Available online at www.sciencedirect.com
ScienceDirect
Systematic and Applied Microbiology 31 (2008) 251–257
SYSTEMATIC AND APPLIED MICROBIOLOGY
www.elsevier.de/syapm

A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes

Jörg Peplies[a], Renzo Kottmann[b,d], Wolfgang Ludwig[c], Frank Oliver Glöckner[b,d,*]

[a]*Ribocon GmbH, D-28359 Bremen, Germany*
[b]*Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany*
[c]*Lehrstuhl für Mikrobiologie, Technische Universität München, D-85350 Freising, Germany*
[d]*Jacobs University Bremen gGmbH, D-28759 Bremen, Germany*

Abstract

Phylogenetic analysis is currently used worldwide for taxonomic classification and identification of microorganisms. However, despite the countless trees that have been reconstructed and published in recent decades, so far, no user-friendly compilation of recommendations to standardize the data analysis and tree reconstruction process has been published. Consequently, this standard operating procedure for phylogenetic inference (SOPPI) offers a helping hand for working through the process from sampling in the field to phylogenetic tree reconstruction and publication. It is not meant to be authoritative or comprehensive, but should help to make phylogenetic inference and diversity analysis more reliable and comparable between different laboratories. It is mainly focused on using the ribosomal RNA as a universal phylogenetic marker, but the principles and recommendations can be applied to any valid marker gene. Feedback and suggestions from the scientific community are welcome in order to improve these guidelines further. Any updates will be made available on the SILVA webpage at http://www.arb-silva.de/projects/soppi.

Keywords: Phylogenetic analysis; Tree reconstruction; Ribosomal RNA; Alignment; Standardization

Introduction

Phylogenetic inference using molecular data has become a standard procedure to indentify and classify (micro)organisms. In the best case, it can even be assumed that a phylogenetic tree provides some information about the history and the ancestors of our currently existing species. It was Carl Woese, a physicist working for General Electrics, who paved the way for stable evolutionary-based systematics using comparative sequence analysis of ribosomal RNA (rRNA) molecules [18]. He propagated the idea Zuckerkandl and Pauling came up with in 1965 [19], that macromolecules, such as DNA or proteins, can be used as molecular clocks for phylogenetic inference. Thirty years have passed since Carl Woese proposed the three primary domains of life based on the phylogenetic analysis of rRNA genes. Despite ongoing discussions about the validity of the concept, the introduction of rRNA as a universal molecular marker (the ''gold standard'') has transformed microbiology to its roots. Cultivation-independent investigations have reported an immense array of completely unexpected microbial diversity in the environment [17]. Today, the rRNA provides the largest collection of any single molecular marker. Currently,

*Corresponding author at: Max Planck Institute for Marine Microbiology Celsiusstr. 1, D-28359 Bremen, Germany. Tel.: +49 421 2028970; fax: +49 421 2028580.
E-mail address: fog@mpi-bremen.de (F.O. Glöckner).

0723-2020/$ - see front matter © 2008 Elsevier GmbH. All rights reserved.
doi:10.1016/j.syapm.2008.08.003

more than 600,000 small (SSU) and 100,000 large (LSU) subunit rRNA variants are available in public databases, such as SILVA [14] or RDP (only SSU) [2], with the vast majority stemming from uncultured bacteria.

This immense amount of available sequence information puts pressure on the individual researcher when starting to reconstruct phylogenetic trees. Although powerful software tools for DNA sequence analysis, such as the ARB package [13], are available, the question that has arisen over and over again at workshops and conferences has been "Can't we get some guidelines to help us navigate through the sequence and tree space?" Consequently, like a protocol for laboratory experiments, this 'standard operating procedure for phylogenetic inference' (SOPPI) should help biologists to improve their workflow when working with phylogenetic marker genes. It reflects the opinion of the authors about the different steps that need to be taken into account when dealing with molecular data (mainly rRNA genes) for phylogenetic inference and tree reconstruction. However, for detailed information about the theory of phylogenetic analysis, please refer to Swofford et al. [16] and Felsenstein [7].

