increase belonged to the novel phylum *Caldiserica* (previouslyknownas candidate phylum OP5), described in 2009 with a single species and a single strain (Mori *et al*., 2009).

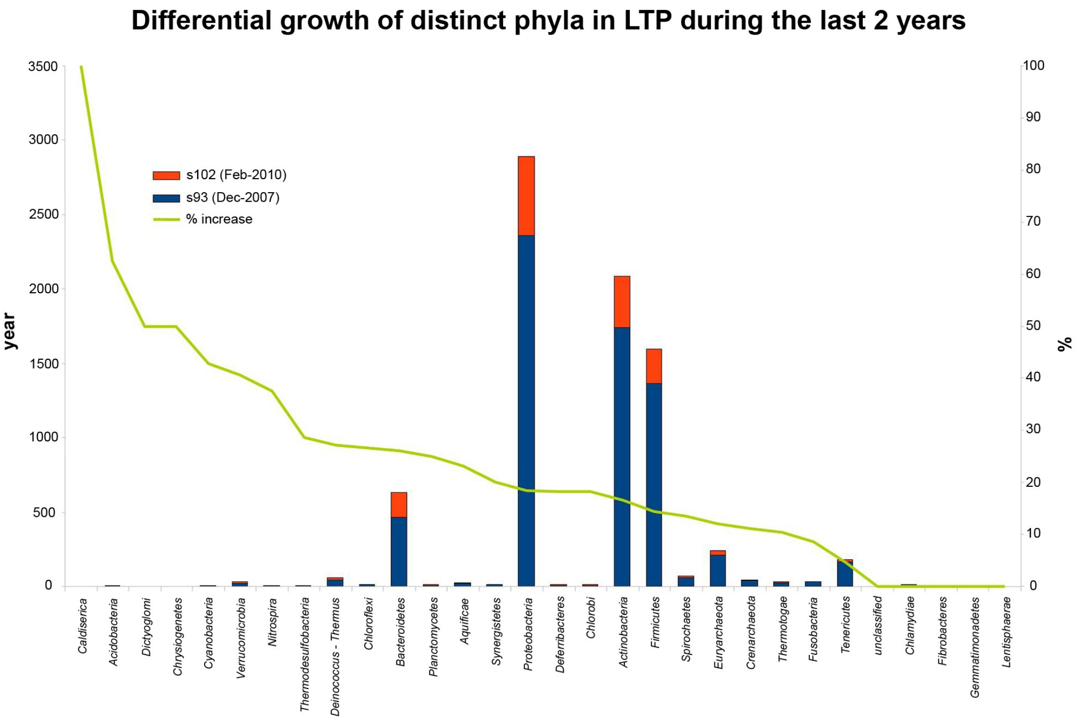


Fig. 2.1. The 29 bacterial and archaeal phyla represented in the All-Species Living Tree. Blue bars indicate classified species until December 2007 and red bars indicate classified species during the period between January 2008 and February 2010. The green line indicates the percentage increase of the red portion according to the overall abundance (red + blue).

Multiple ribosomal operons and LTP

Prokaryotic species may have more than one copy of the ribosomal (rrn) operon in their genomes, with a maximum of 15 found in *Clostridium paradoxum* and *Photobacterium profundum* (Klappenbach *et al*., 2001; Lee *et al*., 2009). Additionally, the distinct copies within the same genome may not be identical in terms of 16S rRNA gene similarity. Such intra-genomic variability has been previously reported and broadly discussed as a harmful phenomenon for phylogenetic inference and diversity measurement in nature (Mylvaganam & Dennis, 1992; Acinas *et al*., 2004; Clayton *et al*., 1995). In order to provide additional support to the All-Species Living Tree project, a comprehensive study was carried out to evaluate the intra-genomic variability on complete type strain genomes. Up to March 2009, we were able to retrieve 943 strains catalogued in the Ribosomal RNA Database (*rrn*DB http://rrndb.cme.msu.edu) and their 2,965 SSU sequences belonging to bacterial and archaeal genomes in SILVA release 98 (http://www.arb-silva.de/). Both the distribution of the *rrn* operon copy number (Fig. 2.2) and the maximum intragenomic divergence (Fig. 2.3) were analyzed using two datasets: the complete group of strains and a subset composed of 173 strains corresponding only to type strains. Inter-operonic divergence was measured in terms of the lowest sequence identity found among the distinct SSU copies within the same genome. Gaps, transversions and transitions were equally penalized and the reliability of the distance matrix was guaranteed by the highly curated alignment. The results showed that: (i) currently, a dataset made up with only type strains constitutes a good representation of the complete list of sequenced genomes, as both display equivalent distributions of the *rrn* operon copy number and intra-genomic divergence; (ii) despite the fact that the vast majority of strains host multiple copies of the *rrn* operon, only 2% of them contain divergences beyond 2% (30 nucleotides) sequence identity. Thus, most likely, the selection of one or another copy should not affect the phylogenetic reconstructions.

It is clear that the intra-genus (94.5%, (Yarza *et al*., 2008)) or intra-species (98.7%, (Stackebrandt & Ebers, 2006)) boundaries calculated might in some cases be exceeded within a single genome. For example, *Haloarcula marismortui* ATCC 43049, with 5.7% of maximum inter-operonic divergence (Mylvaganam & Dennis, 1992), or *Thermoanaerobacter pseudethanolicus* ATCC 33223, with 3.66%, both represent unusual exceptions with rather large sequence divergences, where the selection of one or another sequence might seriously affect the interpretation of a phylogenetic inference. Taking into account such preliminary results, we agreed that for the current SSU and LSU releases we would keep just one sequence for each species. The decision was taken in the first instance because generally most paralogs are nearly identical, and, in the second instance, the inclusion of all rrn copies might cause local branch attraction given the low inter-operonic variation, and, in the third instance, the usefulness of a tree-based taxonomic tool relies on the simplicity in recognizing distinct taxa by the presence of their sequences positioned along a reliable phylogeny.

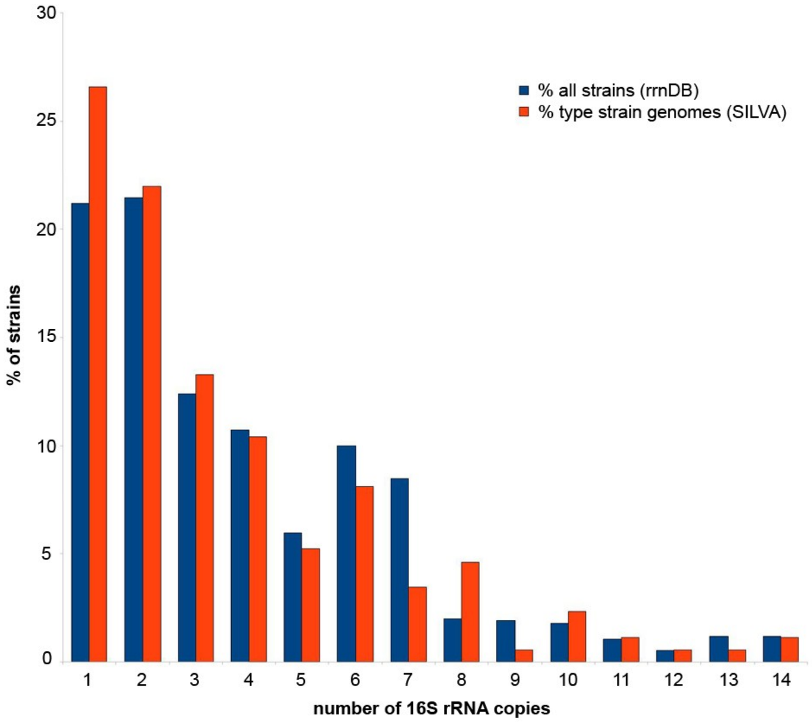


Fig. 2.2. 16s rRNA gene copy number. Up to March 2009, 943 strains (blue bars) were represented in the rrnDB database for which the ribosomal SSU gene copy number was known. Accordingly, 173 type strains (red bars) for which genomic sequences were available could be retrieved from SILVA release 98.

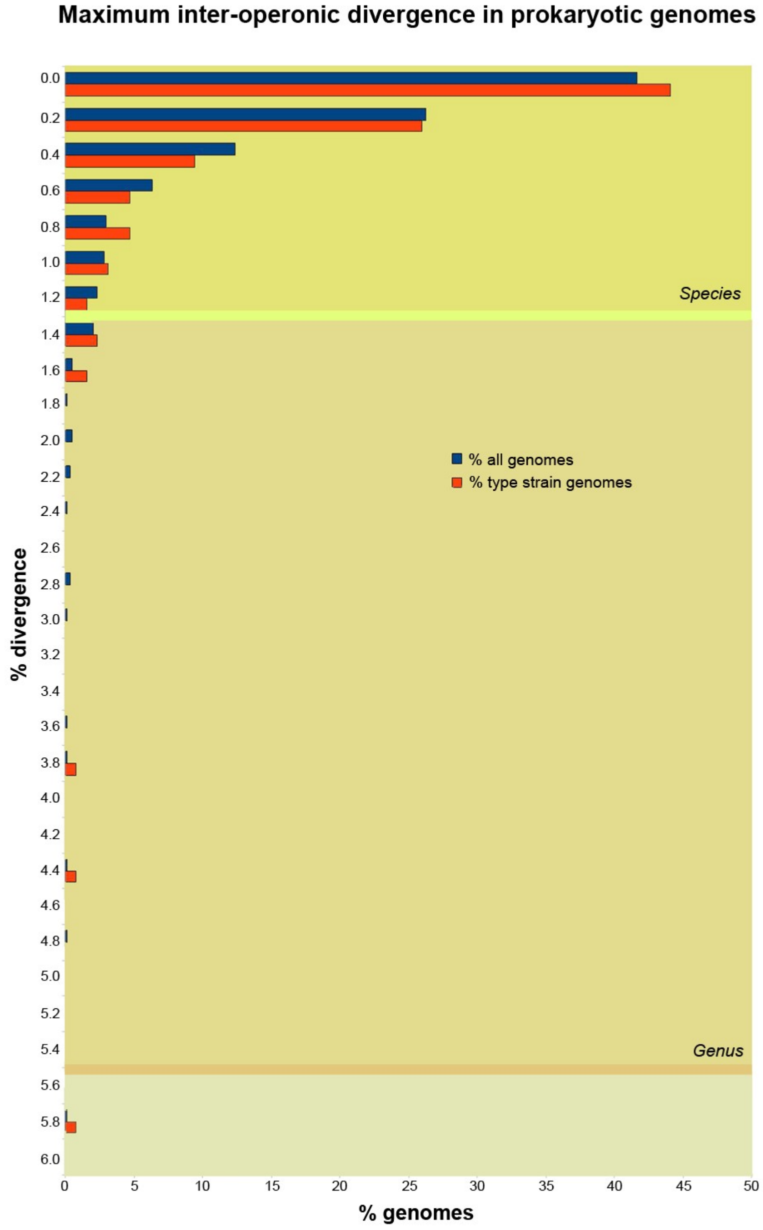


Fig. 2.3. Divergence based on 16S rRNA gene sequence similarity represents the number of ‘changes’ per 100 nucleotides (*Y* axis). Gaps, transversions and transitions were equally penalized. Species and genus boundaries were fixed at 1.3% of the sequence divergence, according to Stackebrandt & Ebers, 2006, and 5.5%, according Yarza *et al*., 2008, respectively. Blue bars represent 598 sequenced genomes having multiple copies of the ribosomal operon and red bars constitute a subset made with 127 type strains.

2.3 The LSU-based All-Species Living Tree

Data selection

As with the SSU, the LTP-LSU construction needed a first manual checking step. The updated list of prokaryotic species with standing in nomenclature was carefully crosschecked against the sequence database (Table 2.2). Making use of the already existing information on type strains in SILVA, regularly synchronized against the Straininfo (Dawyndt *et al*., 2005), RDPII (Cole *et al*., 2009) and EMBL (Kulikova *et al*., 2007) databases, an extensive list of potential candidates to represent the type strains was built automatically with the sequences labeled as type (T), cultured (C) or associated to a genome project (e[g]). From the total number of 629,496 potential LSU candidates in EMBL, only 16,966 were recognized as of acceptable quality (>1900 nt for the Ref database) and among these, 5,654 were considered in the automatic shift. Each sequence was manually assigned to a species by verifying the correct culture collection number assignations. In contrast to what is usually observed in SSU datasets, only about 6% of all entries carried errors either in the organism name (Table S2.1), strain numbers, or both. In this case, the mistakes in the entries were mainly due to the use of basonyms or outdated naming. The crosscheck finished with the number of 1,854 sequences that represented 792 distinct species (Table S2.2). The redundancy in the LSU dataset was reduced using the best sequence available for a given type strain (Yarza *et al*., 2008), where a manual revision of the alignment was always a decisive criterion. Among the 792 species covered, 347 accounted for more than one sequence that added up to 1,409 LSU entries. Among them, 300 type strains had multiple paralogs and 47 were represented by multiple independent submissions. Nearly 91% of the complete list of prokaryotic species seemed to be orphan in the LSU database.

Sequence alignments and phylogenetic reconstruction

The complete dataset of 1,854 type strain sequences had been previously aligned with the SINA aligner (Pruesse *et al*., 2007), using a seed of about 2,800 high quality LSU entries. Despite the overall good quality of the seed alignment, it did not fulfill the requirements of the All-Species LSU database. Hence, we made a complete revision that consisted of increasing the number of sequences in the seed and extending the alignment length. A stringent quality filtering approach was applied for data selection. Sequences comprising at least 2600 nucleotides and covering positions homologous to *E. coli* position 10 to 2895 were used for preliminary tree reconstruction. Clustering the tree topology was carried out by applying the ARB OTU (operational taxonomic unit) tool (Westram *et al*., 2010) followed by quality checking with the ARB sequence quality tool. Quality scores were assigned to the individual sequence entries taking into account the number of ambiguities and deviations from the overall and individual cluster primary and secondary structure consensus. The final number in the seed alignment was 7,900 quality checked almost full length LSU entries. Only 300 of the type strain sequences successfully passed the rigorous quality filtering. The final alignment comprised 80,376 positions. This new seed will be adopted by the SILVA pipeline to improve the quality of forthcoming regular LSU updates.

For tree reconstructions, the seed data set was reduced by replacing clusters of highly similar (>99%) sequences by its single ‘best’ representative using the respective feature of the ARB OTU tool. The resulting data was comprised of 1,900 non-redundant high quality sequences. Maximum likelihood and parsimony-based trees were reconstructed by applying various positional conservation filters. The missing partial or lower quality type strain sequences were added to these trees using the ARB-parsimony tool in combination with the option for keeping the initial topology while inserting additional data. The tree version provided by the LTP project was based on a maximum likelihood tree of the core set generated in combination with a 50% conservation filter for prokaryotes selecting 2,463 alignment positions. Applying such a restrictive filter means losing resolution power at lower taxonomic levels; however, it reduces branch attraction effects, thus potentially influencing topologies for higher ranks (Ludwig & Klenk, 2001; Ludwig *et al*., 2009). The groups shown in the tree are supported by the majority of treeing approaches performed. Group definition in the tree was performed by recognizing the type members according to the taxonomic classification. The tree was carefully compared against previously reported topologies (Garrity, 2001; Ludwig *et al*., 2009) and current taxonomic classifications (http://www.bacterio.cict.fr). In most of the cases, the branching order resulted in strong agreement. However, trees are dynamic structures that can change depending on the quality and abundance of data; hence, by the addition of new sequenced type strains, we can improve the resolution in some areas that may appear ambiguous in this first LSU release.

Some features of the ARB database

The ARB database of the LSU rRNA sequences is composed of just 792 type strain sequences; hence, only 9% of the complete classification list is represented. All the additional supporting sequences used to reconstruct the phylogeny were removed from the final LTPs102-LSU tree (Fig. S2.2), but its topology was kept intact. The type species are labeled as such in the ARB database and assigned to color group 10 for easy tree handling. The LTPs102-LSU database is composed of high quality nearly full length sequences. The minimum length found was 1931 nt for *Pyrobaculum calidifontis*, with a maximum of 3720 nt for *Thermotoga neapolitana* that contained a large insertion of 699 nucleotides at *E. coli* position 1931. The average sequence length was 2866 nucleotides. About 92% of the 792 type strain sequences did not contain any ambiguity.

Topology remarks

It is understood that the present reconstruction does not reflect the correct topology of all represented taxa in the tree. The most important problem faced was the scarce representation of some groups in the EMBL database leading to under-sampling. One negative effect is the branch attraction phenomenon, which is the most plausible explanation for some incoherent affiliations. For example, the order *Bifidobacteriales* affiliates with *Actinomycetales*. At a deeper level along the topology it is possible to recognize some species that, due to the incongruent affiliation, may be susceptible to reclassification. Most of these cases are also supported by the SSU-based trees (Yarza *et al*., 2008). We can find genera adequately classified (e.g. *Aeromonas, Helicobacter, Bartonella, Burkholderia, Campylobacter, Chlamydophila, Mycobacterium, Bifidobacterium, Streptococcus, Pyrobaculum*); species that do not coherently affiliate with the members of the same genus (e.g. *Leuconostoc fallax*); or genera that appear to be paraphyletic or polyphyletic (e.g. *Bacillus, Enterococcus, Rhizobium, Clostridium, Lactobacillus*).

Calculating taxa boundaries

In order to understand how the taxonomic categories genus and family could be circumscribed in terms of 23S rRNA similarities we carried out a statistical approach (Yarza *et al*., 2008). To delineate the distinct taxa numerically, the sequence identity against the type member was used as the single measured parameter. The whole process was automated into a perl-written program, which, in summary, conducted the following steps: (i) the 792 sequences provided in the LTP database were sub-sampled into multiple single-genera, single-families, or single-phyla matrices depending on the taxonomic level under study; (ii) categories having less than three taxa were excluded from the calculation; (iii) the similarity values of each species against their type member were extracted for statistical analysis; (iv) a soft removal of outliers (p50 – 2.5SD) was carried out in order to avoid clearly misclassified taxa and hence remove noise from the data; (v) minimum, maximum, and average similarity, and the corresponding standard deviations were the descriptors calculated for each taxon; (vi) averages of all taxa of a certain taxonomic category and 95% confidence intervals are displayed in Table 3.3. The results showed that the minimum value that guaranteed the circumscription of a new genus would be 93.2% ±1.3 to the corresponding type species. Family boundaries would be fixed by a minimum of 87.7%±2.5 and phylum boundaries by a minimum of 75.3%. Cutoff values based on LSU sequences were slightly lower than those reported previously for the SSU (Yarza *et al*., 2008), which might reflect the higher information content of the LSU. The number of type species represented by good quality LSU sequences in public databases is 257 (14.5% of 1779 classified prokaryotic genera), thus providing a very small number of measurements for the statistical analysis. We expect that the error in the calculated boundaries will decrease as more type strains are sequenced. However, and especially for genus boundaries, we observed a reduced variation in the averages. This may reflect a general consensus among the scientists when creating taxonomic categories. Hence, a similarity cutoff of 92% (LSU) and 94.5% (SSU) maybe confidently used as genus boundaries for prokaryotic organisms. Although these values cannot be used as a strict criterion, they may contribute to the construction of taxonomic schemes and taxa delineations.

Table 2.3. Boundaries at different taxonomic levels

	Genus		Family		Phylum	
	LSU	SSU[a]	LSU	SSU[a]	LSU[b]	SSU[a]
Number of taxa	33	451	12	28	1	10
Number of species	212	4559	39	202	6	195
Maximum identity	97.1% ± 1.0	98% ± 0.2	90.9% ± 2.7	92.5% ± 1.2	84.5%	84.7% ± 1.9
Average identity	95.2% ± 1.1	96.4% ± 0.2	89.3% ± 2.4	90.1% ± 1.1	79.95%	81.7% ± 1.8
Minimum identity	93.2% ± 1.3	94.9% ± 0.4	87.7% ± 2.5	87.5% ± 1.3	75.31%	78.4% ± 2.0

[a] Values taken from Yarza et al., 2008

[b] The single phylum in the LTP-LSU database represented by three or more families, each represented by the type species of the type genera was *Bacteroidetes*. Maximum and minimum identity values belong to single observations and there is no possibility for mean and confidence interval calculation. The average identity was calculated with the similarities shown by the representative species of *Chitinophagaceae*, *Cytophagaceae*, *Prevotellaceae*, *Rhodotermaceae* and *Sphingobacteriaceae* against the type family *Bacteroidaceae*.

