VOL.8 NO.2 September 2019 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Annotation of Bacteria by Greengenes Classifier Using 16S rRNA Gene Hyper Variable Regions

M. Can
Faculty of Engineering and Natural Sciences,
International University of Sarajevo International University of Sarajevo,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
mcan@ius.edu.ba

Article Info

Article history:
Article received on 17 June 2019
Received in revised form 15 August 2019

Keywords:
16S ribosomal RNA; gene segments; diagnosis; bacteria annotation

ABSTRACT: rRNA-genes for phylogenetic classifications started to be used in 1980s first time by Carl Woese which made a ground breaking contribution to microbiome science. rRNA-genes are used to explore microbial diversity as well as a method for bacterial annotation. Many researchers followed rRNA-based analysis track as a central method in microbiology. Similarity based analyses use several new generations of Artificial Neural Networks to create classifiers against bacteria libraries to obtain high accuracies. By the time, the number of bacteria in these libraries increased enormously. In this article the accuracy of a classifier against Greenges library is tested. It has been shown by the author in previous articles that the Greengenes Classifier can be successfully used as a bioinformatics program that performs taxonomic classification of 16S rRNA gene sequences. In a previous article, the accuracy of the program is also tested when it is applied to common PCR products of the 16S rRNA variable regions, which are the only product of laboratories in microbiome projects. In this study, V1–V3 hyper variable regions from 16S rRNA genes of some known bacteria is taken from the work of A. Cosic. In this article we used Longest Common Subsequence similarity measure to classify bacterial 16S rRNA gene sequence short reads against the Greengenes library.

1. INTRODUCTION

Although some bacteria, produce antibiotics; others live symbiotically in the guts of animals including humans, or elsewhere in their bodies, or on the roots of certain plants, bacteria are often found responsible from the human and animal diseases. Helping the breakdown of dead organic matter; they make up the base of the food chain in many environments. Because of their extreme flexibility, capacity for rapid reproduction and growth, and contribution to the processes in the body of humans, and all living creatures, bacteria are of such immense importance in the life on the earth.

Through their activities in the soil, bacteria also contribute immensely to global energy conversion and the recycling of matter. Therefore understanding their life cycles, profiling the microbial community in their quality and quantity are the most important tasks for microbiologists to explore various ecosystems. Only a few percent of bacteria can be cultured or isolated under laboratory conditions (Ash et. al., 1991). For this reason, our understanding of the kingdom of Bacteria remains limited. FISH, fluorescent situ hybridization (Brown, 1999), DGGE, Denaturing gradient gel electrophoresis, (Audic, and. Claverie, 1997), T-RFLP, Terminal restriction fragment length polymorphism

(Benson, et. al., 2000), and Genechips (Bruno, et. al., 2000) were used in the past few decades as mainstream methods in studies of bacterial communities and diversity, until the development of high-throughput sequencing technology. Recently, meta-genomic methods provided by next-generation sequencing technology such as Roche 454 (Cannone, et., al., 2002) and Illumina (Cole, et, al, 2007) have facilitated a remarkable expansion of our knowledge regarding uncultured bacteria (Yang et., a., 2016).

A Brief History of Bacterial Classifications

The genus Bacterium was a taxon described in 1828 by Christian Gottfried Ehrenberg (Ehrenberg, 1828). Ehrenberg also described spiral shaped bacteria Spirillum, in 1832 (Ehrenberg, 1832). A genus of spore-forming rod shaped bacteria, Bacillus, in 1835, and thin spiral shaped bacteria, Spirochaeta, in 1835 (Ehrenberg, 1835).

A genus of comma shaped bacteria, Vibrio, first described in 1854 (Pacini, 1854). In the *Tree of Life* in Generelle Morphologie der Organismen (Haeckel, 1867), Ernst Haeckel, in the year 1866, defining the class Schizomycetes, first classified bacteria as plants. He placed the group in the phylum Moneres in the kingdom Protista and defined them as completely structureless and homogeneous organisms, consisting only of a piece of plasma.

Six genera, Micrococcus, Bacterium, Bacillus, Vibrio, Spirillum, and Spirochaeta (1872) and 4 tribes: Sphero-bacteria, Microbacteria, Desmobacteria, and Spirobacteria. distinguished by Ferdinand Cohn (Cohn, 1875) (Murray, and Holt, 2005), and this classification was influential throughout the nineteenth century.

Erwin F. Smith accepted 33 valid different names of bacterial genera and over 150 invalid names in 1905, (Smith 1905) and in 1913 Paul Vuillemin (Vuillemin, 1913) in a paper concluded that all species of the Bacteria should fall into the genera Planococcus, Streptococcus, Klebsiella, Merista, Planomerista, Neisseria, Sarcina, Planosarcina, Meta bacterium, Clostridium, Serratia, Bacterium and Spirillum.