Sampling

Closely connected to sampling is the recording of contextual data for sequences, such as information on the sampling sites or hosts required for later data integration and interpretation [5]. Only an integrated view of our sequence collections will finally turn molecular data into biological knowledge. However, only a minority of researchers takes this into account when submitting sequence data to the public repositories. Therefore, please make sure that a minimal amount of contextual (meta)data is not only reported in the paper, but also deposited in the databases for every culture or sample devoted to sequencing of phylogenetic marker genes. For environmental samples, at least the GPS position (latitude, longitude) depth/altitude and time of sampling should be made routinely available for every (rRNA) sequence. An extended list of suggested contextual data can be found at http://www.arb-silva.de/projects/contextual-data. More information about the collaborative effort headed by the Genomics Standards Consortium to enrich the contextual data of our genome and metagenome collection, as well as environmental marker genes like rRNA, is available at http://gensc.org [9].

Sequencing

Good phylogenetic inference can only be undertaken based on high quality, nearly full-length sequences because the information content of all genetic markers is limited. For any in-depth phylogenetic rRNA analysis, a minimum length of 1200 and 1900 bases is highly recommended for 16S/18S and 23S/28S, respectively. Please avoid reporting unresolved (ambiguous) bases by rechecking the original chromatogram from the sequencing device. If this does not help, for instance, due to operon heterogeneities, unresolved bases must be handled using the standard international union of pure and applied chemistry (IUPAC) nucleotide code (see http://www.bioinformatics.org/SMS/iupac.html). Finally, do not fill up sequencing gaps with N's. Sequence quality should also be verified by taking into account the alignment that can indicate sequencing errors by considering the secondary structure and the positional variability of the molecule. Several tools exist to assist in raw sequence analysis and assembly. Examples are Sequencher (http://www.genecodes.com) and RNA-Baser (http://www.rnabaser.com). For the verification of sequence quality using secondary structure information we recommend the alignment editor of the free software package ARB (http://www.arb-home.de). Sequences between 300 and 1000 bases can be used for phylogenetic classification, but the results need to be carefully checked taking into account the reduced information content. Partial rRNA sequences shorter than 300 bases are not suited for phylogenetic inference and the only recommendation that can be given in this case is to invest additional efforts to elongate the sequences.

Alignment

Similar to high-quality sequences, a high-quality alignment is a prerequisite for any phylogenetic inference. It assures that every column represents only orthologous bases that have evolved from a common ancestor. The impact of the underlying alignment on the final tree quality should never be underestimated [12].

For rRNA alignments, it is highly recommended to take the secondary structure of the molecule into account. Orthology of the bases in the alignment can be assumed if base changes have been complemented by corresponding exchanges on both sides of the helices (Fig. 1). Experts in the field have dedicated years of their life establishing comprehensive rRNA alignments. For *Bacteria*, you can choose between alignments from SILVA (http://www.arb-silva.de, [14]) and RDP II (http://rdp.cme.msu.edu, [2]). For *Archaea* and *Eukarya*, pre-aligned sequences are only offered by the SILVA databases.

For rRNA, it is not recommended to use a subset of sequences and establish an alignment from scratch using multiple alignment tools like Clustal [1], Mafft [11] or Muscle [3].