2.4 Final considerations

Mistakes in INSDC entries

As previously discussed (Yarza *et al*., 2008), the high rate of mistaken information in public repositories is a source of confusion. This is especially important when taxonomic information, such as organism names and strain number assignations, are required. This is not a trivial issue given that INSDC members, such as EMBL, host the data that many microbiologists and bioinformaticians use daily in their research. At the moment a change in nomenclature is published, all entries associated to this organism will be automatically outdated. Therefore, there is a need for the original author submitting the sequences to update his/her entries. The existence of curated databases, like SILVA or LTP, provides corrected information that prevents users making a wrong choice. However, database refining needs manual and time-consuming inspection, and cannot be completely automated. In the case of SSU entries, the sequence submission generally occurs months before the effective publication of the species name (Fig. 2.4). In many cases, the entry is only identified with a genus-level affiliation (e.g. *Bacillus* sp., *Friedmanniella* sp.). In other cases, the name of the species submitted (e.g. *Xenophilus aerolata*) does not match the bacteriological code rules and the etymology is revised prior to the effective publication (e.g. *Xhenophilus aerolatus*) (Lapage *et al*., 1992). Additionally, human errors during submission might occur and the name of the organism could simply be wrong. Following these examples, the entries in the INSDC databases could be incomplete, or even mistaken. Finally, and even if the entry was originally correct (*Rhodoferax ferrireducens*, Finneran *et al*., 2003, sp. nov.), reclassifications may lead to nomenclatural changes (*Albidiferax ferrireducens* (Finneran *et al*., 2003) Ramana and Sasikala 2009, comb. nov.) without entry update (CP000267).

In contrast to the SSU database, errors in LSU data are much less frequent. As reflected in Fig. 2.4, among the 792 sequences selected in the LSU-LTP database, 280 submissions came from genome projects during the last 15 years (http://www.arb-silva.de; (Liolios *et al*., 2008)). In general, the new LSU submissions are based on already described species, rather than on new descriptions. This fact explains the low degree of mistaken entries in the public repositories. However, the expected increase in complete-genome sequences during the coming years, in some cases with dubious specific name assignation (Richter & Rosselló-Móra, 2009), may lead to an increase of mistaken or outdated meta-data. Therefore, we strongly suggest to the editors of microbial taxonomy journals to encourage the submitters to update the SSU/LSU-associated meta-data as a mandatory step prior to manuscript acceptance.

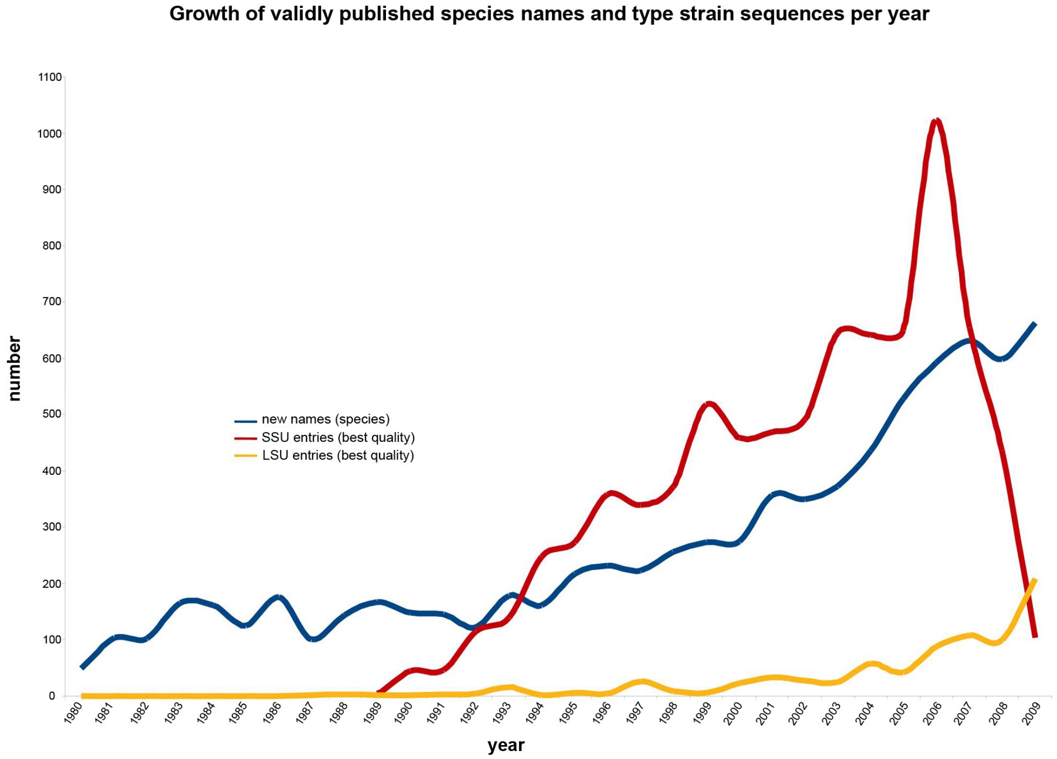


Fig. 2.4. Number of type strain SSU (red) and LSU (yellow) sequences, and validly published species names (blue) per year in the public databases during 1980–2009.

Taxonomic coverage in the All-Species Living Tree

Table 2.4 summarizes the current taxonomic coverage of the SSU/LSU databases. These observations are based on sequences submitted up to SILVA release 102 and the latest update of the validation and notification lists of the February 2010 issue of IJSEM. About 7% of the hitherto classified species are still missing in the SSU dataset, leaving a total number of 108 genera and 11 families unrepresented. However, all higher taxa beyond the family level have at least one representative. The situation for the LSU is almost inverse, where 91% of all species are still absent. In addition, only 4 of the 29 classified phyla were missing. The low coverage at the class, order and family ranks highlights the shortage of LSU data.

Table 2.4. Taxonomic coverage of the All-Species Living Tree

Rank	Complete catalogue[a]	SSU tree	LSU tree
Domain	2	2	2
Phylum	29	29	25 (13.8%)
Class	52	52	43 (17.3%)
Subclass	5	5	5
Order	115	115	89 (22.6%)
Suborder	20	20	18 (10%)
Family	285	274 (3.9%)	194 (31.9%)
Genus	1,779	1,671 (6%)	386 (78.3%)
Species and subspecies[b]	8,602	8,029 (6.7%)	792 (90.8%)

Percentages refer to missing taxa in the LTP due to the absence of a high quality SSU/LSU entry in EMBL release 102.

[a] Complete catalogue of classified species up to the February 2010 issue of IJSEM.

[b] Numbers refer to distinct type strains and not distinct names (i.e. the type subspecies and the species share the same type strain and have been counted once)

Table 2.5. Validly published taxa in the LTPs102

	Total	SSU		LSU	
		Present	Missing	Present	Missing
Proteobacteria	3,141	2,890	251	334	2,807
Actinobacteria	2,146	2,086	60	75	2,071
Firmicutes	1,684	1,592	92	171	1,513
Bacteroidetes	668	633	35	48	620
Euryarchaeota	269	241	28	37	232
Tenericutes	201	176	25	15	186
Spirochaetes	109	67	42	6	103
Deinococcus-Thermus	61	59	2	6	55
Crenarchaeota	56	45	11	19	37
Fusobacteria	37	35	2	14	23
Thermotogae	32	29	3	11	21
Verrucomicrobia	32	32	0	7	25
Aquificae	28	26	2	3	25
Chloroflexi	19	15	4	6	13
Synergistetes	17	15	2	3	14
Chlorobi	15	11	4	7	8
Plantomycetes	15	12	3	8	7
Chlamydiae	13	13	0	9	4
Deferribacteres	11	11	0	1	10
Nitrospira	10	8	2	1	9
Acidobacteria	8	8	0	3	5
Cyanobacteria	8	7	1	2	6
Thermodesulfobacteria	7	7	0	0	7
Unclassified	4	1	3	0	4
Fibrobacteres	3	2	1	1	2
Chrysiogenetes	2	2	0	0	2
Dictyoglomi	2	2	0	2	0
Lentisphaerae	2	2	0	2	0
Caldiserica	1	1	0	0	1
Gemmatimonadetes	1	1	0	1	0

A comparative overview of the taxonomic coverage between SSU and LSU datasets is given in Table 2.4 and Table 2.5. The same taxonomic groups have been the most sequenced ones, either with SSU and LSU, and this is directly related to their overall species content. However, the situation for LSU is inverted, meaning that the completed side is as low as the missing part of the SSU. In order to maximize the taxonomic coverage with the LSU marker, we suggest that the sequencing effort (including forthcoming genome projects) must follow this rationale: first type families of the missing orders, then type genera of the missing families, then type species of missing genera and finally the rest of the type strains.

SOS: sequencing the orphan species

In the course of constructing our type strain database, it was found that about 7% of the classified species of prokaryotes with validly published names did not have an SSU gene sequence entry or the one deposited was of insufficient quality to be considered for our purposes (Table 2.4). Consequently, eleven international culture collections (DSMZ, CCUG, NBRC, CIP, LMG, CECT, ATCC, JCM, NCCB, ICMP and BKF) have already joined us in the initiative to sequence these orphan type strains (unpublished data). Therefore, it is expected that we will soon achieve the complete set of 16SrRNA gene sequences for all classified species of *Bacteria* and *Archaea* with validly published names.

Acknowledgements

The authors acknowledge the Elsevier Publisher, Max Planck Society, and the Spanish Ministry of Science and Innovation (Project Consolider Ingenio 2010 CE-CSD2007-0005 co-financed with FEDER funding) for their economic support to the Project. M. Richter and J. Peplies are acknowledged for their informatics support.

Chapter 3

Taxonomic note: SOS, Sequencing Orphan Species: filling the gaps in the 16S rRNA gene sequence database for all classified species with validly published names

Manuscript in preparation

High quality ribosomal RNA sequences from type strains of all described species with validly published names is a prerequisite for accurate phylogenetic reconstructions, taxonomy as well as sequence classification. Over the last years the Living Tree Project (LTP) has taken care to create a high quality, aligned 16S and 23S gene database of all type strains (Yarza *et al*., 2008; Yarza *et al*., 2010). The worldwide collection of sequences and type strain information revealed that in total 564 "orphan" species had no 16S rRNA gene in the database. For 327 of the validly named species no entry was found in the public repositories. 237 type strains whose 16S rRNA sequence entries were of bad quality needed to be discarded. The undersigning researchers and culture collections have collaborated to improve this situation by (re)-sequencing these "orphan" species in order to report the 16S rRNA gene sequences for the type strains of 390 species. The remaining 174 species could not be retrieved from public collections because either they were non-cultivable (115) or the cultures had been lost (59). The designation of a neotype for the lost cultures or their removal from the validation lists is recommended to achieve a full coverage of the 16S rRNA gene database of all classified prokaryotic species.

3.1 Text

The reconstruction of genealogies among prokaryotes is mainly based on the comparative analysis of the 16S rRNA gene sequences. Microbial systematics and taxonomy have importantly benefited from these developments up to the point that the current classification reflects their genealogical relationships (Garrity, 2001). Despite other conserved functional genes can give similar resolution (Ludwig, 2010), the database of this gene is exponentially increasing and overtaking in orders of magnitude any other gene's rise (Pruesse *et al*., 2007). One of the major drawbacks of the public databases containing the small subunit´s gene sequence is that over 97.5% of the 16S rRNA gene deposited sequences (about 1,304,069 at the time of the completion of this contribution; SILVA release 104 (www.arb-silva.de)) correspond to hitherto uncultured organisms, and just an insignificant minority can be attributed to cultured organisms (about 30,999) (Pruesse *et al*., 2007).

Since 1976, the Bacteriological Code (1975 Revision) (Lapage *et al*., 1976) has been watching for a correct nomenclature recommending a series of rules to follow, which finally will lead to the valid publication of a name. The valid publication of a given name constitutes a form of official registration/indexing that guarantees the uniqueness of the species and the operationallity of the classification system (Tindall *et al*., 2006). One of the most important recommendations made by the Bacteriological Code (1975 Revision) was the designation of a strain that will be representative of the new taxon for any taxonomic purpose, the type strain of the species. Recommendation 30a of the Bacteriological Code (1975 Revision) (Lapage *et al*., 1976) and Recommendation 30a of the Bacteriological Code (1990 Revision) (Lapage *et al*., 1992) state that a culture of the type strain (or, if the species is uncultivable, type material, a photograph, or an illustration) should be deposited in at least one of the permanently established culture collections from which it would be readily available. As of 1st January 2001 the new Rule 30 (De Vos & Trüper, 2000) states that the description of new species, or new combinations previously represented by viable cultures must include the designation of a type strain, and a viable culture of that strain must be deposited in at least two publicly accessible service collections in different countries from which subcultures must be available. The designations allotted to the strain by the culture collections should be quoted in the published description. Evidence must be presented that the cultures are present, viable, and available at the time of publication. Until the mandatory deposition of the type material was effective, the nearly two centuries of microbial taxonomy had produced a classification of 3,153 (non redundant) species whose type strains were not always deposited in culture collections, and thus with difficult access.

The establishment of a classification schema based on 16S rRNA gene genealogical relationships has been an important advance in taxonomy promoting a stable and operative framework. However, the hitherto classified 8,602 (IJSEM - February 2010) species with validly published names contrast with the more than 1,300,000 sequences deposited in the database. The “All species Living-Tree Project” (LTP; (Yarza *et al*., 2008)) intends to provide to the scientific community with a curated database, just made up by high quality sequences, corresponding to the type strains of the species with validly published names. In addition, the LTP provided a high quality alignment in ARB format (8,029 sequences) that may serve as universal reference for any reconstruction directed to the identification of putative new species. In principle, one sequence was selected to represent a type species as the paralogues present in a given genome generally do not diverge more than 1% (Yarza *et al*., 2010). However, in the rare cases in where paralogues may be significantly different, those should appear in the catalogue.

Currently, the best framework for taxonomic purposes will be achieved with the availability of a reliable database in where all type strains are represented by their 16S rRNA gene sequenced in high quality. However, in the first release of the LTP (Yarza *et al*., 2008) just about 91% of the species accounted with a sequence in the database. About 9% of the classified species remained excluded from the database as they did not account with a sequence in the public repositories (362 species), or the available sequences were of a quality below the threshold of the project (276 species). For all the species that no sequence was available we considered to be orphan (Yarza *et al*., 2008).

In 2005 the scientific community expressed the need to accomplish the generation of the full genome sequences of all classified type strains (Buckley & Roberts, 2005). This achievement will surely cause a major breakthrough in many fields of microbiology, and in especial in taxonomy and systematics of prokaryotes. Up to now, there are a series of combined efforts focusing full sequencing programs of type strains (Kyrpides *et al*., submitted). However, in the light of the current progress of the genome database increase, in where most of the organisms sequenced do not respond to the type strains of the species they belong (Kyrpides *et al*., submitted; Wu *et al*., 2009; Richter & Rosselló-Móra, 2009), seems that the ambitious goal is still far to be achieved. Of course, sequencing full genomes will ensure the identification of the 16S rRNA sequences for phylogenetic reconstruction purposes. However, with the current tempo of genome sequencing programs seems to recommend alternative approaches to accomplish a full 16S database for taxonomic purposes.

The accomplishment of a database that contains a reference sequence for all species with a validly published name had been the goal of the team undersigning this contribution. To this purposes the strain collections DSMZ, LMG, ATCC, CIP, NBRC, CECT, JCM, NCCB, ICMP, BZF, CDC, together with the LTP team had revisited the list of orphan species in order to finally sequence the 16S rRNA gene of each missing type strain. Among the 564 species without an entry in the public repositories, there is a set of 115 species (Table S3.1) lacking a culture of their type strain because these organisms have been described as uncultivable (e.g. *Borrelia recurrentis*, *Anaplasma bovis*, *Mycobacterium leprae*, *Polyangium luteum*). These names were validly published under Rule 18a of the Bacteriological Code (1975 Revision) (Lapage *et al*., 1976) and Rule 18a of the Bacteriological Code (1990 Revision) (Lapage *et al*., 1992). Since December 2000 (date of publication of the new Rule 30), species not cultivated or that can not be sustained in culture for more than a few serial passages must be proposed in the category *Candidatus* (Murray & Schleifer, 1994; Murray & Stackebrandt, 1995; Labeda, 1997a; Labeda, 1977b).

On the other hand, we could recognize a list of 59 species (Table S3.2) which type strain had never been deposited in a recognized culture collection or the culture had been lost. The fact that some cultivable species lack type culture, and thus the reference for any taxonomic work, reinforces the request for deposition in public repositories (Euzéby &

Tindall, 2004). In such cases, it would be desirable to designate a neotype strain, as recommended by the rule 18C of the Bacteriological Code (Lapage *et al*., 1992) that closely matches the original description of the species, and that is as well deposited in two international collections. A description without type material prevents the correct classification of close relatives and the identification of new members of the species is not guaranteed.

Finally, the list of species with one or more cultures available in public repositories accounted for 390 (Table S3.3) type strains. At the day of completion of this thesis manuscript 275 type strains have been already sequenced and 115 are still in progress. The joint effort among all collections produced sequences of adequate quality. Direct PCR amplification and sequencing from the cell extracts was the technique of choice. However, in those cases where there was evidence for multiple and divergent paralogs, a previous cloning step was used to discriminate the copies of the SSU. All sequences were submitted (Table S3.3) to INSDC databases (www.insdc.org). In order to simplify the LTP database and to avoid confusion when an organism showed divergent 16S paralogues with less than 2% of identity, just one copy was selected. As studied before, over 98% of the type strains contain paralogues with less than 2% divergence (Yarza *et al*., 2010). Not surprisingly, most of the new sequenced organisms affiliated with the expected relatives. However, in some cases we got evidence that some species had been inadequately placed in a given taxon (Table 3.1; Fig. S3.1; Fig. S3.2; Fig. S3.3; Fig. S3.4). We could detect cases as *Flavobacterium thermophilum* (Loginova & Egorova, 1982) that affiliated with the *Firmicutes* instead of *Bacteroidetes; Rugamonas rubra* (Austin & Moss, 1987) that affiliated with the *Betaproteobacteria* instead of *Gammaproteobacteria; Vampirovibrio chlorellavorus* (Gromov & Mamkayeva, 1980) which is clearly positioned far away from *Proteobacteria* with an unstable rooting along the domain *Bacteria*; or *Anaerorhabdus furcosa* (Veillon & Zuber 1898) (Shah & Collins, 1986) affiliating among *Firmicutes* instead of *Bacteroidetes.* In all such cases, further investigations must be carried out to determine whether they might have been placed in a wrong taxon (i.e. probably based on the sparse information gathered at the time of their classification) and thus reclassification might be recommended, or just contamination issues led to sequencing a wrong strain.

3.2 Final remarks

The identification of the orphan species after the LTP construction, led to recognize eighteen phyla with incomplete sequence catalogue (Yarza *et al*., 2010). In addition to these cases, we could as well identify that among the orphans there were still three species not classified within any phylum, hence considered as unclassified *Bacteria* (i.e. *Stibiobacter senarmontii*, *Bactoderma alba* and *Bactoderma rosea*). From the total 564 orphan species, we have accomplished the sequencing of 275 type strains and expect to achieve soon the remaining complete list of 390 reachable strains. The expected outcome of the SOS initiative is embraces sixteen bacterial and archaeal phyla and is summarized in Figure 1. Among them, *Proteobacteria* (111 sequenced, 53 in progress), *Firmicutes* (62 sequenced, 16 in progress), *Actinobacteria* (36 sequenced, 12 in progress), *Bacteroidetes* (19 sequenced, 10 in progress), *Euryarchaeota* (16 sequenced, 9 in progress) or *Tenericutes* (12 sequenced, 5 in progress). In this regard, the phyla *Synergistetes*, *Fibrobacteres*, *Thermotogae* and *Fusobacteria* will appear to be fully represented in the public repositories for the first time. However, given the unavailability of 174 type strains in the public collections, fourteen phyla will remain incomplete in the SSU databases. For example, none of the type strains that can be sequenced belong to the phyla *Chloroflexi* or *Planctomycetes.* On the other hand, *Spirochaetes*, where nearly 45% of their described species were considered orphan, only 11 type strains are being accessed and sequenced, whereas the remaining 31 species (e.g. *Borrelia recurrentis*, *Spirochaeta plicatilis*, *Treponema pallidum*) can not be cultured or found. In addition to *Spirochaetes,* ten additional phyla will never be completed due to the presence of uncultivable type strains (Table S3.1). For pragmatic reasons, the 59 species listed in Table S3.2 that lack an available type culture, should be considered. For such species it is recommendable to designate a neotype or to request to the ICSP for the invalidation of the names. In addition, it is as well recommendable that further investigation is performed on the species listed in Table 1 whose sequencing revealed misplacement in the taxonomic classification, in order to determine whether an emendation that will consolidate their classification should be proposed. At this point of time we can consider that it is possible to accomplish, for the first ever, the complete catalogue of 16S rRNA gene sequences of all the validly named species with an available type strain in public collections.