Van Niel, (Stanier, and van Niel, 1941) recognized the Kingdom Monera with 2 phyla, Myxophyta and Schizomycetae. The phylum Schizomycetae comprising classes Eubacteriae with 3 orders, Myxobacteriae, 1 order, and Spiroch-etae, 1 order. Bisset (Bisset, 1962) distinguished 1 class and 4 orders: Eubacteriales, Actinomycetales, Strept-omycetales, and Flexibacteriales.

The most widely accepted system of its time was due to Migula, (Migula, 1897) which included all then-known species but was based only on morphology, contained the 3 basic groups, Coccaceae, Bacillaceae, and Spirillaceae but also Trichobacterinae for filamentous bacteria; Orla-Jensen (Orla-Jensen, 1909) established 2 orders: Cephalotrichinae, 7 families, and Peritrichinae, presumably with only 1 family. Bergey (Bergey et al 1925) presented a classification which generally followed the 1920 Final Report of the SAB, Society of American Bacteriologists Committee (Winslow et al, 1917), which divided the class Schizomycetes into 4 orders: Myxobacteriales, Thiobacteriales, Chlamydobacter-iales, and Eubacteriales, with a 5th group being 4 genera considered intermediate between bacteria and protozoans: Spirocheta, Cristospira, Saprospira, and Treponema.

Due to the lack of visible traits to follow, throughout classification history, different authors often reclassified the genera, in different ways. The resulted poor state is summarized in 1915 by Robert Earle Buchanan (Buchanan, 1916).

Relatively recently, in 1980s, Carl Woese brought a new technique to microbiology with his rRNA-based phylo-genetic classification (Woese, et. al, 1990). Today, rRNA-based analysis remains a central method in microbiology, used not only to explore microbial diversity but also as a method for bacterial annotation.

rRNA-based identification methods are conceptually easier to interpret than molecular phylogenetic analyses and are often preferred when the groups are well defined. While phylogenetic methods are clustering techniques, most rRNA classification methods, have been nearest-neighbor-based classification schemes (Maidak, et. al., 1994; DeSantis, et. al., 2003; Brown, 1999). In the past, this was due to the lack of a consistent, higher-level bacterial taxonomies. Several recent events have helped change this situation (Wang, et. al., 2007).

The 16S rRNA gene sequence first used in 1985 for phylogenetic analysis (Lane, et. al., 1985). Because it contains both highly conserved regions for primer design and hypervariable regions to identify phylogenetic characteristics of microorganisms, the 16S rRNA gene sequence became the most widely used marker gene for profiling bacterial communities (Tringe, and Hugenholtz, 2008). Full-length 16S rRNA gene sequences consist of nine hypervariable regions that are separated by nine highly conserved regions (Baker, et. al., 2003; Wang, and Qian, 2009). Limited by sequencing technology, the 16S rRNA gene sequences used in most studies are partial sequences (Yang, et. al, 2016).

2. TAXONOMIES

Microbiome sequencing analysis is mainly concerned with sequencing DNA from microorganisms living in certain environments without cultivating them in laboratory. In a typical taxonomy guided approach (Huson, et. al., 2016),

sequenced reads are first binned into taxonomic units and then the microbial composition of samples is analyzed and compared in detail.

The two main technical ingredients of taxonomic analysis are the reference taxonomy used and the binning approach employed. Binning is usually performed either by aligning reads against reference sequences (Pruesse, et., al., 2012) or using k-mer based techniques (Cole, et. al., 2014). Taxonomic binning of 16S reads is usually based on one of the five taxonomies:

- SILVA (Yilmaz, et. al., 2014),
- RDP (Wang, et. al., 2007),
- Greengenes (McDonald, et. al., 2012)
- NCBI (Federhen, 2012).
- Open Tree of Life Taxonomy (OTT) (Hinchliff, et. al., 2015).

There are inconsistencies of microbial classifications (Beiko, 2016), therefore the choice of reference taxonomy is important in research. In our study we have found that Greengenes is more consistent compared to the first two.

Taxonomic Classifications

Each of the five taxonomies that compared is based on a mixture of sources that have been compiled into taxonomies in different ways. They differ in both size and resolution as in Table 1.

Table1 Overview of five taxonomic classifications

Taxonomy	Type	modes	Lowest	Latest
SILVA	Manual	12,117	Species	2017
RDP	Semi	6,128	Genus	2016
Greengenes	Automatic	3,093	Species	2013
NCBI	Manual	1,522,150	Species	2017
OTT	Automatic	2,627,066	Species	2016

All taxonomies assign ranks to their nodes, the seven main ones being domain, phylum, class, order, family, genus and species. However, RDP only goes down to the genus level, but has two extra levels subclass and suborder, whereas SILVA, Greegenes, NCBI and OTT go down to the species level. In this paper, the taxonomy Greengenes is visited.