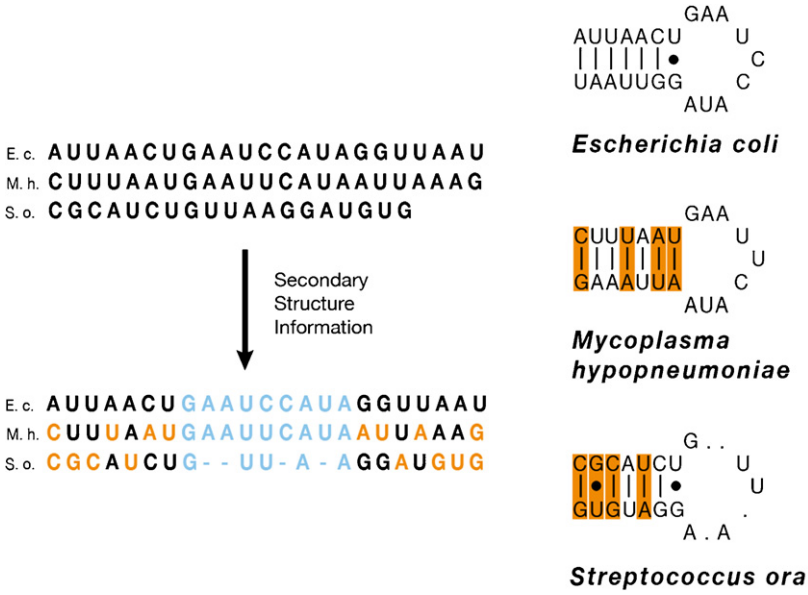

Fig. 1. The figure gives an example of how the secondary structure information of the ribosomal RNA can be used to gain confidence for positional orthology in the alignment columns. On the left side, the raw and aligned sequences for *Escherichia coli* (E. c.), *Mycoplasma hypopneumoniae* (M. h.) and *Streptococcus oralis* (S. o.) are shown. The right panel shows the respective secondary structures. The bases marked in orange indicate the corresponding base exchanges to maintain the helix structure. Dashes indicate canonical, and dots non-canonical base pairings. Especially in the case of *S. oralis*, the consideration of the pairing base pairs in the stem finally allows a reasonable alignment to be built, despite the deletions. Since the exact secondary structure of the 16S rRNA has only been determined for a few reference sequences, the *E. coli* 16S rRNA secondary structure is normally used as a model (framework) for all other bacterial sequences.

The reasons are:

1. Establishing an alignment based on the primary and secondary structure is a tedious job and it will be hard to achieve the already established standards, especially for the rRNA genes.
2. Identity values (matrices) cannot be directly compared if they are based on different alignments.
3. Resulting trees cannot be directly compared if they are based on different alignments.
4. Any alignment optimization can be conserved if you follow one of the existing alignments, for instance, the SILVA database will adopt aligned sequences for incorporation in the reference alignment (Seed) and in the next releases all related sequences will be aligned accordingly.
5. Deposition and exchange of alignments will be facilitated when adhering to existing standards.

Nevertheless, if you decide to build your own alignment, for instance, because no standard alignment is available for your current gene of interest, it is mandatory to document the alignment process exactly. This means you have to describe which method or tool (including the version of the tool) has been used and which parameters have been applied (e.g. how gaps are treated). Depending on the tool or parameters used, the resulting alignments can differ significantly [4] and this directly influences your phylogenetic or statistical analysis. Whatever you do, please take into account that the scientific community must be able to reproduce your results.

In any case, like the sequence information, the alignment needs to be made publicly available in an electronic format for evaluation and tree reproduction (ideally, this should be a multi-FASTA or ARB file including all contextual data). Please do not use cryptic headers containing your personal identifiers. Instead, use the accession number with the start and stop position to make the sequence entry unique. This is mandatory for sequences extracted from (meta)genomes. If the entry is unambiguously described by the accession number, start and stop positions can be omitted. Below is an example for an appropriate header providing the accession number, start, stop, length, and the organism name.

>AF12345678.1.1542 1542 bp *Ultrabacterium headerii*

The corresponding export filters for the ARB software suite are available at www.arb-silva.de/download/archive/imp_exp_filters.

Tree reconstruction

First of all, it is important to realize that there is no easy-to-follow standard recipe of how to produce a

publication-ready tree. The reason is, that the ultimate tree will never exist due to noise in the input data (caused by sequencing or alignment errors), limited resolution power (information content) of the genetic markers and, finally, the extremely large number of potential tree topologies (the "tree space"). Even for small subsets of sequences, not all trees can be evaluated to find the best one according to the model of evolution and tree reconstruction methods used. Testing the 2.8×10^{74} available topologies for only 50 sequences (Table 1) will take longer than the lifetime of a human being. To be able to reconstruct a tree in reasonable time, simplified evolutionary models and the application of heuristics (only a subset of tree topologies is evaluated) are accepted. Nevertheless, as a consequence, the optimal tree topology might escape from the analysis and only a suboptimal topology is reported by the algorithm.