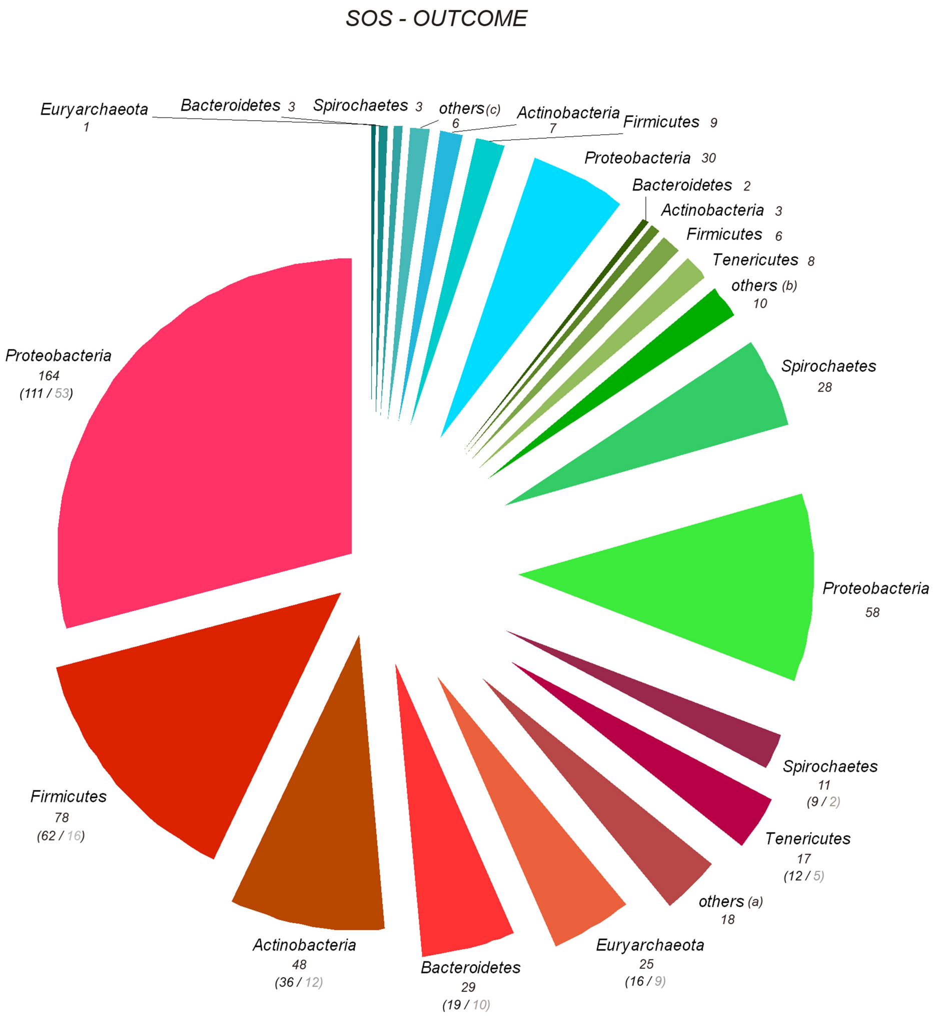


Fig. 3.1. 564 orphan species without a good quality SSU entry in public databases were considered in our sequencing schedule. Among them, 360 type strains (reds) are either properly sequenced (black, in brackets) or in progress (grey, in brackets), whereas 174 could not be achieved. 115 consisted on non-cultivable strains (greens) and 59 could not be found in any recognized culture collection (blues).

(a) Others (18 species): *Crenarchaeota* (4), *Fusobacteria* (3), *Thermotogae* (3), *Chlorobi* (3), *Deinococcus-Thermus* (1), *Fibrobacteres* (1), *Synergistetes* (1), *Nitrospira* (1) and *Bactoderma rosea* (unclassified Bacteria).

(b) Others (10 species): *Chloroflexi* (3), *Plantomycetes* (3), *Chlorobi* (1), *Cyanobacteria* (1), *Crenarchaeota* (1) and *Bactoderma alba* (unclassified Bacteria).

(c) Others (6 species): *Crenarchaeota* (2), *Nitrospira* (1), *Chloroflexi* (1), *Deinococcus-Thermus* (1) and *Stibiobacter senarmontii* (unclassified Bacteria).

Table 3.1 Sequenced type strains with unexpected affiliations

Organism name	Type strain	Expected affiliation	Observed affiliation	Notes
Flavobacterium oceanosedimentum	ATCC 31317	*Flavobacteriaceae* (*Bacteroidetes*)	*Microbacteriaceae* (*Actinonacteria*)	99.8% similarity against *Curtobacterium citreum*
Lactobacillus rogosae	ATCC 27753	*Lactobacillaceae* (*Bacilli*)	*Lachnospiraceae* (*Clostridia*)	94.8% similarity against *Lachnospira multipara* (Fig. S1)
Gemmiger formicilis	ATCC 27749	*Hypomicrobiaceae* (*Proteobacteria*)	*Ruminococcaceae* (*Firmicutes*)	97.5% similarity against *Subdoligranulum variabile* (Fig. S1)
Acetoanaerobium noterae	ATCC 35199	*Ruminococcaceae* (*Clostridia*)	*Peptostreptococcaceae* (*Clostridia*)	88.7% similarity against *Filifactor villosus*. (Fig. S2)
Vitreoscilla filiformis	ATCC 43190	*Neisseriaceae* (*Neisseriales*)	*Comamonadaceae* (*Burkholderiales*)	96% similarity against *Roseateles depolymerans*
Rugamonas rubra	ATCC 43154	*Pseudomonadaceae* (*Gammaproteobacteria*)	*Oxalobacteraceae* (*Betaproteobacteria*)	97.9% similarity against *Duganella zoogloeoides*
Anaerorhabdus furcosa	ATCC 25662	*Bacteroidaceae* (*Bacteroidetes*)	*Erysipelotrichaceae* (*Tenericutes*)	91.6% similarity against *Holdemania filiformis*
Vampirovibrio chlorellavorus	ATCC 29753	*Bdellovibrionaceae* (*Proteobacteria*)	unstable rooting	77% similarity against *Bdellovibrio bacteriovorus*
Meniscus glaucopis	ATCC 29398	*Cytophagaceae* (*Sphingobacteriales*)	deep branch in *Bacteroidales*	80% similarity against *Cytophaga hutchinsonii* (Fig. S3)
Spiroplasma ixodetis	ATCC 33835	*Spiroplasmataceae* (*Entomoplasmatales*)	deep branch in *Entomoplasmatales*	86% similarity against *Spiroplasma citri* (Fig. S4)
Spiroplasma platyhelix	ATCC 51748	*Spiroplasmataceae* (*Entomoplasmatales*)	deep branch in *Entomoplasmatales*	85.6% similarity against *Spiroplasma citri* (Fig. S4)
Thermothrix azorensis	ATCC 51754	*Burkholderiaceae* (*Proteobacteria*)	*Aquificaceae* (*Aquificae*)	99.7% similarity against *Hydrogenobacter subterraneus*
Flavobacterium thermophilum	CCUG 22402	*Flavobacteriaceae* (*Flavobacteria*)	*Bacillales* (*Firmicutes*)	99.7% similarity against *Anoxybacillus kamchatkensis*
Neisseria flava	CCUG 24961	*Neisseriaceae* (*Betaproteobacteria*)	*Moraxellaceae* (*Gammaproteobacteria*)	99.9% similarity against *Moraxella caviae*
Bacteroides xylanolyticus	CCUG 48289	*Bacteroidaceae* (*Bacteroidetes*)	*Lachnospiraceae* (*Firmicutes*)	98.5% similarity against *Clostridium aerotolerans* (Fig. S1)
Nocardia globerula	DSM 44596	*Nocardia* (*Nocardiaceae*)	*Rhodococcus* (*Nocardiaceae*)	99.9% similarity against *Rhodococcus globerulus*
Acetivibrio ethanolgignens	DSM 3005	*Ruminococcaceae* (*Clostridia*)	*Lachnospiraceae* (*Clostridia*)	82% similarity against *Ruminococcus flavefaciens* (Fig. S1)
Acetivibrio multivorans	DSM 6139	*Ruminococcaceae* (*Clostridia*)	*Clostridiaceae* (*Clostridia*)	93.8% similarity against *Clostridium cavendishii*; 84.7% against *Acetivibrio cellulolyticus*
Thermohydrogenium kirishiense	DSM 11055	*Syntrophomonadaceae* (*Clostridia*)	deep branch in *Clostridia*	98.3% similarity against *Thermoanaerobacterium aotearoense*
Acetomicrobium faecale	DSM 20678	*Bacteroidaceae* (*Bacteroidetes*)	deep branch in *Clostridia*	99.46% similarity against *Caldicoprobacter oshimai*
Dichotomicrobium thermohalophilum	DSM 5002	*Hyphomicrobiaceae* (*Alphaproteobacteria*)	deep branch in *Alphaproteobacteria*	93.77% similarity against *Afifella pfennigii*
Aquabacter spiritensis	DSM 9035	*Hyphomicrobiaceae* (*Alphaproteobacteria*)	*Xanthobacteraceae* (*Alphaproteobacteria*)	97.6% similarity against *Xanthobacter tegetidis*

Cytophaga xylanolytica	DSM 6779	*Cytophagaceae* (*Sphingobacteria*)	deep branch in *Bacteroidales*	80.3% similarity against *Cytophaga hutchinsonii* (Fig. S3)
Sphingobacterium antarcticum	DSM 15311	*Sphingobacterium* (*Sphingobacteriaceae*)	*Pedobacter* (*Sphingobacteriaceae*)	99.9% similarity against *Pedobacter piscium*
Nocardia coeliaca	DSM 44595	*Nocardia* (*Nocardiaceae*)	*Rhodococcus* (*Nocardiaceae*)	99.9% similarity against *Rhodococcus erythropolis*
Ilyobacter delafieldii	DSM 5704	*Fusobacteriaceae* (*Fusobacteria*)	*Clostridiaceae* (*Firmicutes*)	98.8% similarity against *Clostridium homopropionicum*
Acetomicrobium flavidum	DSM 20664	*Bacteroidaceae* (*Bacteroidetes*)	deep branch in domain *Bacteria*	71% similarity against *Bacteroides fragilis*. 99.7% against *Anaerobaculum mobile*
Flavobacterium acidificum	LMG 8364	*Flavobacteriaceae* (*Bacteroidetes*)	*Enterobacteriaceae* (*Proteobacteria*)	99.9% similarity against *Pantoea ananatis*
Zavarzinia compransoris	LMG 5821	*Acetobacteraceae* (*Alphaproteobacteria*)	deep branch in *Alphaproteobacteria*	90.9% similarity against *Oceanibaculum indicum*. 86.8% against *Acetobacter aceti*
Haemophilus piscium	CIP 106116	*Pasteurellaceae* (*Gammaproteobacteria*)	*Aeromonadaceae* (*Gammaproteobacteria*)	99.8% similarity against *Aeromonas bestiarum*
Carbophilus carboxidus	CIP 105722	*Rhizobiaceae* (*Alphaproteobacteria*)	*Phyllobacteriaceae* (*Alphaproteobacteria*)	99.9% similarity against *Aminobacter lissarensis*
Aquaspirillum polymorphum	NBRC 13961	*Neisseriaceae* (*Betaproteobacteria*)	deep branch in *Alphaproteobacteria*	98% similarity against *Magnetospirillum gryphiswaldense*

Chapter 4

Empirical circumscription of prokaryotic higher taxa based on comparative analyses of the 16S rRNA gene.

Manuscript in preparation

The current accepted classification of prokaryotes is built up with more than 8,600 species with validly published names. It shows a bias in the distribution of taxa representing distinct taxonomic ranks, given by the difficulties related to culturing some kinds of organisms, together with the different interest in being studied. 71% of the phyla are classified with just one class, 59% of the classes with one order, 56% of orders with one family, 34% of families with one genus and 50% of genera with one species. Additionally, the number of taxa classified with just one species is notably high: 2 phyla, 7 classes, 18 orders, 46 families and 879 genera. During the course of the last 50 years of classification, taxonomists empirically circumscribed taxa following coherent and comparable criteria. In this regard, we have calculated that higher taxonomic categories could be circumscribed by means of sequence similarity thresholds: 94.5% for genus, 86.5% for family, 82% for order, 78.5% for class and 75% for phylum. The sparse representation of validly published higher taxa contrasts with the enormous uncultured diversity reported. By applying the empirical taxonomic thresholds outlined above we have developed a rationale for the classification of candidate taxa within clades of uncultured organisms. The candidate division OP11 was used as model case to apply these thresholds. In this highly diverse candidate division we found evidence for 13 candidate phyla, 29 candidate classes, 54 candidate orders, 122 candidate families and 215 candidate genera.

4.1 Introduction

The discovery of a new prokaryotic organism is followed by a series of steps intended to clarify its placement in the taxonomic classification, in the following order: 1. characterization, 2. classification and, 3. nomenclature (Tindall *et al.*, 2010). The same steps should be considered when revising the taxonomic status of previously described taxa. To assist the researchers along this process, the scientific community accepted an official compilation of principles, rules and recommendations for the nomenclature of prokaryotes gathered in the International Code of Nomenclature of Bacteria (1990 Revision) (Lapage *et al.,* 1992), hereafter Bacteriological Code. The Bacteriological Code built up a hierarchical structure for the naming of prokaryotes, accepting the following categories ordered from the lowest to the highest hierarchies: subspecies, species, subgenus, genus, subtribe, tribe, subfamily, family, suborder, order, subclass and class. Nevertheless, some of them have never been used at all or have fallen into disuse (e.g. subgenus, subtribe, tribe and subfamily). On the other hand, two additional categories above class were established independently from the code's rules, based on 16S rRNA genealogical models (i.e. phylum and domain) (Woese, 1987; Woese *et al.,* 1990). As stated by rules 16-22, all validly published taxa need to have a nomenclatural type designated (a type strain for species and subspecies, a type species for genera, a type genus for families and orders, and a type order for classes and above). These types (always type strains in last place) are the references that must be used to compare against the new isolate in order to measure their relatedness and to discard its assignment to preexisting taxa, hence proving the novelty of a taxon. Just by definition, a prokaryotic species must be included into a genus, and a genus belongs theoretically to a series of successive higher ranks. For example the publication of a new family (e.g. *Caldicoprobacteraceae,* (Yokoyama *et al*. 2010)) entails the publication of a new species and a new genus, by characterizing a number of strains and discarding their assignment to the closest relative type species and type genera. Only comprehensive characterizations based on genetic, genomic and phenotypic traits may lead to unambiguous circumscriptions, guaranteeing a stable classification and a valid nomenclature. As for characterization, there is not an official classification of prokaryotes and both areas remain subjected to current opinion and general agreement (Rosselló-Móra & Amann, 2001; Tindall *et al.,* 2010; Garrity, 2001).

When Carl Woese and coworkers demonstrated in year 1977 the usefulness of the RNA sequence of the small subunit (SSU) of the ribosome as universal phylogenetic marker (Fox *et al.,* 1977), the comparative analysis of this molecule at the primary structure level constituted a breakthrough for microbial taxonomy. 25 years later, it has become the gold standard for reconstructing genealogies and the backbone for a new prokaryotic taxonomy (Ludwig & Klenk, 2001). Currently, the most widely accepted classification scheme is the Taxonomic Outline of the Prokaryotes covered by the Bergey's Manual of Systematic Bacteriology (Garrity, 2001). Here, a SSU-directed classification of prokaryotes has been superimposed to the hierarchical framework provided by the rules of nomenclature (Garrity, 2001). Motivated by microbial taxonomy (Stackebrandt *et al.,* 2002) and microbial ecology (Amann *et al*., 1995), the number of sequence submissions to public databases of the International Nucleotide Sequence Database Collaboration (INSDC; www.insdc.org) started to grow exponentially since early nineties, currently exceeding the number of 2,900,000 entries (http://www.arb-silva.de/documentation/background/release-104). This immense quantity of sequence data generated have promoted the development of software platforms and dedicated databases. The sequences accessible in the INSDC appear in their raw state. However, in order to adequately analyze the sequence similarities to infer genealogies the prerequisite is to establish a reliable alignment for each gene. The basis to set up an alignment relies on recognizing positional orthology of each single base or amino acid residue. In this process homologous characters are positioned in columns during the calculation of phylogenetic trees. The importance to generate a highly curated alignment prior to any analysis was remarked by Wolfgang Ludwig and coworkers during early nineties (Ludwig & Schleifer, 1994). In particular for rRNA sequences, the primary structure in the alignment could be better evaluated and improved by checking for potential secondary structure formation, hence existing helices and loops that would be differently considered by functional and evolutionary constraints. The same group developed the ARB software package for sequence data handling (www.arb-home.de, (Ludwig, *et al.,* 2004)), one of the most relevant tools for complete phylogenetic reconstructions. In addition, three independent surveys in Europe, United States and Australia, SILVA (www.arb-silva.de, (Pruesse *et al*., 2007)), RDP (http://rdp.cme.msu.edu, (Cole *et al.,* 2007)) and Greengenes (http://greengenes.lbl.gov, (DeSantis *et al.,* 2006)), emerged with the aim of: (i) provide the scientific community with updated universal alignments in order to achieve optimal and comparable phylogenetic reconstructions; (ii) produce and maintain curated datasets of nearly full length rRNA sequences to be used for in depth phylogenetic analyses; and (iii) develop a set of bioinformatic tools for on-line sequence data management and analyses.

About 20 years ago, 16S rRNA sequences started to appear in new species descriptions as additional taxonomic information. Due to its common use and its usefulness in recognizing monophyly, it has been suggested that a full length SSU sequence of at least the type strain of the new taxon should be mandatory for a valid species classification (Stackebrandt *et al.,* 2002). Actually, one of the major premises within a polyphasic approach is to prove monophyly of the members of the new taxon by analyzing their 16S rRNA gene (Rosselló-Móra & Amann, 2001, Tindall *et al.,* 2010). Within this framework, three years ago we started the All-Species Living Tree Project (LTP, (Yarza *et al.,* 2008)) as a common initiative between the journal Systematic and Applied Microbiology and the ARB (www.arb-home.de), SILVA (www.arb-silva.de) and LPSN (List of Prokaryotic names with Standing in Nomenclature, www.bacterio.cict.fr) teams. Our goal was to prepare a set of materials intended to facilitate the tasks of microbial taxonomists during the steps of sequence selection and alignment improvement, in order to accommodate new taxa within the existing species classification. Briefly, the aim of the project consisted on to: (i) provide a curated SSU rRNA database of all type strains for which a sequence was available; (ii) maintain an optimized and universally usable alignment; and (iii) reconstruct a single tree harboring a reliable genealogy of all classified species with validly published names. During the construction of the first release, we had to deal with the often incomplete or wrong information found at the INSDC entries (i.e. organism name and/or strain numbers), in many cases hampering recognizing the correct type strain for a species. The current release (LTPs102, September 2010) contains 8,029 type strain SSU sequences, whereas 573 species with a validly published name are still missing due to the absence of entries in public repositories, or due to the bad quality of the deposited SSU sequences. Currently the LTP team leads a new international collaboration together with the culture collections in order to fill the gaps in the SSU catalog (Yarza *et al.*, in preparation). So far the four updates of the LTP have been released, and the complete package of tables, alignments, sets of sequences, ARB databases and trees can be downloaded at the project's web page (www.arb-silva.de/projects/living-tree).

One of the objectives within the LTP project was the calculation of empirical taxa boundaries based on comparative analysis of the 16S rRNA and 23S rRNA gene sequences (Yarza *et al.*, 2008; Yarza *et al.*, 2010). We observed that during the course of the last 50 years of classification, taxonomists circumscribed new genera, families or phyla following coherent and comparable criteria independently of the use of 16S rRNA gene sequences. These coherent circumscription criteria contrast with the fact that microbial systematics has never been static but constantly subjected to methodological changes. The coherence of the empirical circumscription of genera, families and phyla was previously revealed taking the improved alignment of the LTP (Yarza *et al.*, 2008). The evaluation was based on type strains and creating a distance matrix to observe the maximum distance found within each taxon. By averaging all taxa at a certain taxonomic level (i.e. genera, families and phyla) we could obtain a minimum similarity value, or lower-cutoff, that could be applied in general terms as a threshold for discarding assignment of a new taxon to the previously classified taxa. For example, we could observe that a 94.9% (with a confidence interval of 0.4) of sequence identity would constitute the boundary threshold for circumscribing a new genus. In the present work we have carried out a revision of the taxa boundaries by improving the protocol and including the remaining ranks: i.e. suborder, order, subclass and class. We have corroborated the presence of outlier-taxa at all taxonomic levels by means of their phylogenetic affiliations, and found evidences of some taxa in need of reclassification.