2.3 Greengenes (GG)

The Greengenes taxonomy (McDonald, et. al., 2012) is dedicated to Bacteria and Archaea. Classification is based on automatic de novo tree construction and rank mapping from other taxonomy sources (mainly NCBI). Phylogenetic tree is constructed from 16S rRNA sequences that have been obtained from public databases and passed a quality filtering. Sequences are aligned by their characters and secondary structure and then subjected to tree construction with Fast Tree (Price, et. al., 2009). Inner nodes are automatically assigned taxonomic ranks from NCBI supplemented with previous version of Greengenes taxonomy and CyanoDB (Komárek, et. al., 2016). We used a taxonomy associated with the Greengenes database as released on May 2013 with 198.510 bacteria. Although Greengenes is still included in some metagenomic analyses packages, for example QIIME (Caporaso, et. al., 2010), it has not been updated for the last three years.

Table 2. Levels and number of sublevels in Greengenes

Levels	# Sublevels
Phylum	86
Class	232
Order	366
Family	466
Genus	1949
Species	2389

2. MATERIALS AND METHODS

2.1 The Dataset

In Ćosić, and Jahjaefendic (2019) started a research by the following fourteen bacteria

1. Bacillus subtilis
2. Bifidobacterium bifidum
3. Bifidobacterium breve
4. Bifidobacterium infantis
5. Bifidobacterium longum
6. Lactobacillus acidophilus
7. Lactobacillus delbrueckii ssp.bulgaricus
8. Lactobacillus casei
9. Lactobacillus plantarum
10. Lactobacillus rhamnosus
11. Lactobacillus helveticus
12. Lactobacillus salivarius
13. Lactococcus lactis ssp. lactis
14. Streptococcus thermophilus

The taxonomic identities of these fourteen bacteria are as in Table 3.

Table 3.The taxonomic identities of the fourteen bacteria

	Phylum	Class	Order	Family	Genus	Species
1	**Firmicutes**	Bacilli	**Bacillales**	Bacillaceae	**Bacillus**	**subtilis**
2	**Actinobacteria**	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	bifidum
3	**Actinobacteria**	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	breve
4	**Firmicutes**	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	infantis
5	**Actinobacteria**	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	longum
6	**Proteobacteria**	Alphaproteobacteria	Rhizobiales	Methylocystaceae	Rhodoblastus	acidophilus
7	**Firmicutes**	Bacilli	Lactobacillales	Lactobacillaceae	**Lactobacillus**	**delbrueckii**
8	**Actinobacteria**	Actinobacteria	Actinomycetales	Brevibacteriaceae	Brevibacterium	casei
9	**Firmicutes**	Bacilli	Lactobacillales	**Lactobacillaceae**	**Lactobacillus**	**plantarum**
10	**Actinobacteria**	Actinobacteria	Actinomycetales	Kineosporiaceae	Kineosporia	rhamnosa
11	**Firmicutes**	Bacilli	Lactobacillales	Lactobacillaceae	**Lactobacillus**	**helveticus**
12	**Firmicutes**	Bacilli	Lactobacillales	Lactobacillaceae	**Lactobacillus**	**salivarius**
13	**Proteobacteria**	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Camelimonas	lactis
14	**Proteobacteria**	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rhodovulum	thermophilus

Bacterial Growth

The above bacteria are grown in a incubator at 37°C for 24h. Handling of the bacteria during the inoculation and enumeration was done in sterile environment Ćosić, and Jahjaefendic (2019).

DNA Isolation

In the same research by Ćosić, and Jahjaefendic (2019), bacterial DNA's are isolated. The primers are designed with the amplification targets of the whole genes where the point was to get as close as possible to 1500 bp which is the length of the 16S-rRNA gene, and hypervariable regions mainly the V3 and V4 region.

Sequencing

Amplified pieces of DNA's sent for sequencing. The length of sequences are as in the Table 4. Appendix A shows full sequences (Cosic 2019).

3. LONGEST COMMON SUBSEQUENCE TECHNIQUE

To annotate bacteria we define a similarity measure between the two bacteria first as the length of the longest common sub sequence of the two bacteria 16S gene sequences.

Next using a similarity measure, these sequences are annotated against the library Greengenes.

Table 4: Length of sequences

Sequence	Length bp	Sequence	Length bp
1	440	11	261
2	264	12	441
3	262	13	259
4	350	14	437
5	215	15	236
6	262	16	442
7	556	17	259
8	415	18	638
9	262	19	260
10	439	20	726

3.1 The Similarity Measure

To annotate bacteria we define a similarity measure between the two bacteria first as the length of the longest common sub sequence of the two bacteria 16S gene sequences as follows:

st1=CGACG**CTGGCGG**CGT**GCCTAACACATG**CAAG
st2=**GCCTAACACATG**ATTACTAGGT**CTGGCGG**GTC

The longest common subsequence of these two strings is

GCCTAACACATG

Although there are other common subsequence of these two strings, this is the longest, and the length 12 of this common string is a measure of similarity of st1, and st2 (Can, and Ozsoy 2018).