In general, the topology of the reconstructed tree is influenced by the following factors:

1. Quality of the sequences (length, ambiguities, homopolymers)
2. Quality of the alignment (taking into account secondary structure information)
3. Amount and selection of sequences (full length, representative)
4. Amount and selection of alignment positions used for tree reconstruction (application of different filters)
5. The tree reconstruction method and parameters used.

Only if all of these parameters are taken into account and the tree topology is carefully evaluated during the analysis, a high-quality tree can finally be published.

Generally, in-depth phylogenetic analysis should only be performed with nearly full-length sequences (see above as well). The reason is that the information content of standard phylogenetic markers is already limited by its size and allowed character states per position. A further reduction of information will lead to misassignments and unstable topologies [12].

Table 1. Total number of unrooted, bifurcated trees (the "tree space") which theoretically has to be evaluated in a maximum parsimony or maximum likelihood approach to find the optimal tree, depending on the number of sequences used

Number of sequences	Number of trees
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
50	2.8×10^{74}

Calculated according to [6]. For only 50 sequences, the total number of possible trees is nearly in the range of the estimated number of atoms in our universe and trees are normally calculated using many more sequences.

Phylogenetic tree reconstruction represents nothing more than a test of the stability of the tree candidates. To do this, several tree reconstruction methods based on different algorithms need to be applied and the results compared. The three commonly used approaches are based on distance matrix (DM) (e.g. neighbour joining), maximum parsimony (MP) and maximum likelihood (ML) methods. All of these approaches have their advantages and disadvantages and a detailed explanation is beyond the scope of this SOP. To get a basic understanding of the methods and underlying evolutionary models, the book chapter "Phylogenetic Inference" by Swofford et al. [16] is recommended. At the moment, ML is regarded as the most advanced method for phylogenetic inference based on the cost of high computational demands. Nevertheless, tools like PhyML [10] or RAxML [15] provide fast ML implementations which even allow bootstrapping to be considered for ML.

Under the hypothetical assumption that the quality of the sequences and the alignment is in its best state, the amount and selection of sequences can still have a significant influence on the topology of the resulting tree(s). The typical problem that shows up in tree reconstruction is unstable branching patterns, although moving branches or species can often be stabilized by the addition of further reference sequences. This usually helps to reduce branch attraction effects caused by false identities resulting from multiple base changes during the course of evolution. Therefore, you should always go for the maximum number of sequences that can be handled within a given amount of time and computational resources. Even on a normal dual core PC, a tree with 5000 sequences can be calculated with ML in less than 2 days.

Filtering of input data (removal of selected alignment positions) is necessary to improve the signal-to-noise ratio and is an appropriate method to test the topology of the tree by altering the underlying dataset. Nevertheless, care should be taken to find the optimal threshold when removing complete alignment columns for the tree reconstruction process. Noise is superimposed on the dataset mainly by sequencing errors, false identities and a suboptimal alignment which are typically found in highly variable regions and should therefore be excluded from the calculations. However, it remains to the investigator to identify the grade of variability that should be used as a threshold for filtering, although a 50% positional conservatory filter based on the sequence identity of a specific phylogenetic group is definitely a good starting point for rRNA

genes. Such a filter will exclude the complete alignment column, if the frequency of the most abundant nucleotide is below 50%. Calculation (e.g. with the software package ARB) and testing of additional filters, such as 30% and 40% filters, are recommended in order to get a feeling for the stability of the topology. For protein genes, we recommend a 30% filter. For maximum accuracy, positional conservatory filters should always be calculated using all sequences of the corresponding phylum of your group of interest, even if your tree will finally only be calculated on a subset of sequences out of the group.