One of the most challenging facts for the current microbiology is that over 99% of the expected microbial diversity in the biosphere has not yet been cultivated (www.arb-silva.de). Nonetheless, despite our incapability to cultivate the majority of these organisms, we have developed tools to reveal their metabolic traits, their impact on ecosystems or even their putative classification in the taxonomic schema (Pace, 1997; Antón, *et al.,* 2000; Giovannoni *et al.,* 1990). The phylogenetic affiliation is the primary approach for the classification of any new isolate in the frame of microbial taxonomy, and the primary assignation to putative species in microbial ecology. This practice allows revealing genealogical relationships among the complete dataset of cultured and uncultured SSU sequences. The reconstruction, raises a basic taxonomic question regarding the placement of the new sequences in the taxonomic framework: is it close enough to its relatives to be considered as a member of the same genus, or family, order, class, or even the same phylum? Such question is especially fuzzy in molecular microbial ecology. For example, candidate phyla have been typically circumscribed following SSU-based phylogenetic analyses as single criterion (Mori *et al.*, 2010; Hugenholtz *et al.*, 1998). Microbial ecologists generally circumscribe their groups as Operational Taxonomic Units (OTUs; (Rosselló-Mora and López-López, 2008)) using restrictive 16S rRNA sequence similarity thresholds around 97% (Zaballos *et al.*, 2006), and assume that reflect species. Other may prefer using the Operational Phylogenetic Units (OPUs; (López-López *et al.*, 2010)), as unique sequence clades without rigid similarity thresholds, and that could be equalized to genera or even families. In any case, there is a need to understand whether higher taxa can be numerically circumscribed in terms of SSU rRNA similarity, a fact that would simplify the recognition of units for both microbial taxonomists and ecologists.

4.2 Abundance and distribution of higher taxonomic ranks in *Bacteria* and *Archaea.*

The complete list of prokaryotic species with validly published names up to the February issue if the International Journal of Systematic and Evolutionary Microbiology (JSEM) (Yarza *et al.*, 2010) was supplemented with the complete classification obtained from LPSN's “hierarchical classification of prokaryotes” (www.bacterio.cict.fr/classifphyla.fr). This dataset was the basis for the distribution analyses and threshold calculations in the present work. From the classification provided by LPSN the following sources were taken into account: (i) original articles; (ii) latest Taxonomic Outline of *Bacteria* and *Archaea* (TOBA; (Garrity *et al.,* 2007)); (iii) NCBI taxonomy (http://www.ncbi.nlm.nih.gov); (iv) Taxonomic Outlines for volumes 3 and 4 of Bergey's Manual of Systematic Bacteriology (Second Edition) (http://www.bergeys.org/outlines.html, (Garrity *et al.*, 2001)); and (v) the All-Species Living Tree Project (Yarza *et al.,* 2008). The working catalog comprised 8,602 prokaryotic species, a non redundant number in the sense that only type strains have been taken into account (i.e. a species and its first subspecies share the same type). Those were classified in 1,779 genera, 285 families, 115 orders, 52 classes, 29 phyla and 2 domains. Along this chapter we have avoided the use of the ranks suborder and subclass given the low number of such classified categories (20 and 5 respectively).

Genera. By definition, all species are classified into a genus. About 90% of the genera showed to encompass between 1 and 10 species. Also, 50% of all classified genera have been described just with a single species. Only 22 genera embraced 50 or more species and just 8 appeared with more than 100 classified species. Very large genera rarely show clear cut monophyletic clades according to ribosomal SSU and LSU markers, and often appear para- or polyphyletic. Due to this reason we can expect important reclassifications in the future for the eight largest genera *Streptomyces* (514 species), *Clostridium* (178), *Bacillus* (159 species), *Mycobacterium* (149 species), *Lactobacillus* (146 species), *Pseudomonas* (133 species), *Mycoplasma* (118 species) and *Paenibacillus* (110 species). For example, para- or polyphyletic structures have been observed in rRNA trees for *Clostridium*, *Bacillus* or *Pseudomonas*, and up to 22 species currently belonging to the genus *Mycoplasma* are closely related to the genera *Mesoplasma*, *Ureaplasma* and *Acholeplasma* (Yarza *et al.,* 2008, Yarza *et al.*, 2010).

Families. Nearly 50% of all classified families embraced between 1 and 10 genera (Fig. 4.1A). Besides, about 16% of all classified families are composed by just a single species and genus. There is no correlation between the number of genera and the number of species classified within a family, as shown by linear regression coefficients below 0.4 (data not shown). However, in general families with a low number of genera are expected to harbor a low number of species (e.g *Kineosporiaceae*, 3 genera and 14 species). But the classification exhibits plenty of exceptions, for example the most abundant family in terms of species content is *Streptomycetaceae* with 545 species, but only 3 genera. Contrarily, some families with high number of classified genera reveal a low number of species e.g. *Coriobacteriaceae* (13 genera and 24 species). However, the 10 largest families following the rule were: *Flavobacteriaceae* (85 genera), *Rhodobacteraceae* (75 genera), *Enterobacteriaceae* (46 genera), *Bacillaceae* (37 genera), *Comamonadaceae* (30 genera), *Microbacteriaceae* (30 genera), *Veillonellaceae* (30 genera), *Neisseriaceae* (30 genera), *Halobacteriaceae* (27 genera) and *Acetobacteraceae* (27 genera). During the updatings of the LTP (Yarza *et al.*, 2008, Yarza *et al*., 2010), we have observed that the clades

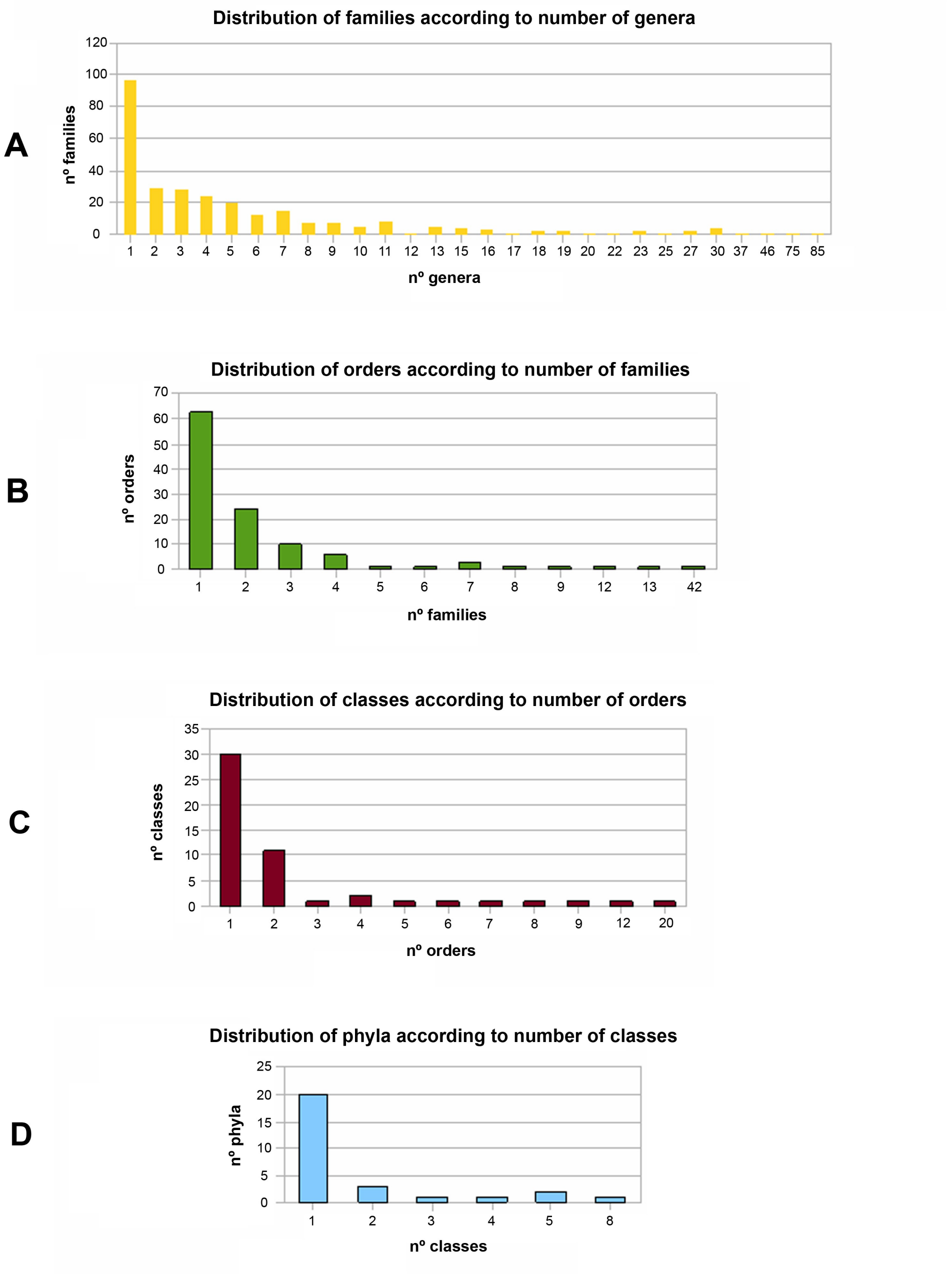


Fig 4.1. Distribution of number taxa according to its content of genera (A), families (B), orders (C) and classes (D).

representing single families were quite stable in the reconstructed trees, whereas the genera circumscription was sometimes influenced by the effects of plesiomorphy and branch attraction (Ludwig & Klenk, 2001). In most cases, families were strongly supported as monophyletic clusters unless there was a problem of misclassification (e.g. *Bacillaceae*, *Clostridiaceae*, *Eubacteriaceae*, *Rhodobacteraceae, Flammeovirgaceae*).

Orders. Half of the classified orders (63 of 115) contain currently only a single family. In general, orders usually show a number between 1 and 14 families (Fig. 4.1B). The most remarkable exception is the order *Actinomycetales* represented by 42 families and 2,063 species. However, about 16% of all orders are composed by just one species (e.g. *Methanocellales*, *Thermomicrobiales*, *Methanopyrales*), hence being single-genus and single-family taxa as well. Due to the fact that taxonomy and phylogeny developed harmoniously just recently, higher taxa above the rank of family (with the exceptions of phylum and domain) do not necessarily behave as monophyletic structures in comprehensive trees (Ludwig & Schleifer, 1994; Ludwig & Klenk, 2001; Yarza *et al.*, 2010).

Classes. About 86% of all classes (44 of 52) encompass currently between 1 and 4 orders (Fig. 4.1C) and 13% actually consist of just one species (e.g. *Caldisericia*, *Ktedonobacteria*). The largest classes corresponded to *Actinobacteria* (8 orders and 2,146 species), *Gammaproteobacteria* (20 orders, 1,434 species), *Bacilli* (3 orders, 1,032 species), *Alphaproteobacteria* (12 orders, 913 species), *Deltaproteobacteria* (9 orders, 302 species), *Betaproteobacteria* (7 orders, 451 species) and *Clostridia* (6 orders, 640 species). In comprehensive trees, classes generally clustered with higher coherence than orders, hence allowing more reliable phylogenetic affiliations.

Phyla. Contrarily to the lower taxonomic categories, a new phylum is solely defined by segregation of a new branch in a SSU-based phylogenetic reconstruction (Mori *et al.*, 2009). In spite of their size most of the phyla are formed by only one class (Fig. 4.1D) and just eight reveal two or more classes (e.g. *Bacteroidetes*, *Euryarchaeota*). “Actinobacteria” phylum was the second largest in the catalog in terms of species content (2,146), however it embraces only one class (*Actinobacteria*). It looks like if a fuzzy phylogenetic boundary exists between the levels class and phylum and 16S rRNA tree reconstructions evidence a tendency to segregate classes of some bacterial phyla according to the low significance of their branches. For example in “Proteobacteria”, *Epsilonproteobacteria* and *Deltaproteobacteria* tend to move away from their counterparts while *Betaproteobacteria* and *Gammaproteobacteria* constitute a group with high significance and clearly separated from *Alphaproteobacteria*. A similiar situation is found within the “Firmicutes” whose classes *Bacilli* and *Clostridia* cannot clearly grouped together (Yarza *et al.,* 2010) (Fig. 4.2).

Unclassified taxa. Quite often, descriptions of new taxa do not explicitly indicate classification at their successive higher taxonomic ranks. Indeed, at the date of compilation of our database (February 2010), 207 species remained without a classification at the family level (i.e. whose genera were not classified into a family), 33 species without an order-level classification, 16 species not assigned to a class and 3 species not assigned to any phylum. Many of these “unclassified” taxa correspond to “orphan” species: they have never been sequenced or their SSU sequences were of too low quality to be included in the LTP. For the rest, being represented in the LTP, many could indeed be affiliated to a family, order, class, or phylum with high confidence (e.g. *Spongiispira norvegica* (Kaesler *et al.,* 2008) (Fig. S4.1); Genus *Tepidimonas* (Moreira *et al.,* 2000) (Fig. S4.2); *Solimonas soli* (Kim *et al*., 2007) (Fig. S4.3)*; Thiohalobacter thiocyanaticus* (Sorokin *et al*., 2010) (Fig. S5); However, in some cases the 16S rRNA gene sequences led to ambiguous or deep phylogenetic placements (Yarza et al., 2008, Yarza et al., 2010), thus not providing sharp enough phylogenetic affiliations. As examples, unresolved affiliations at the family level occurred for *Sedimenticola* (1 species)*, Alkalimonas* (3 species)*, Salicola* (2 species)*, Spongiibacter* (2 species) *of the class Gammaproteobacteria, or Methanocalculus* (4 species)*, Methanolinea* (1 species) *of the class Methanomicrobia.*

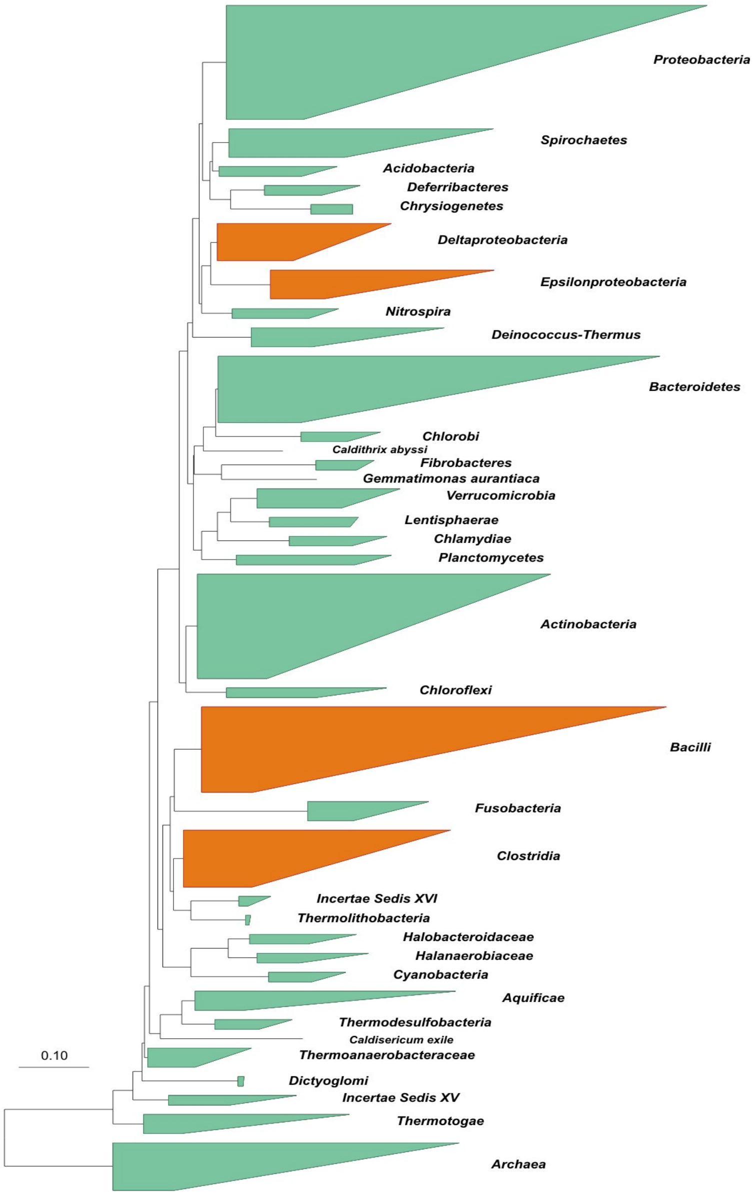


Fig. 4.2. Collapsed overview of the All-Species Living Tree, version LTPs102 (Yarza et al., 2010). Scale bar indicates substitutions per site.

4.3 Fine-tuning taxa boundaries*.*

Finding some numerical recommendations on the putative taxa boundaries, may be useful to simplify the classification process of new taxa, and also to recognize equivalent categories for uncultured organisms. One of the disadvantages of the previous calculations of boundaries (Yarza *et al.,* 2008) was that they took into account only similarity measures against the type member of a given taxon. As a consequence, we disregarded many taxa because their type members were "orphan" (i.e. without sequence in the LTP). Additionally, the earlier calculation of the family and phylum boundaries was based on a manual selection of 28 and 10 clear cut families and phyla represented by only 202 and 195 species, respectively. We now have applied now a new approach that takes into account all similarity values that can be found within a taxon, independently of the presence of the nomenclatural type. To calculate the boundaries, the complete alignment of the LTP, composed of 8,029 sequences each one belonging to a type strain of a classified species up to February 2010 issue of the IJSEM, was used to create a single distance matrix. As each one of the sequences in the LTP database had been complemented with the taxonomy provided by LPSN (Yarza *et al.,* 2010) we got a fully endowed source file with the sequence identity data. A new routine was programmed to conduct the following main operations: (1) it generated a list of all distinct taxa found in the whole matrix at a user-selected taxonomic rank (genus, family, suborder, order, subclass, class or phylum); (2) the name of each taxon was used to match all its members and to generate a sub-matrix; (3) taxa with less than three members were discarded from the study as they could only provide one measure; (4) a list of measures was obtained from an all-against-all comparison (excluding reciprocal matches); (5) two distinct corrections were applied for the treatment of outliers, one restrictive (p50-1.5SD) and another permissive correction (p50-2.5SD), each one producing a list of outliers for further manual inspection (Table 4.1); (6) for each taxon, the descriptors maximum value, minimum, mean, SD and median were obtained; and, (7) all computed taxa at a given taxonomic rank were returned in separated rows in the main output file. Finally, median and its confidence interval were the descriptors of choice to represent the taxa by single and meaningful values.

Table 4.1. Outliers removed from the dataset show incongruent phylogenetic affiliations.

Organism name	Type strain	Expected affiliation	Observed affiliation	Notes
Aeromonas sharmana	GPTSA-6	*Aeromonas* (*Aeromonadaceae*)	Deep branch in *Aeromonadaceae*	Clearly separated from the genus *Aeromonas* (Fig. S5)
Arthrobacter viscosus	LMG 16473	*Micrococcaceae* (*Actinobacteria*)	*Rhizobiaceae* (*Proteobacteria*)	(Heyrman *et al.,* 2005) (Fig. S6)
Azospirillum irakense	KBC1	*Azospirillum (Rhodospirillaceae)*	*Rhodocista (Rhodospirillaceae)*	Clearly separated from the genus *Azospirillum and closely related to Rhodocista.* (Fig. S7)
Campylobacter ureolyticus [1]	ATCC 33387	*Bacteroidaceae* (*Bacteroidetes*)	*Compylobacteraceae* (*Proteobacteria*)	(Vandamme *et al.,* 2010)
Flavobacterium mizutaii	DSM 11724	*Flavobacteria* (*Bacteroidetes*)	*Sphingobacteria* (*Bacteroidetes*)	(Gherna & Woese, 1992) Closely related to *Sphingobacterium* (Fig. S8)
Leuconostoc fallax	DSM 20189	*Leuconostoc* (*Leuconostocaceae*)	Deep branch in *Leuconostocaceae*	Clearly separated from genus *Leuconostoc* (Fig. S9) (Antunes et al., 2002)
Moraxella boevrei	ATCC 700022	*Moraxella* (*Moraxellaceae*)	Deep branch in *Moraxellaceae*	Clearly separated from the genus *Moraxella* (Fig. S10)
Pseudomonas beteli	ATCC 19861	*Pseudomonadales* (*Gammaproteobacteria*)	*Xanthomonadales* (*Gammaproteobacteria*)	(Van Den Mooter & Swings, 1990)
Pseudomonas boreopolis	ATCC 33662	*Pseudomonadales* (*Gammaproteobacteria*)	*Xanthomonadales* (*Gammaproteobacteria*)	(Anzai *et al.,* 2000)
Pseudomonas cissicola	ATCC 33616	*Pseudomonadales* (*Gammaproteobacteria*)	*Xanthomonadales* (*Gammaproteobacteria*)	(Hu et *al.,* 1997)
Pseudomonas flectens	ATCC 12775	*Pseudomonadales* (*Gammaproteobacteria*)	*Enterobacteriaceae* (*Gammaproteobacteria*)	(Anzai *et al.,* 2000)
Pseudomonas geniculata	ATCC 19374	*Pseudomonadales* (*Gammaproteobacteria*)	*Xanthomonadales* (*Gammaproteobacteria*)	(Anzai *et al.,* 2000)
Pseudomonas hibiscicola	ATCC 19867	*Pseudomonadales* (*Gammaproteobacteria*)	*Xanthomonadales* (*Gammaproteobacteria*)	(Van Den Mooter & Swings, 1990)
Pseudomonas pictorum	LMG 981	*Pseudomonadales* (*Gammaproteobacteria*)	*Xanthomonadales* (*Gammaproteobacteria*)	(Anzai *et al.,* 2000)
Rhizobium lupini	DSM 30140	*Rhizobiaceae*	*Bradyrhizobiaceae*	(Ludwig *et al.,* 1995)

The outliers' correction performed by the program was just a control point to exclude species that did not follow common trends, or with dubious classification, thus removing noise from the dataset. These examples correspond to some type strains with wrong phylogenetic affiliations that were automatically rejected. Phylogenetic diagnosis was made using the maximum likelihood reconstruction of the All-Species Living Tree (release LTPs102. Yarza *et al.,* 2010).