Then we define the affinity of a bacteria to a taxonomic class.

Similarity of the a bacteria to a taxonomic class

Let Q is the query bacteria, and a taxonomic class consists of bacteria

$$TC = \{B_1, B_2, B_3, \dots, B_n\}. \quad (1)$$

Let the sequence of similarities of Q to the bacteria in TC is

$$A = \{S_1, S_2, S_3, \dots, S_n\}. \quad (2)$$

The maximum of the sequence A, is the affinity F of the query Q, to the taxonomic class TC.

$$F(Q, TC) = Max(A). \quad (3)$$

3.2 Annotation of Bacteria

To annotate unknown bacteria Q, to taxonomic classes, the affinity of this unknown bacterium to all taxonomic classes, at a level of the taxonomy, are computed. To decrease the computational workload, 50 bacteria are randomly sampled from groups with bacteria more then 50. Let at a taxonomic level, the sublevels are

$$C = \{C_1, C_2, C_3, \dots, C_m\}. \quad (4)$$

and the affinity of Q to those classes be

$$F = \{F_1, F_2, F_3, \dots, F_m\}. \quad (5)$$

If the maximum of the sequence F is F_k, it is concluded that the unknown bacteria Q, belongs to the taxonomic class C_k.

To test this technique in a previous research (Can, and Ozsoy 2018), at all levels of taxonomies SILVA, RDP, and Greengenes, from each sublevel one random bacteria is chosen, then using the longest common subsequence similarity measure, these bacteria are re annotated. The results for the Greengenes are repeated here.

3.3 Annotation Accuracies for Greengenes

The taxonomy Greengenes has phylum, class, order, family, genus, and species levels. Accuracies obtained in re annotations are as in Table 4.

Table 4. Accuracies obtained in re annotations in Greengenes

query	# Subgroups	Accuracy %
Phylum	85	91.63
Class	223	91.03
Order	366	92.90
Family	466	91.63
Genus	*1949	87.36
Species	**2389	70.51

*Sublevels with only 1 and 2 bacteria are disregarded.
** Sublevels with less than 50 bacteria are disregarded.

3.4 The effect of sampling

The effect of sampling is studied at phylum levels. It is seen that Greengenes data is the one who effected by sampling most.

Table 5. The effect of sampling at phylum levels on percent accuracies

Sample Size	SILVA	RDP	Greengenes
50	90.12	82.00	84.88
100	96.30	90.00	88.37
200	93.83	90.00	88.37
500	95.06	91.63	91.63
1000	96.30	96.00	84.88
5000	98.76	98.00	82.56
Full	94.87	94.00	80.23

4. ANNOTATION OF TWENTY SHORT READS

Amplified pieces of DNA's sent for sequencing. The length of sequences are as in the Table 4. Appendix A shows full sequences (Cosic 2019).

To annotate these twenty short reads of length 300 base pairs in average are first tested against phylum classes of the Greengenes taxonomy. Each of 20 sequences annotated to one of the 86 phylum classes according their similarities. The phylum annotation of a sequence is to the phylum class with maximum similarity to the sequence. The same is done for all taxonomic classes. The results are shown in Table 6.

Table 6. Annotation of 20 Bacteria Sequences Against Greengenes Library

	Phylum	Class	Order	Family	Genus	Species	Main
1	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	cohnii	1
2	TPD-58	Bacilli	Lactobacillales	Enterococcaceae	Lactobacillus	plantarum	9
3	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	selenatarsenatis	1
4	FBP	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	plantarum	9
5	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Bacillus	endophyticus	1
6	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	plantarum	9
7	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	foraminis	1
8	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	firmus	1
9	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	flexus	1
10	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	firmus	1
11	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	flexus	1
12	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	firmus	1
13	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	selenatarsenatis	1
14	Actinobacteria	Bacilli	Bacillales	Bacillaceae	Bacillus	firmus	1
15	Firmicutes	AHT28	Bacillales	Bacillaceae	Bacillus	flexus	1
16	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	firmus	1
17	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	selenatarsenatis	1
18	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	plantarum	9
19	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	plantarum	9
20	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	coagulans	1

5. CONCLUSION

From Table 6, we may conclude the following: Among the purchased bacteria in the list, there is one with genus identity Basillus. Therefore, during the bacterial growth, DNA isolation and sequencing stages, fifteen sequenced 16S rRNA gene parts denoted by "1" in the last column belong to the same bacteria Bacillus subtilis first in the bacteria list.