Tree reconstruction should be performed with different methods, parameters and filters, and the resulting topologies need to be compared to find out which part of the tree is stable and which one is not. This is not an easy task because no automatic method exists. A very basic possibility is to print out the different trees and mark inconsistencies to get a visual impression of the stability of the phylogenetic branchings. ARB users can also ''mark'' selected clusters in a tree and by switching between alternative trees it can be easily checked if the sequences under investigation show the same branching order.

Finally, how to deal with partial sequences? The worst-case scenario is to truncate full-length sequences to fit the length of the shortest sequences. If partial sequences need to be part of a tree, and extension by additional sequencing is not possible, incremental adding procedures, as offered by the ARB Parsimony system, can be used. In this case, the tree needs to be built first on full-length sequences, using MP, ML or DM, and the partial sequences are later added without allowing changes of the overall tree topology.

Tree reconstruction can be a very time consuming and computationally intensive task. If your laboratory is not prepared to deal with large datasets and different phylogenetic tools, professional assistance in publication-ready tree reconstruction and training is offered by the company Ribocon (www.ribocon.com).

Presenting the tree

For the presentation or publication of an estimated phylogeny, the tree which is considered to be the ''best'' (e.g. the ML tree) should be taken as the template for the introduction of multifurcations, in case, the respective branching pattern is not unambiguously supported by the different approaches [12] despite of a bifurcation. A multifurcation represents a separation from a single entity to three or even more entities. Or, in other words, the exact evolutionary history in this part of the tree remains unresolved. In the case where bootstrapping is undertaken, the respective values can be added to the unambiguously resolved branches. It is well accepted to finally show just a (small) subset of the sequences in the paper, as long as it is properly documented. This is a valid trade-off between the clarity of presentation (and the limited space of a journal page) and the accuracy of the topology, which can often only be reached with a larger dataset. To make comparisons and reproducibility of the tree easier, please avoid cryptic names on the tip of the branches. Use the accession number and the name of the organism (if available) for identification. For sequences from uncultivated organisms, try to show entries in your final tree which at least provide some additional information on the origin (e.g. ''marine surface water clone'' instead of ''uncultivated organism''). Nevertheless, again, for your calculations you can also use high-quality ''unannotated'' sequences. The quality of your tree will be better the more sequences you use. In case your tree consists of sequences which significantly differ in length, the number of bases should also be provided. If you show several trees, organize them in a similar way to facilitate easy comparisons of the topologies. If you define new clusters or groups, indicate them in all trees with brackets and the corresponding cluster names.

Documentation

Irrespective of what has been carried out, the most important point is always to document accurately all steps undertaken to create the presented tree. The minimal information required to evaluate your results are the method(s), tool(s) and number of sequences that have been used for alignment and tree reconstruction, the evolutionary model (e.g. GTR, F84, JC) and the parameters (e.g. gamma distribution, filters) that have been applied and, very importantly, the number of valid columns the tree is based on (due to filtering). If multifurcations have been introduced, this needs to be clearly stated in the results and discussion, as well as in the figure legends.

Sequence submission and access to sequences

Sequences need to be submitted to one of the partners of the international nucleotide sequence database collaboration (INSDC, http://www.insdc.org), comprising DDBJ, EMBL and Genbank, to make them publicly available.

Please do not forget to add as much contextual data as possible (see paragraph on sampling) when submitting the sequences. Unfortunately, not all desired parameters can currently be deposited in the INSDC databases. Nevertheless, it is worthwhile to store them consistently, for instance, in your personal ARB/SILVA

database, and send them to the "note" field of the INSDC databases. Respective export filters for sequence and contextual data will be made available for the ARB package, therefore, please check www.arb-silva.de for forthcoming information.

Later, after acceptance of your paper, all sequences used for analysis must be immediately made publicly available via the INSDC databases. The respective accession numbers need to be clearly stated in the paper. For the review process, all sequences that are at this stage not yet available from the INSDC databases need to be sent, together with the alignments, in multi-FASTA format to the Editor of the journal considered for publication.