[1] Previously known as *Bacteroides ureolyticus.*

The taxonomic boundaries

Median and its 95% confidence interval (CI) for the minimum and median similarity of all taxa at each taxonomic rank is shown in Table 4.2A. In general, 94.5% is the minimum similarity between two ribosomal SSU sequences that guarantees the circumscription of a new genus, 86.5% a new family, 82% a new order, 78.5% a new class and 75% a new phylum. These re-calculated values are in accordance with previously reported boundaries at the genus, family and phylum levels (Yarza *et al.,* 2008; Cole *et al.,* 2010). The maximum similarity found within the genera was of 98.95 (±0.1), where a possible species boundary would be, in accordance with the previously proposed values of 98.7-99% (Stackebrandt & Ebers, 2006). The lowest identity found between a pair of prokaryotic SSU rRNA sequences was of 55.45% and corresponded to the pair *Haloquadratum walsbyi* (*Archaea*) – *Mycoplasma penetrans* (*Bacteria*), represented by accession numbers AY676200 and L10839 respectively. On the other hand, three of the 850 genera of *Bacteria* and *Archaea* harboring two or more species, did not show a range of sequence similarity. The nine species of the genus *Brucella* (using the "six species" concept for *B. melitensis*, *B. abortus*, *B. canis*, *B. neotomae*, *B.suis* and *B.ovis* (Osterman & Moriyón, 2006)), represented by AM158979, L37584, AM158982, AY594215, AM392286, AY594216, L26168, AM158981 and AM158980, showed 100% identical ribosomal SSU genes. The same occurred with the two species of the genus *Caldimonas* (*C. manganoxidans*, *C. taiwanensis*) represented by AB008801 and AY845052, and the three species of the genus *Stigmtella* (*S. aurantiaca*, *S. erecta, S. hybrida*) represented by DQ768127, AJ970180 and DQ768129. In general, the reduced confidence intervals (CIs) observed (Table 4.2A) indicated that there has been a harmonious practice when creating taxonomic categories. The fact that CIs turn broader as higher we move in the taxonomic hierarchy can be interpreted as a consequence of the lower number of taxa at those categories. Ranks suborder and subclass appeared with the least reliable cutoff values, indicated by wide CIs and great shortage of taxa. Additionally, their lower-cutoff values of ~86.5% (suborder) and ~82% (subclass) indicated a full overlap with the categories family and order, respectively. These results indicate that categories suborder and subclass might be introducing redundancy in the taxonomic schema, as they might be equalized to other existing taxonomic ranks.

Table 4.2. Prokaryotic taxa boundaries.

A. Bacteria + Archaea

	Genus		Family		Suborder		Order		Subclass		Class		Phylum	
No. Taxa	568		201		18		85		4		39		23	
No. Species	6644		7749		2067		7960		2086		7999		8018	
Median [a]	96.4	(96.2, 96.55)	92.25	(91.65, 92.9)	92.6	(90.9, 94.1)	89.2	(88.25, 90.1)	88.54	(87.45, 89.4) [b]	86.35	(84.7, 87.95)	83.68	(81.6, 85.93)
Minimum	94.8	(94.55, 95.05)	87.65	(86.8, 88.4)	88.5	(86.8, 90.35)	83.55	(82.25, 84.8)	83.53	(82, 87.1) [b]	80.38	(78.55, 82.5)	77.43	(74.95, 79.9)
CUTOFF	**94.5%**		**86.5%**		**86.5%**		**82%**		**82%**		**78.5%**		**75%**	

B. Bacteria

	Genus		Family		Suborder		Order		Subclass		Class		Phylum	
No. Taxa	539		188		18		73		4		31		21	
Median [a]	96.38	(96.2, 96.55)	92.15	(91.48, 92.8)	92.6	(90.9, 94.1)	88.7	(87.62, 89.7)	88.54	(87.45, 89.4) [b]	85.19	(83.7, 87.08)	83.8	(81.6, 86.38)
Minimum	94.75	(94.5, 95)	87.4	(86.55, 88.25)	88.5	(86.8, 90.35)	82.85	(81.45, 84.15)	83.53	(82, 87.1) [b]	79.3	(77.4, 81.2)	77.6	(75.05, 80.3)
CUTOFF	94.5%		86.5%		86.5%		81.5%		82%		77.5%		75%	

C. Archaea

	Genus		Family		Suborder[c]		Order		Subclass[c]		Class		Phylum	
No. Taxa	29		13		0		12		0		8		2	
Median [a]	96.7	(95.83, 97.33)	93.6	(91.75, 95.5)			91.44	(90.15, 93.6)			90.6	(86.5, 94.85)	81.9	(77.3, 86.5)
Minimum	95.3	(94.05, 96.3)	90.3	(87.75, 93.15)			87.1	(85.35,89.35)			85.83	(79.95, 90.8)	74	(68.3, 79.7)
CUTOFF	94%		87.7%				85.3%				80%		68.3%	

[a] Median values for all taxa were approximated with the pseudomedian and the 95% confidence interval using the Wilcoxon signed rank test. [b] 88% confidence interval was applied for these cases due to data shortage. [c] Categories suborder and subclass have never been proposed for archaeal species.

When comparing the values produced within the domain *Bacteria* to the domain *Archaea* we found slightly different results. The bacterial dataset (Table 4.2B) produced equivalent values as observed in the full dataset for all the taxonomic categories. *Archaea*, instead, yielded different results (e.g. 94% for genera, 87.7% for families) (Table 4.2C). However, these values might be highly biased due to the low number of described archaeal species (325 so far). The fact that all archaeal classes, excepting "Methanomicrobia", have been classified with just one order, might explain the high similarity values at the class level close to the order boundary. It is to expect that as new archaeal species are classified, these taxa gain internal structure tending to decrease the confidence intervals and to match the bacterial taxonomic cutoffs calculated.

Special taxa with particular thresholds?

The taxonomic boundaries showed in Table 4.2A can be safely applied to most of the prokaryotic taxa. Those taxa relying beyond the 95% confidence intervals, constitute exceptions that can be explained by means of the following reasons: internal taxon structure, total available number of species or sequences, classification age, and phylogenetic coherence of the taxon. ***Internal taxon structure*** makes reference to the internal fractioning of a group into different taxonomic levels. As we showed before, there are many higher taxa classified with a single genus, single family, etc. In these cases, the boundaries at distinct taxonomic ranks might overlap. For example, if an order is composed by only one family, then its particular order-level similarity cutoff would match the one found for its family (e.g. *Pasteurellales* and *Pasteurellaceae,* order and family boundaries = 89.15%; *Planctomycetales* and *Planctomycetaceae,* order and family boundaries = 76.7%). The same occurs for those classes or phyla composed by a single family, e.g. phylum “Synergistetes” = class *Synergista* = order *Synergistales* = family *Synergistaceae* = 79.9%. ***Total available number of species or sequences*** that can be used to form datasets influence statistical analysis (i.e. scarce input datasets may lead to poor conclusions whereas large datasets lead to robust statistical outcomes). Such is the case of the class *Acidobacteria*, that contains just eight species. ***Classification age*** of a taxon makes reference to the date when it was originally classified. Early classifications were performed with currently obsolete techniques and knowledge that grouped together species that modern taxonomy would show that they belonged to different taxa, e.g. family *Bacillaceae*, genus *Mycoplana* (*Shida et al., 1996;* Kang *et al.,* 2009). Finally, the ***phylogenetic coherence*** makes reference to their phylogenetic support in reliable tree reconstructions. At a certain rank, taxa that are not monophyletic (e.g. genus *Mycoplasma* = 73%, genus *Eubacterium* = 74.6%) will necessarily show higher 16S rRNA gene sequence divergences than clear cut monophyletic taxa (e.g. genus *Glycomyces* = 94.4%, genus *Staphylococcus*

= 95.7%).

Biased boundaries for some taxa might not only be understood with a combination of the four reasons explained above. For example, the major problem found behind the phylum *Spirochaetes* seems to be a classification schema that could be entirely updated so that taxonomy fits more naturally over a phylogenetic framework. Indeed all validly published higher taxa (i.e. with ranks above genus) for this phylum were described based on merely phenotypic traits (Cavalier-Smith, 2002; Buchanan, 1917; Hovind-Hougen, 1979; Swellengrebel, 1907). Here, its 109 described species are organized into 15 genera, 4 families, 1 order and 1 class. Many of these species correspond to uncultivable organisms (hence, orphan species), and just 67 type strains can be represented by a good SSU rRNA sequence in the LTP dataset. From the phylogenetic point of view, this phylum reveals a particularly clear separation of the distinct families and genera, indicated by long branches and high distance between clades (Fig. S4.11). By means of its SSU sequence identity, phylum "Spirochaetes", class *Spirochaetes* and order *Spirochaetales* showed equal cutoff values (72.3%) as they were composed by identical sequence datasets. This threshold would be appropriate for the rank phylum (75%), but still very low for ranks class (78.5%) and order (82%) (Table 4.2A). The family *Brachyspiraceae* showed a genus-like lower-cutoff value of 95.9% as was only represented by one genus (*Brachyspira*) in the LTP. Family *Spirochaetaceae* showed anomalous median similarity of 81.9% and a minimum of 74.9% due to the high distance between genus *Borrelia*, *Spirochaeta* and *Treponema*. Genus *Spirochaeta* showed anomalous value of 81.9% due to its paraphylethic structure (Fig. S4.11). Family *Leptospiraceae* showed a boundary at 79.2%, due to the high divergences found among its genera *Leptospira*, *Leptonema* and *Turneriella*. Overall, these results indicate that although a high phylogenetic support exists for current taxonomy on the phylum "Spirochaetes", its internal structure might be revised in order to reflect a more natural classification. Hence, solely taking into account 16S rRNA sequence data, we have found support for the following rearrangements: (i) families of the phylum "Spirochaetes" could be elevated to orders, including *Exilispira thermophila* as a fifth independent order; (ii) genera *Borrelia*, *Treponema*, *Brachyspira*, *Exilispira*, *Leptospira*, *Leptonema* and *Turneriella* could be elevated to separated Families; (iii) new orders could be arranged to create 4 distinct classes, one of them harboring two orders.

4.4 Phylogenetic classification of uncultured prokaryotes. The case of candidate division OP11.

One of the major problems that microbial molecular ecologists face when analyzing their 16S rRNA gene sequences directly retrieved from the environment is to find a rationale for clustering (Rosselló-Móra & López-López, 2008). Most often, the sequences found are grouped into units with a similarity >97% (Zaballos *et al.,* 2006) which are then treated as Operational Taxonomic Units (OTUs). This rather conservative grouping responds to the need to hidden microdiversity and sequencing errors in order to simplify the observations, and also to find a putative unit that may resemble taxonomic species (Acinas *et al.,* 2004). In order to understand whether a clade of uncultured microorganisms could be classified as putative (or “candidate“) taxa, we carried out an OTU-based hierarchical clustering analysis using our calculated taxonomic thresholds, followed by a detailed comparison against a reliable phylogenetic reconstruction. As a model, we selected the candidate division OP11 (Hugenholtz *et al.,* 1998) due to its apparently high phylogenetic diversity (Harris *et al.,* 2004) present in a relatively small set of full length SSU sequences. In December 2009, the candidate division OP11 consisted on a total of 422 SSU rRNA sequences within the SILVA SSURef102 database. Sequence quality parameters provided by SILVA allowed us to identify 62 sequences with high probability for being chimeric or just with very anomalous alignments, and therefore these were removed from the dataset. A distance similarity matrix was generated and submitted to DOTUR (Schloss *et al.,* 2009) using complete linkage clustering and analyzed at the following Euclidean distance levels: 0.055 (genus), 0.135 (family), 0.18 (order), 0.215 (class) and 0.25 (phylum) (obtained from similarity values shown in Table 4.2A). Suborder and subclass categories were not examined as their similarity cutoffs overlapped with those for family and order, respectively. Species category was also excluded from the experiment due to our suggested threshold was too narrow to compartmentalize groups. The low resolution power of 16S rRNA sequence for species recognition has been extensively discussed (Rosselló-Móra & Amann, 2001; Fox *et al.,* 1992). In the light of our taxa boundary calculations, the commonly used threshold in molecular microbial ecology of 97% identity may be closer to the genus category than to that of species. A phylogenetic tree was calculated and optimized with ARB_parsimony using the final dataset of 369 sequences. Its topology was further validated with an additional Neighbor Joining calculation using the jukes-cantor correction.

The rarefaction analyses (Fig. 4.3) indicate an important diversity coverage at the lowest categories as genus and family curves are far to reach the *plateau* of saturation. However, the artificial structuring using taxonomically meaningful thresholds indicated that the candidate division OP11 might be very compartmentalized. The results yielded 218 genera, 131 families, 70 orders, 39 classes and 20 phyla (Fig. 4.3). One of the main questions to address was whether the OTUs' delivery would be supported by coherent phylogenetic clades. Given that five OTUs classifications were modeled (each one at a distinct taxonomic boundary), each individual sequence was simultaneously assigned to a putative genus, family, order, class and phylum. At this point, the sequences of the distinct OTUs were

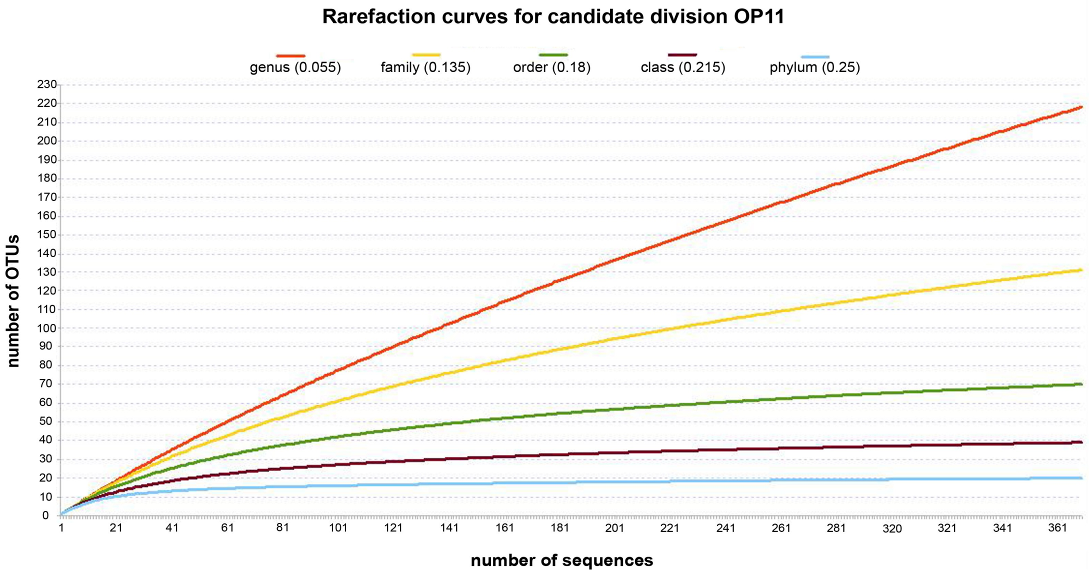


Fig 4.3. Rarefaction curves at distinct taxonomic thresholds.

recognized in the maximum parsimony reconstruction of the OP11 (Fig. S4.12) and grouped if they formed monophylethic clusters. Groups recognition was carried out from the highest category (phylum) to the lowest (genus). A code was designed to give human-readable names to the generated clades, in the following format: 'X.YN'. *X* stands for the name of the candidate division under study, *Y* is the taxonomic category at which OTUs were modeled (P = phylum, C = class, O = order, F = family and G = genus) and *N* is the name of the OTU which in our case was a numeric label automatically given by DOTUR. For example, OTU number 26 at the order level corresponds to the clade “OP11.O26” in the tree (Fig. S4.12). Overall, there was a good correlation between the OTU classification and clade discrimination (Table S4.1). Indeed most of the OTUs were either single-sequence branches (e.g. Uncultured bacterium, accession number = EU644101) or highly supported monophyletic clusters (e.g. “Candidate family OP11.F1”). However, in some cases (especially at rank order), OTUs appeared with a poly/paraphylethic pattern (e.g. OTU number 22 at rank order, in candidate phylum OP11.P5). For such cases, it is better to consider a monophyletic clade, disregarding rigid identity thresholds, in order to circumscribe a stable operational structure (what has been called Operational Phylogenetic Unit, OPU; (López-López *et al.,* 2010)). Such was the case of candidate class “OP11.C4-5-6-7” in the tree (Fig. S4.12).

In summary, the current taxonomic schema of the candidate division OP11 could be understood by the existence of 13 candidate phyla, 29 candidate classes, 54 candidate orders, 122 candidate families and 215 candidate genera (Table S4.1). Is important to note that this classification is entirely based on 16S rRNA data from environmental clones, for which very little is known about their physiology. OP11 candidate division has been retrieved from a variety of environments (sea and fresh-water sediments, hydrothermal systems, water column) but with the current data available it was not possible to detect any environmental factor that may support the monophyly of these taxa. We have found evidences for a possible multi-phylum structure of OP11 division (Fig. 4.4). It was supported by the maximum internal divergence of 64.1% (between sequences AJ583211 from candidate phylum OP11.P1-2-3-4 and AY193198 from candidate phylum OP11.P14-15). This value contrasts with the general observation for this taxonomic category (75%, Table 4.2). Although it is possible that we have overestimated the real number of taxa within the

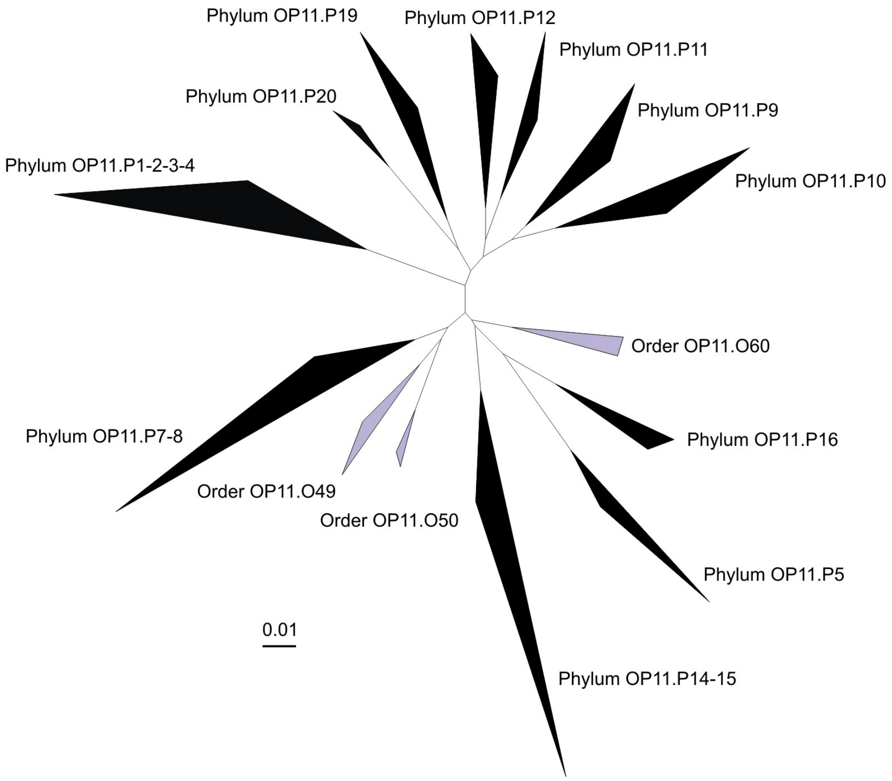


Fig. 4.4. Maximum-Parsimony phylogenetic reconstruction of the candidate division OP11. Collapsed overview showing candidate phyla. Scale bar indicates substitutions per site.