Five sequenced 16S rRNA gene parts denoted by "9" in the last column belong to the same bacteria Bacillus plantarum nineth in the bacteria list.

Longest common subsequence is a novel similarity measure. It is seen that the re annotation accuracies are comparable with the accuracies of more sophisticate tools.

REFERENCES

Ash, C., J. A. E. Farrow, S. Wallbanks, and M. D. Collins. (1991) Phylogenetic heterogeneity of the genus bacillus revealed by comparative analysis of small subunit ribosomal RNA sequences. Lett. Appl. Microbiol. 13:202–206.

Audic, S., and J. M. Claverie (1997) The significance of digital gene expression profiles. Genome Res. 7:986–995.

Baker GC, Smith JJ, Cowan DA. (2003) Review and re-analysis of domain- specific 16S primers. J Microbiol Methods. 55(3): 541–55.

Balvociute, M., and Huson, DH. (2017) SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?, BMC Genomics 2017, 18 (Suppl 2):114

Beiko RG. (2016) Microbial malaise: How can we classify the microbiome? Trends Microbiol. 23(11):671–9.

Benson, D., Lipman, D.J., and Ostell, J. (1993) Genbank Nucleic Acids Res., 21, 2963-2965.

Bergey, D.H. (1925) Bergey's Manual of Determinative Bacteriology, Baltimore : Williams & Wilkins Co. (with many subsequent editions)

Bisset, K. A. (1962) Bacteria, 2nd ed., Livingston, London

Bonnie L.Maidak, Niels Larsen, Michael J.McCaughey, Ross Overbeekl, Gary J.Olsen, Karl Fogel, James Blandy2 and Carl R.Woese (1994) The Ribosomal Database project, Nucleic Acids Research, Vol. 22, No. 17 3485-3487

Brown, M. P. S. (1999) RNA modeling using stochastic context-free grammars. Ph.D. thesis. University of California, Santa Cruz.

Brown,M.P.S. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammars. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB 2000), San Diego, CA. pp. 57–66.

Bruno, M. D., Korfhagen, T. R., Liu, C., Morrisey, E. E. and Whitsett, J. A. (2000). GATA-6 activates transcription of surfactant protein A. J. Biol. Chem. 275, 1043-1049.

Buchanan, R E J (1916) Bacteriol. Nov;1(6):591-6. ISSN 0021-9193 PMC 378679

Can, M., and Gursoy, O. (2018) Artificial Neural Networks in Bacteria Taxonomic Classification, Southeast Europe Journal of Soft Computing Vol.7 No.2 (1-7)

Cannone, J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:2.

Cohn, F. (1875) Untersuchungen uber Bakterien, II. Beitrage zur Biologie der Pfanzen 1: 141-207

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. (2014) Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42 (Database issue): 633–42.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., etal., (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. 35, D169–D172.

Ćosić, A. (2019) Investigation of 16s Rrna Gene and Gene Segments for Determination of Probiotics, Master Thesis, International University of Sarajevo September, 2019

Ćosić, A., and Jahjefendic, A. H. Investigation Of 16S rRNA Gene And Gene Segments For The Determination Of Probiotics, Southeast Eur. J. Soft Comput., vol. 8, no. 1, Apr. 2019.

DeSantis, T. Z., I. Dubosarskiy, S. R. Murray, and G. L. Andersen (2003) Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. Bioinformatics 19:1461–1468.

Ehrenberg C.G. (1832) Beiträge zur Kenntnis der Organization der Infusorien und ihrer geographischen Verbreitung besonders in Sibirien. Abhandlungen der Koniglichen Akademie der Wissenschaften zu Berlin, 1832, 1830, 1-88.

Ehrenberg C.G. (1833-1835) Dritter Beitrag zur Erkenntniss grosser Organisation in der Richtung des kleinsten Raumes. Abhandlungen der Preussischen Akademie der Wissenschaften (Berlin) aus den Jahre 1833-1835, pp. 143-336.

Ehrenberg C. G. (1828 [plates], 1831 [text]). Symbolae physicae animalia evertebrata. In: Symbolae physicae, seu Icones adhue ineditae corporum naturalium novorum aut minus cognitorum, quae ex itineribus per Libyam, Aegyptum, Nubiam, Dengalam, Syriam, Arabiam et Habessiniam. Pars Zoologica. Hemprich F. G. & Ehrenberg C. G. (eds.). Officina Academica: Berlin. pp. 2 and 8, plate 2, fig. 6, [1].

Smith, E. F. (1905) Nomenclature and Classification in Bacteria in Relation to Plant Diseases vol. 1

Federhen S. (2012) The NCBI taxonomy database. Nucleic Acids Res. 40(D1):136–43.