Example for materials and methods

"Sequences have been analysed using the ARB software package (version December 2007) [13] and the corresponding SILVA SSURef 94 database [14]. After importing, all sequences were automatically aligned according to the SILVA SSU reference alignment. Manual refinement of the alignment was carried out taking into account the secondary structure information of the rRNA. Tree reconstruction was performed with up to 1000 sequences using the neighbour joining (ARB), MP (DNAPars v1.8, [8]) and ML (RAxML v7.04, [15]) methods. Tree topology was further tested by the application of 30%, 40% and 50% positional conservatory filters. The final tree was calculated with 500 sequences based on 1280 valid columns (50% conservatory filtering) with RAxML (model: GTRGAMMA). Partial sequences have been added to the tree using the ARB parsimony tool. Multifurcations have been manually introduced in the case where tree topology could not be unambiguously resolved based on the different treeing methods and the underlying dataset. For better clarity, only selected subsets of the sequences used for treeing are shown in the figure.

The respective alignments are available in multi-FASTA format and as an ARB file at ftp.xyz.de/data. All sequences described in the paper are available from the databases of the INSDC, comprising DDBJ, EMBL and Genbank, under EX123456, EX123458–EX123470."

Acknowledgements

We thank all speakers and participants of the International Workshop on rRNA Technology, April 7–9, 2008, in Bremen, Germany for providing the final spark to compile these guidelines. This study was supported by the Max Planck Society and Ribocon GmbH.

References

[1] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, J.D. Thompson, Multiple sequence alignment with the Clustal series of programs, Nucleic Acid Res. 31 (2003) 3497–3500.
[2] J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, A.M. Bandela, E. Cardenas, G.M. Garrity, J.M. Tiedje, The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data, Nucleic Acid Res. 35 (2007) D169–172.
[3] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 1–19.
[4] R.C. Edgar, S. Batzoglou, Multiple sequence alignment, Curr. Opin. Struct. Biol. 16 (2006) 368–373.
[5] N. Editorial, A place for everything, Nature 453 (2008), (2–2).
[6] J. Felsenstein, Number of evolutionary trees, Syst. Zool. 27 (1978) 27–33.
[7] J. Felsenstein, Inferring Phylogenies, Sinauer Associates Inc., Sunderland, MA, 2004.
[8] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.67, Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2005.
[9] D. Field, G. Garrity, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, T. Gray, M. Ashburner, S. Baldauf, J. Boore, G. Cochrane, J. Cole, C. dePamphilis, R. Edwards, N. Faruque, R. Feldmann, F.O. Glöckner, et al., Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification, Nat. Biotechnol. 26 (2008) 541–547.
[10] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[11] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acid Res. 30 (2002) 3059–3066.
[12] W. Ludwig, H.P. Klenk, A phylogenetic backbone and taxonomic framework for prokaryotic systematics, in: D.R. Boone, R.W. Castenholz (Eds.), The *Archaea* and the Deeply Branching and Phototrophic *Bacteria*, Springer, New York, 2001, pp. 49–65.
[13] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A.W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, K.H. Schleifer, ARB: a software environment for sequence data, Nucleic Acid Res. 32 (2004) 1363–1371.

[14] E. Pruesse, C. Quast, K. Knittel, B.M. Fuchs, W.G. Ludwig, J. Peplies, F.O. Glöckner, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB, Nucleic Acid Res. 35 (2007) 7188–7196.
[15] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690.
[16] D.L. Swofford, G.J. Olsen, P.J. Waddel, D.M. Hillis, Phylogenetic inference, in: D.M. Hillis, C. Moritz, B.K. Marble (Eds.), Molecular Systematics, second ed., Sinauer Associates Inc., Sunderland, MA, 1996, pp. 407–514.
[17] V. Torsvik, J. Goksoyr, F.L. Daae, High diversity in DNA of soil bacteria, Appl. Environ. Microbiol. 56 (1990) 782–787.
[18] C.R. Woese, E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms, Proc. Natl. Acad. Sci. USA 74 (1977) 5088–5090.
[19] E. Zuckerkandl, L. Pauling, Molecules as documents of evolutionary history, J. Theor. Biol. 8 (1965) 357–366.