OP11 clade, this classification is, to our opinion, a useful and natural starting point for a future research on this group of microorganisms. By recognizing candidate taxa, uncultured organisms are placed into a provisional taxonomic schema that can be finally resolved after polyphasic characterization of type strains. Compartmentalizing candidate divisions into putative taxonomic categories would help to explain diversity found in studied environments, as well as to evaluate the degree of taxonomic novelty of the newly sequenced operational units.

4.5 Final remarks

Besides the genus-level classification that implicitly exists at every new species circumscription, classification at all the other taxonomic levels should be explicitly indicated in new species descriptions. Ideally, family-level classification can be indicated when describing a new genus (e.g. *Thalasosspira lucentensis,* (López-López *et al.,* 2002)), order-level when describing a new family (e.g. *Picrophilus torridus,* (Schleper *et. al.* 1996)), class-level when describing a new order (e.g. *Euzebia tangerina*, (Kurahashi *et al.,* 2010)), phylum-level when describing a new class (e.g. *Anaerolinea thermolimosa*, (Yamada *et al.,* 2006)) and domain-level when describing a new phylum (e.g. *Gemmatimonas auratiaca*, (Zhang *et al.,* 2003)). Phylogenetic reconstructions made up with large datasets, representing the whole diversity of the cultured and validly described prokaryotic species, have revealed a particular high coherence of the taxa at the family level (i.e. indicated by stable and well defined monophylethic branches).

The empirical taxonomic boundaries reported in this study reflect a summary of what taxonomists have always considered what a higher taxon might be. Nevertheless, they cannot be used as strict criterion but are especially recommended for prospective studies in clone libraries and as additional support for the circumscription of new taxa, or emendation of existing ones (e.g. *Spirochaetes*). Although suborder and subclass categories may show phylogenetic coherence (Stackebrandt *et al.,* 1997; Zhi *et al*., 2009), their sequence similarity cut-offs matched those for family and order, respectively. These apparently fuzzy boundaries between taxa might be obscured by data shortage (i.e 20 suborders and 5 subclasses classified so far). On the other hand, most of taxonomists have considered that families are grouped into orders, and orders into classes without intermediate levels.

A detailed comparison between an OTU-based classification (i.e. using rigid thresholds of sequence similarity to generate clusters) and an OPU-based classification (i.e. using reliable phylogenetic trees to distinguish meaningful clades) constitutes an approach to recognize coherent putative higher taxa. It can be applied to clades of uncultured microorganisms in order to lay down a provisional but common classification system usable for all scientists. Conclusions as those obtained for candidate division OP11 should be taken with care and submitted to revision in a future due to phylogenetic trees are dynamic structures that change on the basis of the quantity and quality of sequence data used (Ludwig & Klenk, 2001).

Acknowledgements

The authors acknowledge the Elsevier Publisher and the Max Planck Society. As well acknowledge their funding received from the Spanish Ministry of Science and Innovation with the projects Consolider Ingenio 2010 CE-CSD2007-0005 and CLG2009_12651-C02-01 and 02 (all co-financed with FEDER funding).

Discussion

The way microbiologists build the taxonomy of prokaryotes is linked to the belief that the order in nature is driven by evolutionary forces. Subsequently, the taxonomic classification has been constructed in the way that reflects the genealogies of taxa. However, unless we can understand all mechanisms involved in bacterial evolution, the classification will remain rigid and artificial at a certain extant (Rosselló-Móra, 2003). Fortunately, a general agreement in microbiology exists today for the nomenclature, classification and identification of prokaryotes. It is mainly due to: (i) a broadly accepted species concept for prokayotes (Rosselló-Móra & Amann, 2001), (ii) the existence of an official Bacteriological Code (Lapage *et al*., 1992), (iii) accepted standard techniques for strains characterization (Tindall *et al*., 2010) and, (iv) a reference phylogenetic framework for the placement of novel taxa (Garrity, 2001). Altogether, these four pillars may hold up our current view of taxonomy in microbiology.

The responsible teams of the ARB, SILVA and LPSN projects (www.arb-home.de, www.arb-silva.de, and www.bacterio.cict.fr) together with the journal Systematic and Applied Microbiology (SAM) started the “All-Species Living Tree Project” (LTP), a project conceived to provide a tool especially designed for the microbial taxonomist scientific community. The aim of the project was to set up a type-strain sequences database and single 16S and 23S rRNA trees harbouring all sequenced type strains of the hitherto classified species of *Archaea* and *Bacteria*. By updating the database and trees with the new validly published names that appear monthly in the International Journal of Systematic and Evolutionary Microbiology (IJSEM) the project was intended to be periodically released, in a manner that allowed users to download the new files and incorporate them into their work-flows. To that end a web platform was created where all additional material (i.e trees, alignments, datasets, tables of species, etc.) could be downloaded (www.arb-silva.de/projects/living-tree).

Departing from the same concept, another project called EzTaxon arose in parallel to the LTP at the Seoul National University (Chun *et al*., 2007). Although, both projects cover the same scope and target the same user community, it is necessary to point out some substantial deficiencies of the Korean project, based on which the LTP can offer a much robust and comprehensive service: (i) in the curation of a list of type-strain SSU sequences, EzTaxon did not provide a detailed explanation of the process itself nor about the difficulties found during the sieving of public DNA databases; (ii) the existence of type strains that can not be represented in high quality SSU datasets (i.e. either because they are 'orphan' or just account with bad quality sequence entries) is a very relevant issue that was sparsely explained in their manuscript and most probably not considered at all by the EzTaxon; (iii) EzTaxon was conceived as an on-line tool where sequence database is hidden, just accessible for queries trough a graphic user interface and without the possibility of being downloaded; (iv) neither curated alignments nor reference trees are offered by the EzTaxon.

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, (Ludwig & Klenk, 2001)), or combinations of them (Wu & Eisen, 2008), might better perform in reflecting genealogies of certain groups given the higher information content. Although far from reaching SSU levels, submission of alternative markers is growing fast, in part because: (i) the number of metagenomes and complete genomes (Liolios *et al*., 2008) 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 (Wu *et al*., 2009). 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 of time the number of sequenced genomes of type strains is still very low (Richter & Rosselló-Móra, 2009) and therefore the current possibilities for an in-depth taxonomic study are sparse.

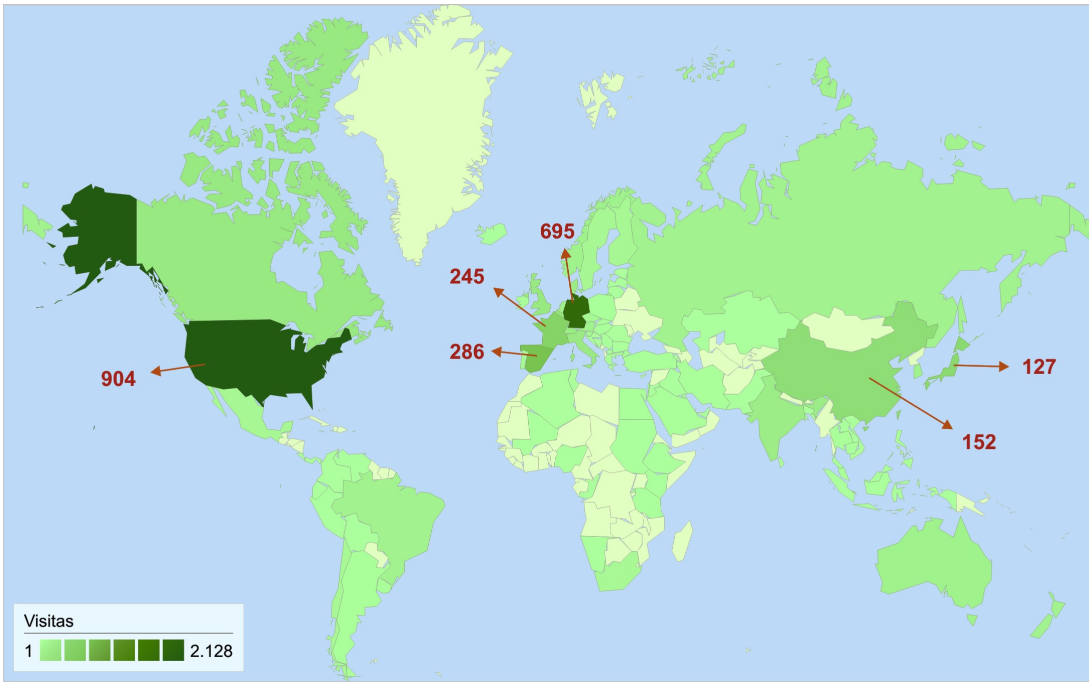
The creation of the first 16S and 23S rRNA-based LTP releases plus the consequent updates has been described in detail along the present manuscript. Setting up a type-strain sequences database passed trough the sieving of the public DNA databases whose sequence entries appeared often misleading or outdated. It involved the manual cross-check of nearly 14,000 SSU and 6,000 LSU entries against the catalogue of distinct species with validly published names retrieved from LPSN. Over 500 "orphan" species were not present in the SSU datasets since they did not account with an entry of adequate quality. The final SSU and LSU datasets that we provide may serve to facilitate the collection of sequences for the reconstruction of taxa genealogies. Despite the high quality of the original alignments provided by SILVA, a detailed manual revision was carried out to further preserve positional orthology. This improved alignment has been introduced into the SILVA seed, hence improving the accuracy of the SINA aligner (www.arb-silva.de) in forthcoming releases. The first maximum-likelihood reconstruction of the all-species tree accounted for 6,728 type strain SSU sequences (i.e. each one representing a single species) plus 3,247 additional supporting ones needed to stabilize the topology. On the other hand, the LSU tree was calculated upon a carefully selection of 1,900 sequences, 792 of them belonging to type strains. In order to validate the resulting topologies for both SSU and LSU trees, a detailed comparison against other published trees and topologies was carried out. It was particularly challenging for the case of the LSU tree due to data shortage in many groups. The tree topologies revealed cases of incoherent taxa (e.g. species that do not coherently affiliate with the rest of the members of their genus), as well as coherent and thus adequately classified ones. Is important to remark that trees are dynamic structures that change on the basis of the quality and quantity of the sequences used for their calculation. The aim of the LTP project was not to reconstruct the currently described species genealogy with total fidelity, but to provide a curated taxonomic tool for the scientific community.

In order to provide additional support to the All-Species Living Tree Project, a comprehensive study was conducted to evaluate the intra-genomic variability on complete type strain genomes. We observed that in very unusual exceptions the intra-genus (94.5%, (Yarza *et al*., 2008)) or intra-species (98.7%, (Stackebrandt & 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 host multiple copies of the *rrn* operon, only 2% of them contain divergences beyond 2% (30 nucleotides) sequence identity. Thus, most likely, the selection of one or another copy should not affect the phylogenetic reconstructions. For the first time, in the fifth LTP release (LTPs104, in preparation) we have decided to include all paralogs with higher divergences than 2%. This involved two SSU entries for *Haloarcula marismortui* ATCC 43049^{T} (with 5.7% of maximum inter-operonic divergence) and other two for *Thermoanaerobacter pseudethanolicus* ATCC 33223^{T} (with 3.66%).

The understanding that 7% of all classified species were missing from the ribosomal SSU sequence catalogues, motivated us to start the “Sequencing the Orphan Species” (SOS), a novel initiative on type strain sequencing. Eleven international culture collections joined us on this project consisting on the sequencing of 564 “orphan” type strains that were still missing from the LTP dataset. As coordinators, some of our tasks consisted on to: spread advertisement, contact collaborators, distribute the strains among the different partners (including ourselves) and, supervise the quality of the sequences obtained by means of their primary sequence (i.e. length and number of ambiguities) and their phylogenetic position. Currently, we have covered the 70% of the task list and we expect to achieve the remaining 30% along this year. 174 type strains have been excluded from our schedule since they could not be retrieved from public collections. They were either non-cultivable strains (115) or their cultures had been lost (59). As a consequence, fourteen phyla will never be completely represented on SSU datasets. For those species whose type strains were never deposited into culture collections or they cultures were lost, it is recommendable to designate a neotype or to request to the ICSP for the invalidation of the names.

The rich set of data generated by the LTP allowed us to deepen into still open questions regarding the coherence between current taxonomy and phylogenetic schemes. We made a revision of the higher taxonomic categories (i.e. from genus to phylum) in terms of their taxa content and phylogenetic coherence. In first instance, it was surprising the high abundance of taxa composed by just a single species, whereas quite often, descriptions of new taxa did not explicitly indicate classification at their successive higher taxonomic ranks. However, in some of these cases, phylogenetic analyses may support affiliation to an existing taxon. 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. By taking into account all the taxa at a particular taxonomic rank, we obtained lower-cutoff values of sequence identity for genus, family, order, class and phylum. These values are no more than the reflect of the empirical decisions of the responsible scientist creating categories. As shown by the low errors (specially at lower categories like genus and family), historically used criteria are homogeneous and do not lead to unambiguous circumscriptions. Nevertheless, they cannot be used as strict criterion but are especially recommended for prospective studies in clone libraries and as additional support for the circumscription of new taxa, or emendation of existing ones. One of our interests was to test the calculated empirical taxonomic thresholds for the classification of candidate taxa within clades of uncultured organisms. The candidate division OP11 was used as model case to apply these cutoffs. Operational Taxonomic Units (OTUs) delivered by using the thresholds, were compared to the Operational Phylogenetic Units (OPUs) delineated on a reliable topology. As a result, we found evidence that candidate division OP11 might be provisionally classified into 13 candidate phyla, 29 candidate classes, 54 candidate orders, 122 candidate families and 215 candidate genera. The example of OP11 may serve to lay down a start-up classification of "uncultured-sequence" clouds usable for all scientists.

In the present thesis manuscript we have presented the main aspects of the All-Species Living Tree Project, an initiative conceived to provide the scientific community with a curated taxonomic tool. During the course of more than three years of developing and updating effort, many new interesting questions about taxonomy and phylogeny of prokaryotes arose and were investigated in detail. Some of them consisted on contributions to other scientific works related to microbial ecology, systematics, genomics and metagenomics, directly related to the central purpose of the project (See appendix). We are proud to have received a very active input from the scientific community, who sent us requests and suggestions to help us with improving the final product. The complete set of materials that is released each time has been modified according to the needs of a growing user community. Accordingly new fields of information have been curated and included in the database. As a taxonomic tool the LTP must be understood as a collection of reference materials, including: (i) the sequence database of type-strain SSU and LSU sequences complemented with curated meta-data, (ii) the complete dataset of type strain sequences with the refined aligned in a text file, (iii) the whole species classification into a single phylogenetic tree and, (iv) a set of tables about species included in the LTP, orphan species, mistakes found in INSDC entries, etc. (Supplementary material). Apart from a good number of citations, the LTP web site has been visited by more than 3,500 users, distributed among 99 countries around the world, who recorded more than 4,100 downloads of one of the available files. The LTP continues its activity, and the fifth update of the project will be soon available, including the type strain sequences of all classified species up to December 2010.



LTP-web statistics. Map overlay showing the number of visits since 01-September-2008. In red, number of absolute unique visitors.

Conclusions

1. We have created a database containing the SSU and LSU sequences of the type strain of all species with validly published names up to February 2010.

2. From the complete catalogue of 8,602 species, 564 and 7,810 type strains could not be represented as they show lack of a good quality SSU and LSU sequence entry in public repositories, respectively. These numbers embrace a high number of type species of genera and type genera of families.

3. Each type strain sequence in LTP databases has been complemented with the following information: valid species name, type species designation, accepted classification into higher taxonomic ranks and risk group classification.

4. Over 12% of the type-strain SSU rRNA sequences that can be found in the INSDC databases carry outdated information in important fields for taxonomy (e.g. organism name, strain numbers). It is a reflect of the lack of updating effort of the entries at the time of a valid species publication which, typically, gets effective more than one year after the sequence submission.

5. We have provided high quality alignments for the 16S and 23S rRNA genes that may serve as reference for any reconstruction directed to the recognition of putative new species. The procedure implied a detailed revision of nearly 10,000 (SSU) and 2,000 (LSU) preliminary aligned sequences coming from SILVA databases.

6. To our knowledge, we have reported 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 (9,975 entries) representing the whole diversity of the cultured and validly described prokaryotic species.

7. Although vast majority of strains host multiple copies of the *rrn* operon, only 2% of them contain divergences beyond 2% (30 nucleotides) sequence identity. Thus, most likely, the selection of one or another copy should not significantly affect the phylogenetic reconstructions.

8. We have empirically proven that taxonomic categories genus, family, order, class and phylum can be circumscribed by means of their SSU and LSU rRNA sequence similarities with high reliability. Overall, a sequence identity below 94.5% entails the description of a new genus, 86.5% a new family, 82% a new order, 78.5% a new class and 75% a new phylum.

9. The LTP has started a novel initiative in collaboration with 11 international culture collections in order to sequence the orphan species. For the first time ever, we can complete the catalogue of 16S rRNA gene sequences of all the validly named species with an available type strain in public collections. By now, 275 strains have been sequenced and 115 are still in progress.

10. A total number of 174 type strains are not available in any known culture collection. As a consequence, fourteen phyla will remain incomplete in the SSU databases. For the 59 type strains whose cultures have been lost or never deposited in recognized culture collections, the designation of a neotype is recommended.

11. We have described a combined approach between an OTU-based classification (i.e. using rigid similarity thresholds to generate clusters) and an OPU-based classification (i.e. using reliable phylogenetic trees to distinguish meaningful clades) to recognize coherent putative higher taxa. We have applied this rationale to the candidate division OP11 and demonstrated that is possible to lay down a provisional classification scheme for “uncultured” sequences clouds based on candidate phyla, classes, orders, families and candidate genera.

Bibliography

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., Polz, M.F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J. Bacteriol. 186, 2629-2635.

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

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

Antón, J., Rosselló-Móra, R., Rodríguez-Valera, F., Amann, R. (2000) Extremely halophilic bacteria in crystallizer ponds from solar salterns. Appl. Environ. Microbiol. 66, 3052-3057.

Antunes, A., Rainey, F.A., Nobre, M.F., Schumann, P., Ferreira, A.M., Ramos, A., Santos, H., da Costa, M.S. (2002) *Leuconostoc ficulneum* sp. nov., a novel lactic acid bacterium isolated from a ripe fig, and reclassification of *Lactobacillus fructosus* as *Leuconostoc fructosum* comb. nov. Int. J. Syst. Evol. Microbiol. 52, 647-655.

Anzai, Y., Kim, H., Park, J.Y., Wakabayashi, H. Oyaizu, H. (2000) Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. Int. J. Syst. Evol. Microbiol. 50, 1563-1589.

Austin, D.A., Moss, M.O. (1986) Numerical taxonomy of red-pigmented bacteria isolated from lowland river, with the description of a new taxon *Rugamonas rubra* gen. nov., sp. Nov. J. Gen. Microbiol. 132, 1899-1909.

Breed, R.S., Murray, E.G.D., Smith, N.R. (Eds.) (1923) Bergey's manual of determiantive bacteriology, first edition. The Williams & Wilkins Co, Baltimore.

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

Buchanan R.E. (1917) Studies in the nomenclature and classification of the bacteria. II. The primary subdivisions of the Schizomycetes. J. Bacteriol. 2, 155-164.

Buckley, M., Roberts, R.J. (2005) Reconciling microbial systematics and genomics, American Academy of Microbiology, ASEM, Washington, DC, http://academy.asm.org/index.php/colloquia-reports/browse-all/249-reconciling-microbial-systematics-and-genomics-march-2007-b.

Cavalier-Smith, T. (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int. J. Syst. Evol. Microbiol. 52, 7-76.

Christensen H., Bisgaard M., Frederiksen W., Mutters R., Kuhnert P., Olsen J.E. (2001) Is characterization of a single isolate sufficient for valid publication of a new genus or species? Proposal to modify recommendation 30b of the Bacteriological Code (1990 Revision). Int. J. Syst. Evol. Microbiol. 51:2221-2225.

Chun, J., Lee, J.-H., Jung, Y., Kim, M., Kim, S., Kim, B.K., Lim, Y.-W. (2007) EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. Int. J. Syst. Evol. Microbiol. 57, 2259–2261.

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

Clayton, R., Sutton, G., Hinkle, P.S., Jr., Bult, C., Fields, C. (1995) Intraspecific variation in small-subunit rRNA sequences in genbank: why single sequences may not adequately represent prokaryotic taxa. Int. J. Syst. Bacteriol. 45, 595-599.