Haeckel, Ernst (1867). Generelle Morphologie der Organismen. Reimer, Berlin. ISBN 1-144-00186-2.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA. (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci USA. 112 (41): 12764–9.

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. (2016) MEGAN Community Edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. 12(6):1004957. doi:10.1371/journal. pcbi.1004957.

Komárek J, Hauer T. CyanoDB.cz – On-line database of cyanobacterial genera. (2016) Word-wide electronic publication. http://www.cyanodb.cz, Univ. of South Bohemia & Inst. of Botany AS CR.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 1985; 82(20): 6955–9.

Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olsen, K.Fogel, J. Blandy, and C. R. Woese. (1994) The Ribosomal Database Project. Nucleic Acids Res. 22:3485–3487.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. (2012) An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6(3):610–8.

Migula W. (1897) System der Bakterien. Gustav Fischer, Jena

Murray, R.G.E., Holt, J.G. (2005). The history of Bergey's Manual. In: Garrity, G.M., Boone, D.R. & Castenholz, R.W. (eds., 2001). Bergey's Manual of Systematic Bacteriology, 2nd ed., vol. 1, Springer-Verlag, New York, p. 1-14. link. [See p. 2.]

Orla-Jensen S., (1909) Die Hauptlinien des naturalischen Bakteriensystems nebst einer Ubersicht der Garungsphenomene. Zentr. Bakt. Parasitenk., II, 22: 305-346

Pacini, F. (1854) Osservazione microscopiche e deduzioni patologiche sul cholera asiatico. Gazette Medicale de Italiana Toscano Firenze, 6, 405-412.

Price MN, Dehal PS, Arkin AP. (2009) Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26 (7): 1641– 50. doi: 10.1093/ molbev/msp077.

Stanier, R. Y., and van Niel, C. B. (1941) The main outlines of bacterial classification. J. Bact., 42, 437-466,

Tringe SG, Hugenholtz P. (2008) A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008;11(5):442–6.

Vuillemin, P. (1913) Genera Schizomycetum. Annales Mycologici. 11,512-527.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7. doi:10.1128/AEM.00062-07.

Wang Y, Qian P-Y. (2009)Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One. 4(10):e7401.

Winslow, C.E. A., J. Broadhurst, R. E. Buchanan, C. Krumwiede, Jr., L. A. Rogers, and G. H. Smith. (1917) The families and genera of the bacteria. Preliminary report of the Committee of the Society of American Bacteriologists on Characterization and Classification of Bacterial Types. J. Bacteriol. 2505-566.

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

Yang, B,, Wang, Y., and Qian, P.-Y. (2016) Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis , BMC Bioinformatics 17:135-142

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 42(Database issue):643–8.

APPENDIX A

16 S rRNA SEQUENCES

Sequence_1

GATCGCATGAGAGTCTGACGGAGCACGCCGCGTGAGTGA
TGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAG
AACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTAC
CTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCG
CGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTG
GGCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATGT
GAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACT
GGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCACGT
GTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGT
GGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAGGA
GCGAAAGCGTGGGGAGCGAACAGGATTAGATATCCTGTG
TAGAATTCGG

Sequence_2

GGGACGAGTCGGATATTGGGCGTAAGCGATCGCAGGCGG
TTTTTTAAGTCTGATGTGAAAGCCCCTCGGCTCAACCGGG
GAGGTGTCATCTGGAAACTGGGGAACTTGAGTGCAGAAG
AGGACAAGTGGAACTCCATGTGTAGCGGTGAAATGCGTA
GATATATGGAAGAACACCAGTGGCGAAGGCGGCTCTCTG
GTCTGTAACTGACGCTGAGGCTCGAAAGTATGGGTAGCA
AAACAGGATTAGATACCCTGGTAGTCCAG

Sequence_3

AGCACGTGTCGGATTATTGGGCGTAAGGGCTCGCAGGCG
GTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGG
GGAGGGTCATTGGAAACTGGGGAACTTGAGTGCAGAAG
AGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTAG
AGATGTGGAGGAACACCAGTGGCGAAGGCGACTCTCTGG
TCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGA
ACAGGATTAGATACCCTGGTAGTACAGGT

Sequence_4
GCTCATGGAGAGTCTGATGGAGCACGCCGCGTGAGTGAA
GAAGGGTTTCGGCTCGTAAAACTCTGTTGTTAAAGAAGA
ACATATCTGAGAGTAACTGTTCAGGTATTGACGGTATTTA
ACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGG
TAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGC
GTAAAGCGAGCGCAGGCGGTTTTTTAAGTCTGATGTGAA
AGCCTTCGGCTCAACCGAAGAAGTGCATCGGAAACTGGG
AAACTTGAGTGCAGAAGAGGACAGTGGAACTCCCTGTGT
AGCGGTGAAATGCGTAAATATATGGAAAAAAACCGAA