Coenye, T., Gevers, D., Van de Peer, Y., Vandamme, P., Swings, J. (2005) Towards a prokaryotic genomic taxonomy. FEMS Microbiol. Rev. 29:147–167.

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., 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.

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., Tiedje, J.M. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141-D145.

Cole, J.R., Konstantinidis, K., Farris, R.J., Tiedje, J.M. (2010) Microbial diversity and phylogeny: extending from rRNAs to genomes. In: W.-T. Liu and J.K. Jansson (Eds.), Environmental molecular microbiology. Caister Academic Press, pp. 1-19.

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

De Vos, P., Trüper, H.G. (2000) Judicial Commission of the International Committee on Systematic Bacteriology. IXth International (IUMS) Congress of Bacteriology and Applied Microbiology. Minutes of the meetings, 14, 15 and 18 August 1999, Sydney, Australia. Int. J. Syst. Evol. Microbiol. 50, 2239-2244.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol, 72, 5069-5072.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. Science 284, 2124-2128.

Euzéby, J. (1997) List of bacterial names with standing in nomenclature: a folder available on the Internet. Int. J. Syst. Bacteriol. 47, 590-592.

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

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

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

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

Garrity, G.M. (2001) Bergey’s Manual of Systematic Bacteriology, second ed. Springer, New York, 2001.

Garrity G.M., Lilburn T.G., Cole J.R., Harrison S.H., Euzéby J. and Tindall B.J. (2007) Taxonomic Outline of the Bacteria and Archaea, Release 7.7. Michigan State University Board of Trustees.

Geissinger, O., Herlemann, D.P.R., Morschel, E., Maier, U.G., Brune, A. (2009) The ultramicrobacterium “*Elusimicrobium minutum*” gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. Appl. Environ. Microbiol. 75, 2831–2840.

Gherna, R., Woese, C.R. (1992) A partial phylogenetic analysis of the "flavobacter-bacteroides" phylum: basis for taxonomic restructuring. Syst. Appl. Microbiol. 15, 513-521.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., Field, K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345, 60-63.

Gromov, B.V., Mamkayeva, K.A. (1980) Proposal of a new genus *Vampirovibrio* for chlorellavorus bacteria previously assigned to *Bdellovibrio*. Mikrobiologia 49, 165-167.

Harris, J.K., Kelley, S.T., Pace, N.R. (2004) New Perspective on Uncultured Bacterial Phylogenetic Division OP11. Appl. Environ. Microbiol. 70, 845-849.

Heyrman, J., Verbeeren, J., Schumann, P., Swings, J., De Vos, P. (2005) Six novel *Arthrobacter* species isolated from deteriorated mural paintings. Int. J. Syst. Evol. Microbiol. 55, 1457-1464.

Hovind-Hougen, K. (1979) *Leptospiraceae*, a new family to include *Leptospira* Noguchi 1917 and *Leptonema* gen. nov. J. Syst. Bacteriol. 29, 245-251.

Hu, F.P., Young, J.M., Stead, D.E., Goto, M. (1997) Transfer of *Pseudomonas cissicola* (Takimoto 1939) Burkholder 1948 to the genus *Xanthomonas*. Int. J. Syst. Bacteriol. 47, 228-230.

Hugenholtz, P., Pitulle, C., Hershberger, K.L., Pace, N.R. (1998) Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. 180, 366-376.

International Committee on Bacteriological Nomenclature (Eds.) (1958) International code of nomenclature of bacteria and viruses. Ames, Iowa State College Press.

Judicial Commission of the International Committee on Systematics of Prokaryotes (2008) 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, 1755-1763.

Kaesler, I., Graeber, I., Borchert, M.S., Pape, T., Dieckmann, R., Von Döhren, H., Nielsen, P., Lurz, R., Michaelis, W., Szewzyk, U. (2008) *Spongiispira norvegica* gen. nov., sp. nov., a marine bacterium isolated from the boreal sponge *Isops phlegraei*. Int. J. Syst. Evol. Microbiol. 58, 1815-1820.

Kang, S.-J., Choi, N.-S., Choi, J.H., Lee, J.-S., Yoon, J.-H., Song, J.J. (2009) *Brevundimonas naejangsanensis* sp. nov., a proteolytic bacterium isolated from soil, and reclassification of *Mycoplana bullata* into the genus *Brevundimonas* as *Brevundimonas bullata* comb. nov. Int. J. Syst. Evol. Microbiol. 59, 3155-3160.

Kim, M.K., Kim, Y.J., Cho, D.H., Yi, T.H., Soung, N.K. and Yang, D.C. (2007) *Solimonas soli* gen. nov., sp. nov., isolated from soil of a ginseng field. Int. J. Syst. Evol. Microbiol. 57, 2591-2594.

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

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Garcia-Pastor, M.P., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W., Apweiler, R. (2007) EMBL nucleotide sequence database in 2006. Nucleic Acids Res. 35, D16–D20.

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

Kurahashi, M., Fukunaga, Y., Sakiyama, Y., Harayama, S., Yokota, A. (2010) *Euzebya tangerina* gen. nov., sp. nov., a deeply branching marine actinobacterium isolated from the sea cucumber *Holothuria edulis*, and proposal of *Euzebyaceae* fam. nov., *Euzebyales* ord. nov. and *Nitriliruptoridae subclassis* nov. Int. J. Syst. Evol. Microbiol. 60, 2314-2319.

Kurland, C.G. (2005) The paradigm lost. In: Sapp, J. (Ed.), Microbial Phylogeny and Evolution Concepts and Controversies. Oxford University Press, Oxford, pp. 207-223.

Labeda, D.P. (1997a) Judicial Commission of the International Committee on Systematic Bacteriology: VIIth International Congress of Microbiology and Applied Bacteriology. Minutes of the meetings, 17 and 22 August 1996, Jerusalem, Israel. Int. J. Syst. Bacteriol. 47, 240-241.

Labeda, D.P. (1997b) International Committee on Systematic Bacteriology: VIIth International Congress of Microbiology and Applied Bacteriology. Minutes of the meetings, 17, 18, and 22 August 1996, Jerusalem, Israel. Int. J. Syst. Bacteriol., 1997, 47, 597-600.

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

Lee, Z.M.-P., Bussema 3rd, C., Schmidt, T.M. (2009) rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res. 37, D489–D493.

Liolios, K., Mavromatis, K., Tavernarakis, N., Kyrpides, N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 36, D475–D479.

Lapage, S. P., Sneath, P. H. A., Lessel, E. F., Skerman, V. B. D., Seeliger, H. P. R., Clark, W. A. (Eds.) (1976) International code of nomenclature of Bacteria (1975 revision). American Society for Microbiology. Washington, DC.

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

Loginova, G.B.U., Egorova, L.A. (1978) A new species of obligate thermophilic bacteria, *Flavobacterium thermophilum*. Mikrobiologia 47, 1081-1084.

López-López, A., Yarza, P., Suárez-Suárez, A., Antón, J., Niemann, H., and Rosselló-Móra, R. (2010) Extremely halophilic microbial communities in anaerobic sediments from a solar saltern. Environ. Microbiol. Rep. 2: 258-271.

López-López, A., Pujalte, M.J., Benlloch, S., Mata-Roig, M., Rossello-Mora, R., Garay, E., Rodriguez-Valera, F. (2002) *Thalassospira lucentensis* gen. nov., sp. nov., a new marine member of the alpha-Proteobacteria. Int. J. Syst. Evol. Microbiol. 52, 1277-1283.

Ludwig, W., Schleifer, K.H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol. Reviews 15, 155-173.

Ludwig, W., Rosselló-Móra, 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., Schleifer, K.H. (1995) Comparative sequence analysis of 23S rRNA from proteobacteria. Syst. Appl. Microbiol. 18, 164-188.

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

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

Ludwig, W., Schleifer, K.-H. (2005) The molecular phylogeny of Bacteria based on Conserved genes. In: Sapp, J. (Ed.), Microbial Phylogeny and Evolution Concepts and Controversies. Oxford University Press, Oxford, pp. 70-98.

Ludwig, W., Schleifer, K.-H., Whitman, W.B. (2009) Revised road map to the phylum Firmicutes, in: Whitman, W.B. (Ed.), Bergey's Manual of Systematic Bacteriology, vol. 3, second ed., Springer, New York, pp. 1-13.

Ludwig, W. (2010) Molecular Phylogeny of Microorganisms: Is rRNA still a useful marker? In: Oren, A., Papke, R.T. (Eds.), Molecular Phylogeny of Microorganisms, Caister Academic Press, pp. 65-83.

Moreira, C., Rainey, F.A., Nobre, M.F., da Silva, M.T., da Costa, M.S. (2000) *Tepidimonas ignava* gen. nov., sp. nov., a new chemolithoheterotrophic and slightly thermophilic member of the beta-Proteobacteria. Int. J. Syst. Evol. Microbiol. 50, 735-742.

Mori, K., Yamaguchi, K., Sakiyama, Y., Urabe, T., Suzuki, K. (2009) *Caldisericum exile* gen. nov., sp. nov., an anaerobic, thermophilic, filamentous bacterium of a novel bacterial phylum, *Caldiserica phyl*. nov., originally called the candidate phylum OP5, and description of *Caldisericaceae fam*. nov., *Caldisericales ord*. nov. and *Caldisericia classis* nov. Int. J. Syst. Evol. Microbiol. 53, 2894-2898.

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

Murray, R.G.E., Stackebrandt, E. (1995) Taxonomic Note: implementation of the provisional status Candidatus for incompletely described procaryotes. Int. J. Syst. Bacteriol. 45, 186-187.

Mylvaganam, S., Dennis, P. (1992) Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics 130, 399-410.

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

Osterman, B., Moriyón, I. (2006) International Committee on Systematics of Prokaryotes Subcommittee on the taxonomy of *Brucella*. Minutes of the meeting, 17 September 2003, Pamplona, Spain. Int. J. Syst. Evol. Microbiol. 56, 1173-1175.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. Science 276, 734-740.

Peplies, J., Kottmann, R., Ludwig, W., Glöckner, F.O. (2008) A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. Syst. Appl. Microbiol. 31, 251-257.

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

Richter, M., Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U.S.A. 106, 19126-19131.

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

Rosselló-Móra, R. (2003) The species problem, can we achieve a universal concept? Syst. Appl. Microbiol. 26, 323-326.

Rosselló-Móra, R. (2005) Updating prokaryotic taxonomy. J. Bacteriol. 187, 6255-6257.

Rosselló-Móra, R., and López-López, A. (2008) The last common denominator: species or operational taxonomic units? In: Accessing uncultivated microorganisms: from the environment to organisms and genomes and back. Zengler, K. (ed). Washington, DC: ASM Press, pp. 117-130.

Schleper, C., Puehler, G., Holz, I., Gambacorta, A., Janekovic, D., Santarius, U., Klenk, H.P., Zillig, W. (1996) *Picrophilus oshimae* and *Picrophilus torridus* fam. nov., gen. nov., sp. nov., two species of hyperacidophilic, thermophilic, heterotrophic, aerobic Archaea. Int. J. Syst. Bacteriol. 46, 814-816.

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., Weber, C.F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537-7541.

Shah, H.N., Collins, M.D. (1986) Reclassification of *Bacteroides furcosus* Veillon and Zuber (Haudoroy, Ehringer, Urbain, Guillot and Magrou) in a new genus *Anaerorhabdus* as *Anaerorhabdus furcosus* comb. nov. System. Appl. Microbiol. 8, 86-88.

Shida, O., Takagi, H., Kadowaki, K., Komagata, K. (1996) Proposal for Two New Genera, *Brevibacillus* gen. nov. and *Aneurinibacillus* gen. nov. Int. J. Syst. Bacteriol. 46, 939-946.

Skerman, V.B.D., McGowan, V., Sneath, P.H.A. (1980) Approved Lists of Bacterial Names. Int. J. Syst. Bacteriol. 30, 225-420.

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

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

Sorokin, D.Y., Kovaleva, O.L., Tourova, T.P., Muyzer, G. (2010) *Thiohalobacter thiocyanaticus* gen. nov., sp. nov., a moderately halophilic, sulfur-oxidizing gammaproteobacterium from hypersaline lakes, that utilizes thiocyanate. Int. J. Syst. Evol. Microbiol. 60, 444-450.

Stackebrandt, E., Goebel, B.M. (1994) 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, 846-849.

Stackebrandt, E., Rainey, F.A., Ward-Rainey, N.L. (1997) Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. Int. J. Syst. Bacteriol. 47, 479-491.

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

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

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

Swellengrebel, N.H. (1907) Sur la cytologie comparée des spirochètes et des spirilles. Ann. Inst. Pasteur (Paris) 21, 562-586.

Tindall, B.J., Kämpfer, P., Euzéby, J.P., Oren, A. (2006) Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. Int. J. Syst. Evol. Microbiol. 56, 2715-2720.

Tindall, B.J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W., Kämpfer, P. (2010) Notes onthe characterization of prokaryote strains for taxonomic purposes. Int. J. Syst. Evol. Microbiol. 60, 249–266.

Van Den Mooter, M., Swings, J. (1990) Numerical analysis of 295 phenotypic features of 266 *Xanthomonas* strain and related strains and an improved taxonomy of the genus. Int. J. Syst. Bacteriol. 40, 348-369.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., Swings, J. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Rev. 60, 407–438.

Vandamme, P., Debruyne, L., Debrandt, E., Falsen, E. (2010) Reclassification of *Bacteroides ureolyticus* as *Campylobacter ureolyticus* comb. nov., and emended description of the genus *Campylobacter*. Int. J. Syst. Evol. Microbiol. 60, 2016-2022.

Westram, R., Bader, K., Prüsse, E., Kumar, Y., Meier, H., Glöckner, F.O., Ludwig, W. (2010) ARB: a software environment for sequence data. In: F.J. de Bruijn (Ed.), Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches. John Wiley & Sons, Hoboken, New Jersey. In press.

Woese, C.R. (1987) Bacterial evolution. Microbiol. Rev. 51, 2, 221-271.

Woese, C.R., Kandler, O., Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc. Natl. Acad. Sci. USA 87, 4576-4579.

Wu, M., Eisen, J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.

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, L.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., Eisen, J.A. (2009)A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. Nature 462, 1056-1060.

Yamada, T., Sekiguchi, Y., Hanada, S., Imachi, H., Ohashi, A., Harada, H., Kamagata, Y. (2006) *Anaerolinea thermolimosa* sp. nov., *Levilinea saccharolytica* gen. nov., sp. nov. and *Leptolinea tardivitalis* gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes *Anaerolineae* classis nov. and *Caldilineae* classis nov. in the bacterial phylum Chloroflexi. Int. J. Syst. Evol. Microbiol. 56, 1331-1340.

Yarza, P., Richter, M., Peplies, J., Euzéby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., 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, 241–250.

Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F. O., Rosselló-Móra, R. (2010) Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. System. Appl. Microbiol. 33, 291-299.

Yokoyama, H., Wagner, I.D., Wiegel, J. (2010) *Caldicoprobacter oshimai* gen. nov., sp. nov., an anaerobic, xylanolytic, extremely thermophilic bacterium isolated from sheep faeces, and proposal of *Caldicoprobacteraceae* fam. nov. Int. J. Syst. Evol. Microbiol. 60, 67-71.

Zaballos, M., López-López, A., Ovreas, L., Bartual, S.G., D'Auria, G., Alba, J.C., Legault, B., Pushker, R., Daae, F.L., Rodríguez-Valera, F. (2006) Comparison of prokaryotic diversity at offshore oceanic locations reveals a different microbiota in the Mediterranean Sea. FEMS Microbiol. Ecol. 56, 389-405.

Zhang, H., Sekiguchi, Y., Hanada, S., Hugenholtz, P., Kim, H., Kamagata, Y., Nakamura, K. (2003) *Gemmatimonas aurantiaca* gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum *Gemmatimonadetes phyl*. nov. Int. J. Syst. Evol. Microbiol. 53, 1155–1163.

Zhi, X.-Y., Li, W.-J., Stackebrandt, E. (2009) An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class *Actinobacteria*, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. Int. J. Syst. Evol. Microbiol. 59, 589-608.

Resumen

Breve revisión sobre taxonomía microbiana.

La clasificación de los organismos vivos ha sido uno de los tópicos más recurrentes en ciencia, debido a la necesidad de obtener y compartir conocimiento en biología. Aristóteles fue el primero en enfrentarse al concepto de especie hace aproximadamente unos 400 a.c., y nuestra actual percepción del orden en la naturaleza está basada en las definiciones de Lineo desde hace más de dos siglos. Al contrario que la botánica o la zoología, la microbiología es una disciplina reciente que debe su nacimiento y desarrollo a los avances tecnológicos (como el microscopio). Los microorganismos no aparecían en los registros fósiles, eran invisibles a simple vista y mostraban un muy bajo polimorfismo fenotípico. Sin embargo, la primera parte de la historia de la taxonomía de procariotas estuvo enteramente influenciada por los sistemas botánicos y zoológicos, y con una circunscripción de nuevas categorías basada solamente en morfología. En este sentido, los avances en técnicas de cultivo y la capacidad de obtener cultivos puros fueron cruciales para empezar a entender sus rasgos fisiológicos. Hasta el principio del siglo XX varias clasificaciones independientes habían surgido siguiendo criterios diferentes, pero no fue hasta el año 1923 cuando se publicó el primer trabajo de referencia llamado *Bergey's Manual of Determinative Bacteriology*, el cual, durante las siguientes ediciones, se consolidaría como el marco común para los taxónomos microbianos hasta nuestro nuestros días (actualmente conocido como *Bergey's Manual of Systematic Bacteriology*). A mitad del siglo XX, los microbiólogos empezaron a sacar provecho de los avances tecnológicos de otras disciplinas con tal de mejorar los viejos métodos de caracterización de bacterias, y por lo tanto, acercarse a estos microorganismos desde un punto de vista más bioquímico. Con el creciente conocimiento sobre el papel que los ácidos nucleicos juegan en la vida celular, los científicos comenzaron por primera vez a usar parámetros genómicos tales como el contenido molar en G+C y la hibridación DNA-DNA (DDH). La DDH se convirtió en una técnica estándar para la circunscripción de nuevos taxones y ha sido la principal directora del esquema taxonómico procariótico tal y como lo conocemos hoy en día. Sin embargo, a mediados de los años 1970, una última contribución científica fue la que más significativamente ha cambiado nuestra visión en la clasificación de procariotas. Fue el uso de “relojes moleculares”, tales como genes ribosómicos or ciertas proteínas conservadas, para inferir relaciones genealógicas a partir de árboles filogenéticos. La secuencia de la subunidad pequeña del ribosoma (SSU) fue la clave para establecer un marco más fiable y natural para la clasificación de los procariotas, y es el que actualmente aceptamos.

Situación actual.

El actual panorama de la clasificación de los procariotas está basado principalmente en afiliaciones genealógicas, y la circunscripción de cualquier nuevo taxón con una jerarquía superior a la especie (i.e. género y superiores) está basada en parentesco genealógico. Aunque el gen de la subunidad grande (LSU) del ribosoma está considerado como el cronómetro molecular más informativo, puesto que dobla en tamaño y contenido informativo al de la subunidad pequeña (SSU), limitaciones técnicas y económicas han favorecido el análisis de la segunda, que se ha establecido como el estándar para la reconstrucción de relaciones filogenéticas entre procariotas con propósitos de clasificación. Como consecuencia, el depósito de secuencias de SSU en bases de datos públicas ha aumentado exponencialmente en torno a tres órdenes de magnitud en aproximadamente 15 años.

La cepa tipo de una especie es un aislado que se debe depositar en diversas colecciones de cultivo internacionales, y es la cepa a la cual se hace referencia en el protólogo que describe al nuevo taxón. Esta cepa es la verdadera referencia que garantiza la correcta identificación de nuevos posibles miembros del mismo taxón. Es una aproximación común el identificar la singularidad de una nueva especie mediante el chequeo previo de que no existe una secuencia públicamente disponible de la cepa tipo. Por esta razón, la mayoría de las descripciones de nuevas especies y géneros están generalmente acompañadas por la secuencia del gen del SSU de las cepas tipo.

Una de las premisas más importantes para poder reconocer la singularidad de nuevos taxones es la identificación de las secuencias de cepas tipo disponibles en bases de datos públicas. Por desgracia, este paso está actualmente dificultado por la información inexacta de las secuencias alojadas en la “International Nucleotide Sequence Database Consortium” (INSDC; www.insdc.org), que engloba a EMBL, GenBank y DDBJ. Los errores más frecuentes están relacionados con nombres incorrectos de especies, números de acceso mal asignados o números de colección equivocados. Además, la correspondiente secuencia primaria depositada puede ser de baja calidad, traduciéndose en difíciles o imposibles reconstrucciones filogenéticas.

Objetivos.

En esta memoria presentamos los aspectos más relevantes del “All-Species Living Tree Project (LTP)”, una colaboración internacional entre la revista científica Systematic and Applied Microbiology (ELSEVIER) y el grupo de científicos responsables de los proyectos LPSN (www.bacterio.cict.fr), ARB (www.arb-home.de) y SILVA (www.arb-silva.de), cuya motivación es la de proporcionar a la comunidad científica una herramienta taxonómica de gran utilidad. El trabajo de creación, mantenimiento y gestión del LTP ha estado enfocado a la superación de los siguientes objetivos principales:

- ✔ Proporcionar una base de datos depurada de SSU y LSU con todas las cepas tipo, de las especies con nombres válidamente publicados, para las cuales existan entradas de calidad adecuada.
- ✔ Poner a punto un alineamiento optimizado y de carácter universal.
- ✔ Reconstruir un único árbol filogenético que albergue topologías fiables.
- ✔ Proporcionar actualizaciones regulares de la base de datos, alineamientos y árboles con los nuevos taxones válidamente publicados.
- ✔ Crear una página web para el proyecto, donde almacenar el conjunto completo de materiales para que puedan ser descargados gratuitamente.
- ✔ Investigar, mediante el uso de la base de datos, aspectos fundamentales sobre la taxonomía de procariotas, tales como: umbrales filogenéticos en la circunscripción de

nuevos taxones, coherencia de la taxonomía actual en cuanto a su esquema filogenético y, la relevancia del ARNr 16S en estudios taxonómicos.

✔ Conseguir por primera vez el catálogo completo de secuencias de SSU para todas las especies clasificadas hasta el momento, mediante la secuenciación de las más de 500 especies que todavía permanecen ausentes.

Relación de capítulos.

Todos los capítulos de la presente memoria han sido originalmente escritos en inglés para su publicación en revistas científicas de ámbito internacional. Por lo tanto, cada uno de los capítulos ha sido presentado del mismo modo en que deben ser enviados para su publicación.

Chapter 1 **Yarza, P., Richter, M., Peplies, J., Euzéby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., 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, 241-250.

Chapter 2 **Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.H., Glöckner, F.O., Rosselló-Móra, R.** (2010) Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst. Appl. Microbiol., 33, 291-299.

Chapter 3 **Yarza, P., Euzéby, J., Spröer, C., Mrotzek, N., Swiderski, J., Tindall, B.J., Pukall, R., Spring, S., Lang, E., Gronow, S., Verbarg, S., Klenk, H.-P., Crouch, A., Beck, B., Unosson, C., Moore, E.R.B., Nakagawa, Y., Clermont, D., Janssens, D., Sakamoto, M., Iida, T., Kudo, T., Kosako, Y., Oshida, Y., Ohkuma, M., Arahal, D.R., Spieck, E., Pommerening-Roeser,**

A., Figge, M., Park, D., Buchanan, P., Nicholson, A., Cifuentes, A., Schleifer, K.-H., Amann, R., Glöckner, F.O., Rosselló-Móra, R. Taxonomic note: SOS, Sequencing Orphan Species: filling the gaps in the 16S rRNA gene sequence database for all classified species with validly published names. In preparation.

Chapter 4 **Yarza, P., Euzéby, J., Ludwig, W., Amann, R., Glöckner, F.O., Schleifer, K.-H., Rosselló-Móra, R.** Empirical circumscription of prokaryotic higher taxa based on comparative analyses of the 16S rRNA gene. In preparation.

Conclusiones.

1. Hemos creado una base de datos de secuencias de SSU y LSU de las cepas tipo de todas las especies con nombres válidamente publicados hasta febrero de 2010.

2. Del catálogo completo de 8.602 especies, 564 y 7.810 no se pudieron representar ya que no disponían de una secuencia de SSU y LSU de buena calidad en repositorios públicos, respectivamente. Estos números abarcan un alto número de especies tipo de géneros y géneros tipo de familias.

3. Cada una de las secuencias de cepa tipo en las bases de datos del LTP se ha complementado con la siguiente información adicional: nombre válido de la especie, designación de especie tipo, clasificación actualmente aceptada en rangos taxonómicos superiores a género y, clasificación en grupos de riesgo.

4. Más del 12% de las secuencias de SSU de cepas tipo que se pueden encontrar en las bases de datos del INSDC contienen errores en campos importantes para la taxonomía (por ejemplo, nombre del organismo y números de cepa). Esto es un

reflejo de la falta de actualización de las entradas en el momento de la publicación válida de la especie que, normalmente, sucede en torno a un año más tarde que el depósito de la secuencia.

5. Hemos desarrollado alineamientos de alta calidad para los genes de ARNr 16S y 23S, que pueden servir como referencia universal para cualquier reconstrucción dirigida a la identificación de posibles nuevas especies procarióticas. El procedimiento implicó una revisión detallada de cerca de 10.000 secuencias de SSU y 2.000 secuencias de LSU procedentes de las bases de datos del SILVA.

6. A nuestro saber, hemos proporcionado la primera reconstrucción de todas las especies basada en una selección cuidadosa de secuencias de cepas tipo de *Bacteria* y *Archaea*. El producto final tiene dos valores añadidos: (i) un conjunto de datos depurado hecho a partir de secuencias representantes de las cepas tipo de todas las especies descritas hasta la fecha y, (ii) la primera reconstrucción por *maximum-likelihood* a partir de un conjunto de secuencias tan grande (9.975), y representativo de la diversidad procariótica completa de especies cultivadas y válidamente publicadas.

7. Aunque la gran mayoría de los genomas alberga múltiples copias del operón ribosómico, solamente el 2% presenta divergencias superiores al 2% (~ 30 nucleótidos) de identidad de secuencia. Por lo tanto, muy probablemente la selección de una copia u otra no debe afectar significativamente a las reconstrucciones filogenéticas.

8. Hemos demostrado empíricamente que las categorías taxonómicas género, familia, orden, clase y filo, pueden ser circunscritas con alta fiabilidad en función de sus niveles de similitud de secuencias de SSU y LSU. En general, una identidad por debajo de 94,5% conduce a la descripción de un nuevo género, 86,5% una nueva familia, 82% un nuevo orden, 78.5% una nueva clase y 75% un nuevo filo.

9. El LTP ha comenzado una nueva colaboración internacional con 11 colecciones de cultivo internacionales para secuenciar las especies huérfanas. Por primera vez, podemos completar el catálogo de secuencias del gen del ARNr 16S con todas las especies válidamente publicadas y presentes en colecciones de cultivo públicas. Por ahora, 275 cepas tipo ya han sido secuenciadas y 115 están proceso.

10. Un número total de 174 cepas tipo no está disponible en ninguna colección de cultivo conocida. Como consecuencia, 14 filos permanecerán incompletos en las bases de datos de SSU. Al parecer 59 cepas tipo nunca fueron depositadas en una colección o se han perdido, y por lo tanto recomendamos la designación de un neotipo.

11. Hemos descrito una metodología que combina una clasificación basada en OTUs (usando umbrales rígidos de identidad de secuencia) y una clasificación basada en OPUs (usando reconstrucciones filogenéticas para distinguir clados significativos), para reconocer la coherencia de taxones candidatos. Hemos empleado esta aproximación sobre la división candidata OP11 y demostrado que es posible establecer un esquema de clasificación provisional basado en filos candidatos, clases, órdenes, familias y géneros candidatos para los grupos de secuencias de organismos no cultivados.

Perspectiva.

En el transcurso de más de tres años de desarrollo y esfuerzo de actualización, han ido surgiendo cuestiones muy interesantes sobre taxonomía y filogenia de procariotas. Algunas de ellas se investigaron en profundidad y consistieron en contribuciones a otros trabajos científicos relacionados con ecología, sistemática, genómica y metagenómica microbianas, directamente relacionadas con el propósito central del proyecto (ver apéndice). Estamos orgullosos de haber obtenido una respuesta muy activa por parte de la comunidad científica, habiendo recibido sugerencias y peticiones que en definitiva nos ayudaron a mejorar el producto final. El set de materiales que se publica, se ha ido modificando cada vez de acuerdo a las necesidades de una comunidad creciente de usuarios. Por consiguiente, nuevos campos con información depurada se han incluido en la base de datos. Como herramienta taxonómica, el LTP se debe de entender como una colección de materiales de referencia, incluyendo: (i) la base de datos de secuencias de SSU y LSU de cepas tipo con información adicional depurada, (ii) el conjunto entero de secuencias de cepas tipo alineadas en un archivo de texto, (iii) la clasificación completa de las especies representada en un árbol filogenético y, (iv) el conjunto de tablas de especies incluidas en el LTP, especies "huérfanas", errores encontrados en las bases de datos del INSDC, etc. (ver material suplementario). Aparte de un buen número de citas, la página web del LTP ha sido visitada por más de 3.500 usuarios distintos, distribuidos en 99 países alrededor del mundo, los cuales han registrado un total de 4.100 descargas de alguno de los archivos disponibles. El LTP continúa su actividad y muy pronto estará disponible la quinta actualización que incluirá las secuencias de las cepas tipo de las especies clasificadas hasta diciembre de 2010.

Appendix

Extremely halophilic microbial communities in anaerobic sediments from a solar saltern

Arantxa López-López, Pablo Yarza, Michael Richter, Ana Suárez-Suárez, Josefa Antón, Helge Niemann and Ramón Rosselló-Móra

The prokaryotic communities inhabiting hypersaline sediments underlying a crystallizer pond of a Mediterranean solar saltern have been studied in a polyphasic approach including 16S rRNA and *dsrAB* gene libraries analysis [the last encoding for dissimilatory (bi)sulfite reductase], most probable number of cultivable counts, and metabolic measurements of sulfate reduction. The samples studied here represent one of the most hypersaline anoxic environments sampled worldwide that harbour a highly diverse microbial community different from those previously reported in other hypersaline sediments. Both bacterial and archaeal types are present but, contrarily to the overlying brine system, the former dominates. Molecular analyses indicated that the bacterial fraction is highly diverse and mostly composed by groups related to sulfate-reducing bacteria (SRB). In good agreement with this, sulfate-reducing activity was detected in the sediment, as well as the metabolic diversity within SRB (as indicated by the use of different electron donors in enrichments). On the other hand, the archaeal fraction was phylogenetically homogeneous and, surprisingly, strongly affiliated with the MBSl-1 candidate division, an euryarchaeotal group only reported in deep-sea hypersaline anoxic basins of the Western Mediterranean, for which a methanogenic metabolism was hypothesized. The hypersaline studied samples constitute a valuable source of new prokaryotic types with metabolisms adapted to the prevalent *in situ* extreme conditions.

Published. Environmental Microbiology Reports (2010) 2: 258–271.
DOI 10.1111/j.1758-2229.2009.00108.x

Occurrence of *Halococcus* spp. in the nostrils salt glands of the seabird *Calonectris diomedea*.

Jocelyn Brito-Echeverrıa, Arantxa Lopez-Lopez, Pablo Yarza, Josefa Anton, Ramon Rossello-Mora

The nostrils of the seabird *Calonectris diomedea* are endowed with a salt-excreting gland that could produce a suitable environment for the colonization of extreme halophilic prokaryotes. We have studied in this organ the presence of extreme halophiles by means of culturing techniques. We could easily cultivate members of haloarchaea, and all cultures studied were identified as members of one of the two species *Halococcus morrhuae* and *Hcc. dombrowskii*. In order to reveal the diversity of these colonizers, we undertook a taxonomic study. Altogether, the results indicated that members of the genus *Halococcus* may constitute a part of the natural epizootic microbiota of *C. diomedea*, and that they exhibit such an important degree of taxonomic variability that appeals for a pragmatic species definition. This seabird nests in the west Mediterranean coasts, but its migratory habits, reaching locations as distant from the Mediterranean as the South Atlantic, may help in the dispersal mechanisms of haloarchaea through the Earth's surface.

Published. Extremophiles (2009) 13:557–565 . DOI 10.1007/s00792-009-0238-2

Pseudomonas arsenicoxydans sp nov., an arsenite-oxidizing strain isolated from the Atacama desert

Victor L. Camposa, Cristian Valenzuela, Pablo Yarza, Peter Kämpfer , Roberto Vidal, C. Zaror, Maria-Angelica Mondaca, Arantxa Lopez-Lopez, Ramon Rosselló-Móra

A Gram-negative, arsenite-oxidizing bacterial strain, designated VC-1, was isolated from sediment samples from the Camarones Valley in the Atacama Desert, Chile. Strain VC-1 was strictly aerobic, oxidase and catalase positive, rod shaped, of about 5.5 μm in length and 0.5–1.0 μm in diameter. It was motile by means of multiple polar flagella. The phylogenetic reconstruction of the 16S rRNA gene sequence, an MLSA study by concatenating six genes, and DDH studies indicated that the strain differed genotypically from its closest relatives and was therefore recognized as a new species within the genus *Pseudomonas*. Phenotypic analysis combining metabolic tests, fatty acid profiles and MALDI-TOF profiles of total cell extracts supported the classification of the new species for which we propose the designation *Pseudomonas arsenicoxydans* sp. nov. The type strain is accessible under the culture collection numbers CCUG 58201^{T} and CECT 7543^{T}.

Published. Systematic and Applied Microbiology (2010).
DOI 10.1016/j.syapm.2010.02.007

Evaluation of the 18S rRNA clone library approach to study the diversity of the macroeukaryotic leaf-epiphytic community of the seagrass *Posidonia oceanica* (L.) Delile.

F. J. Medina-Pons, J. Terrados, A. López-López, P. Yarza, R. Rosselló-Móra

The sequence comparisons among genes codifying for the RNA component of the small ribosomal subunit (16S rRNA or 18S rRNA) in cellular organisms have been largely used to reconstruct their phylogenies, and hence the identification of taxa by means of a molecular approach. Furthermore, the direct DNA isolation from environmental samples and the PCR amplification of the pool of rRNA genes with the subsequent cloning and sequencing have opened the door to the description of naturally occurring microbial communities independently from any culturing technique or morphological identification. These studies have unveiled an enormous hidden diversity in a wide variety of microbial communities. Our main objective was to evaluate the usefulness of the 18S rRNA gene clone libraries to describe the structure of the macroeukaryotic leaf-epiphytic assemblage of the seagrass *Posidonia oceanica*, and monitor the changes occurring in different stages of its seasonal succession (winter, spring and summer). To that end, we compared the results of these libraries with those provided by classical microscopy techniques. Among both approaches, the screening of clone libraries rendered the highest number of distinct units named operational phylogenetic units. However, diversity estimates provided by both methods were comparable and rendered the highest Shannon Diversity Index (*H'*) at the end of the succession. The major discrepancies were on the different occurrence of some groups. For example, macroalgae were the most frequent epiphytes counted by microscopy, whereas metazoa (specially, bryozoa) dominated the clone libraries. Altogether the results indicate that clone libraries constitute an excellent complementary approach to classical microscopy methods. To the best of our knowledge, this is the first attempt to describe a marine macro- eukaryotic community using a molecular approach such as the analysis of 18S rRNA gene clone libraries.

Published. Marine Biology (2009) 156:1963–1976. DOI 10.1007/s00227-009-1221-2

Complete genome sequence of *Marinobacter adhaerens* type strain (HP15), a diatom-interacting marine microorganism

Astrid Gärdes, Eva Kaeppel, Aamir Shehzad, Shalin Seebah, Hanno Teeling, Pablo Yarza, Frank Oliver Glöckner, Hans-Peter Grossart and Matthias S. Ullrich

Marinobacter adhaerens HP15 is the type strain of a newly identified marine species, which is phylogenetically related to *M. flavimaris, M. algicola*, and *M. aquaeolei*. It is of special in-terest for research on marine aggregate formation because it showed specific attachment to diatom cells. *In vitro* it led to exopolymer formation and aggregation of these algal cells to form marine snow particles. *M. adhaerens* HP15 is a free-living, motile, rod-shaped, Gram-negative gammaproteobacterium, which was originally isolated from marine particles sam-pled in the German Wadden Sea. *M. adhaerens* HP15 grows heterotrophically on various media, is easy to access genetically, and serves as a model organism to investigate the cellu-lar and molecular interactions with the diatom *Thalassiosira weissflogii*. Here we describe the complete and annotated genome sequence of *M. adhaerens* HP15 as well as some details on flagella-associated genes. *M. adhaerens* HP15 possesses three replicons; the chromosome comprises 4,422,725 bp and codes for 4,180 protein-coding genes, 51 tRNAs and three rRNA operons, while the two circular plasmids are -187 kb and -42 kb in size and contain 178 and 52 protein-coding genes, respectively.

Published. Standards in Genomic Sciences (2010) 3:97-107. DOI 10.4056/sigs.922139

Taxonomic and functional metagenomic profiling of the microbial community in the anoxic sediment of a sub-saline shallow lake (Laguna de Carrizo – Central Spain)

María-Eugenia Guazzaroni*, Michael Richter*, Adela García-Salamanca*, Pablo Yarza*, Ana Suárez-Suárez, Jennifer Solano, Pieter van Dillewijn, M. Antonia Molina-Henares, Nieves López-Cortés, Yamal Al-Ramahi, Carmen Guerrero, Alejandro Acosta, Laura I. de Eugenio, Virginia Martínez, Silvia Marques, Fernando Rojo, Eduardo Santero, Olga Genilloud, Julian Pérez-Pérez, Ramón Rosselló-Móra, Juan Luis Ramos and Manuel Ferrer

* these authors contributed equally to the work

The phylogenetic and functional structure of the microbial community residing in the anoxic sediment of a sub-saline shallow lake (Laguna de Carrizo) was estimated by analyzing the diversity in 16S rRNA cloned genes and 2.6 Mb of consensus metagenome sequence. 16S rRNA sequences revealed a diverse community with about 22% of the bacterial rRNAs being less than 94.5% similar to any rRNA currently in Genbank and 79% of the archaeal rRNAs mostly related to uncultivated *Euryarchaeota* of CCA47 group. Accordingly, 26% of the open reading frames had no hits in the environmental database. Interestingly, compared to other aquatic systems, Carrizo shallow contains an unusual and versatile range of (specialized) prokaryotes and of genes (some of them clustered) coding enzymes for the assimilation cycle of sulfur (serine, cysteine, SO_4^{2-}, $SO_3^{2-,}$ S^{2-}, S_n^{2-}, H_2S, HSO_{3-}, SPO_{7-}, sulfate esters and sulfolactate) and nitrogen (N_2, NH_{4+}, NO_{3-}, R-NO_2, HC≡N, R-C≡N, and R-NO_2-R) containing molecules, the use of (ubi)quinone as an electron carrier and for he synthesis and utilization of a wide range of storage polymers. A number of antibiotic gene transfer and phage-like insertion events associated with nitrogen and sulfur assimilation, suggesting their functional importance in the community. To the best of our knowledge, this is the first metagenome analysis of the anoxic sediment of a shallow lake originated by human acivity, namely, $CaSO_4$ – H_2O extraction.

Submitted to Environmental microbiology (December 2009)

Response of sulphate-reducing bacteria inhabiting a calcareous sandy sediment to hydrocarbon contamination

Ana Suárez-Suárez, Arantxa López-López, Antonio Tovar-Sánchez, Pablo Yarza, Alejandro Orfila, Jorge Terrados, Julia Arnds, Silvia Marqués, Helge Niemann, Philippe Schmitt-Kopplin, Rudolf Amann and Ramón Rosselló-Móra.

In situ mesocosm experiments using a calcareous sandflat from the coastal areas of Mallorca in the Mediterranean Sea were performed to study the sulphate-reducing bacterial (SRB) response to a controlled crude oil spill or a heavy contamination with naphthalene. The changes in the community due to the contamination were monitored by a combination of molecular techniques, culturing approaches, and metabolic activity rates. The results showed that both, crude oil and naphthalene promoted a negative effect onto the total microbial population as the natural increase in cell numbers due to the seasonal dynamics was attenuated. However, both contaminants enhanced the sulfate reduction rates as well as the culturable fraction of SRB. In addition the results suggested the presence of an autochthonous microbiota, related to the deltaproteobacterial families *Desulfobacteraceae*, *Desulfobulbaceae* and *Desulfovibrionaceae,* able to face the potential contamination with oil or PAHs as naphthalene.

Submitted to Environmental Microbiology (December 2009).

Supplementary materials

For practical reasons, all supplementary tables, figures, files and other materials referenced in the text have been included in the attached CD-rom. Its contents are detailed below.

CD - contents

/S1	Supplementary materials for chapter 1
/S2	Supplementary materials for chapter 2
/S3	Supplementary materials for chapter 3
/S4	Supplementary materials for chapter 4
/LTPs93	Complete LTP release s93 (September 2008)
/LTPs95	Complete LTP release s95 (September 2008)
/LTPs100	Complete LTP release s100 (September 2008)
/LTPs102	Complete LTP release s102 (September 2008)