Sequence_5

GATCGAATCTCGAGTCTGACGGAGCACGCCGCGTGAGTG
ATGAAGGTTTTCGGCATCGTAAAAGCTCTGTTGTTACGCG
GTAAGAACATATCCCATTCGAATAGGGCGGTATCTTGAC
GGTACCTCACCAAAAGCCCGGTTAACTACTTGCCCAAGC
CCGGGAATTACAAGGGGGCAAGCGTGGTCCGGAATTTTG
GGGGTAAAGGGGTCCAGGG

Sequence_6

GGGACGTGTCGGATTTATTGGGCGTAAGCGAGCGCAGGC
GGTTTTTTAAGTCTGATGTGAAAGCCTTCGGCTCAACCGA
AGAAGTGCATCGGAAACTGGGAAACTTGAGTGCAGAAG
AGGACAGTGGAACTCCATGTGTAGCGGTGAAATGCGTAG
ATATATGGAAGAACACCAGTGGCGAAGGCGGCTGTCTGG
TCTGTAACTGACGCTGAGGCTCGAAAGTATGGGTAGCAA
ACAGGATTAGATACCCTGGTAGTCAACG

Sequence_7

GACGCAGAGAGTCTGACGGAGCACGCCGCGTGAGTGAT
GAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAGA
ACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTACC
TAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGC
GGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTGG
GCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATGTG
AAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACTG
GGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCACGT
GTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGT
GGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAGGA
GCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTGGT
AGTCACAGGGAATCTTCCGCAATGGACGAAAGTCTGACG
GAGCAACGCGCTTGAGTGATGAAAGGTTTTCGGATCGTA
AAGCTCTATTGTTAGGGATTAATCATGTTTCTTTTGGATA
AAGGGGGGTA

Sequence_8

GATGGCATGAGAAGTCTGACGGAGCACGCCGCGTGAGTG
ATGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAA
GAACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTA
CCTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCC
GCGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATT
GGGCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATG
TGAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAAC
TGGGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCAC
GTGTAGCGGTGAAATGCGTAAAGATGTGGAGGAACACCA
GTGGCGAAGGCGACTCTCTGGGCTGTAACTGACGCTGAT
GACCGTCAGTGTGGGGAAATACCTT

Sequence_9

AGGACGGATTCGGATTATTGGGCGTAAGGGCTCGCAGGC
GGTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCG
GGGAGGGTCATTGGAAACTGGGGAACTTGAGTGCAGAA
GAGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTA
GAGATGTGGAGGAACACCAGTGGCGAAGGCGACTCTCTG
GTCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCG
AACAGGATTAAATACCCTGGTAGTCCCAG

Sequence_10

GATGGCATGAGAGTCTGACGGAGCACGCCGCGTGAGTGA
TGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAG
AACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTAC
CTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCG
CGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTG
GGCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATGT
GAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACT
GGGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCAC

GTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCA
GTGGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAG
GAGCGAAAGCGTGGGGAGAGAACACGAGAATAGACACC
CTGGGAAACCAG

Sequence_11

AGTACGTGTCGGATTATTGGGCGTAAGGGCTCGCAGGCG
GTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGG
GGAGGGTCATTGGAAACTGGGGAACTTGAGTGCAGAAG
AGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTAG
AGATGTGGAGGAACACCAGTGGCGAAGGCGACTCTCTGG
TCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGA
ACAGGATTAAATACCCTGGTAGTCCAAG

Sequence_12

GACGCATGAGAGTCTGACGGAGCACGCCGCGTGAGTGAT
GAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAGA
ACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTACC
TAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGC
GGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTGG
GCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATGTG
AAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACTG
GGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCACGT
GTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGT
GGCGAGGCGACTCTCTGGTCTGTAACTGACGCTGAGGAG
CGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTGGGT
AGAGCTTCGGG

Sequence_13

AGCACGTGTCGGATTATTGGGCGTAAGGGCTCGCAGGCG
GTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGG
GGAGGGTCATTGGAAACTGGGGAACTTGAGTGCAGAAG
AGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTAG
AGATGTGGAGGAACACCAGTGGCGAAGGCGACTCTCTGG
TCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGA
ACAGGATTAGATACCCTGGTAGTCCA

Sequence_14
GATCGCAGTCGAGTCTGACGGAGCACGCCGCGTGAGTGA
TGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAG
AACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTAC
CTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCG
CGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTG
GGCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATGT
GAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACT
GGGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCAC
GTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCA
GTGGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAG
GAGCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCT
GGGTAGTCTA

Sequence_15

ATACGTTCTCGGCACGATCGACCACTAAAGTTTGTAACA
GCCGATTCTCTAGTGATGTTACCTTTTACGACCCCTCCGC
CGGGGGTGGGACATATGATTGGGGTGAAGTCGTTGCAAG
GTAACCGAGTGGAATTCCACGTGTAGCGGTGAAATGCGT
AGAGATGTGGAGGAACACCAGTGGCGAAGGCGACTCTCT
GGTCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGC
GAACAGGATTAGATACCCTGGTAGTCCA

Sequence_16

GATGCATGAGAAGTCTGACGGAGCAACGCCGCGTGAGTG
ATGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAA
GAACAAGTACCGTTCGAATAGGGCGGTACCTTGACGGTA
CCTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCC
GCGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATT
GGGCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATG
TGAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAAC
TGGGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCAC
GTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCA
GTGGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAG
GAGCGAAAGCGTGGGGAGCGAACAGGAGTAGATACCCT
GGTAGGCGTCGGGA

Sequence_17

AGCACGAGTCGGATTATTGGGCGTAAGGGCTCGCAGGCG
GTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGG
GGAGGGTCATTGGAAACTGGGGAACTTGAGTGCAGAAG
AGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTAG
AGATGTGGAGGAACACCAGTGGCGAAGGCGACTCTCTGG
TCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGA
ACAGGATTAGATACCCTGGTAGTCCA

Sequence_18

GCTGAATGAGAAGTCTGATGGAGCACGCCGCGTGAGTGA
TGAAGGGTTTCGGCTCGTAAAACTCTGTTGTTAAAGAAG
AACATATCTGAGAGTAACTGTTCAGGTATTGACGGTATTT
AACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCG
GTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGG
CGTAAAGCGAGCGCAGGCGGTTTTTTAAGTCTGATGTGA
AAGCCTTCGGCTCAACCGAAGAAGTGCATCGGAAACTGG
GAAACTTGAGTGCAGAAAAGGACAGTGGAACTCCATGTG
TAGCGGTGAAATGCGTAAATATATGGAAGAACACCAGTG
GCGAAGGCGGCTGTCTGGTCTGTAACTGACCCTGAAGCT
CGAAAGTATGGGTAACTAACAGGATTAGATACCCTGGTA
GTCAAAGGAATCTTCCACAATGGACGAAAGTCTGATGGA
GCAACGCCGCGTGAGTGAAGAAGGGTTTCCGCTCTTAAA
ACTCTGTTTGTTTAAGAAAAACATATCTAGATTTAACTGT
TCCGGTATTGTTCGGCATTTCACCAGAAAGCCCGGGTTAC
TACTTGCCCCCACCCCCGTAATATCTATTGTGGACCCCTT
TTTCGAAATT

Sequence_19

GGCACGTATCGGATTCTTGGGCGTAAGCGAGCGCAGGCG
GTTTTTTAAGTCTGATGTGAAAGCCTTCGGCTCAACCGAA
GAAGTGCATCGGAAACTGGGAAACTTGAGTGCAGAAGA
GGACAGTGGAACTCCATGTGTAGCGGTGAAATGCGTAGA
TATATGGAAGAACACCAGTGGCGAAGGCGGCTGTCTGGT
CTGTAACTGACGCTGAGGCTCGAAAGTATGGGTAGCAAA
CAGGATTAGATACCCCTGGTAGTCCA

Sequence_20

GACGGCATGAGAAGTCTGACGGAGCAACGCCGCGTGAG
TGAAGAAGGCCTTCGGGTCGTAAAACTCTGTTGCCGGGG
AAGAACAAGTGCCGTTCGAACAGGGCGGCGCCTTGACGG
TACCCGGCCAGAAAGCCACGGCTAACTACGTGCCAGCAG
CCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTA

TTGGGCGTAAAGCGCGCGCAGGCGGCTTCTTAAGTCTGA
TGTGAAATCTTGCGGCTCAACCGCAAGCGGTCATTGGAA
ACTGGGAGGCTTGAGTGCAGAAGAGGAGAGTGGAATTCC
ACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACAC
CAGTGGCGAAGGCGGCTCTCTGGTCTGTAACTGACGCTG
AGGCGCGAAAGCGTGGGGAGCAAACAGGATTAGATACC
CTGGTAGTCAGGGAATCTTCCGCAATGGACGAAAGTCTG
ACGGAGCAACGCCGCGTGAGTGAAGAAGGCCTTCGGGTC
GTAAAACTCTGTTGCCGGGGAAGAACAAGTGCCGTTCGA
ACAGGGCGGCGCCTTGACGGTACCGGCCAAAAAGCCCCG
GTTAATTACGTGCCCGCGCCCGCGGTAATACGTAAGGTT
GGCAAGCGCTTTTCCGGAATTATTGGGGCGTAAAGGGCG
CGCAGGCGGGTTTTCTTAAATCTGAATGTGAAAATCTTGG
GGCTAACCCCAGCGGTCATTGGGAA