VOL.7 NO.2 September 2018 - ISSN 2233 – 1859: DOI: 10.21533/scjournal.v7i2.144

UOIuBIH
ORSinBIH
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

Longest Common Subsequences in Bacteria Taxonomic Classification

M. Can
O. Gürsoy

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
ogursoy@ius.edu.ba

Article Info

Article history:
Article received on 10 June 2018
Received in revised form 1 August 2018

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

ABSTRACT: In 1980s, Carl Woese made a ground breaking contribution to microbiology using rRNA-genes for phylogenetic classifications. He used it not only to explore microbial diversity but also as a method for bacterial annotation. Today, rRNA-based analysis remains a central method in microbiology. Many researchers followed this track, using several new generations of Artificial Neural Networks obtained high accuracies using available datasets of their time. By the time, the number of bacteria increased enormously. In this article we used Longest Common Subsequence similarity measure to classify bacterial 16S rRNA gene sequences of 1.820.414 bacteria in SILVA, 3.196.038 bacteria in RDP, and 198.509 bacteria in Greengenes. The last two taxonomy have six taxonomical levels, phylum, class, order, family, genus, and species, while SILVA has two more levels subclass and suborder, but lacks species level. The majority of classifications (98%) were of high accuracy (98%).

1. INTRODUCTION

Bacteria are often identified as the causes of human and animal diseases. However, 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. They help to break down dead organic matter; make up the base of the food web in many environments. Bacteria are of such immense importance because of their extreme flexibility, capacity for rapid growth and reproduction, and contribution to the processes in the body of humans.

Bacteria also contribute immensely to global energy conversion and the recycling of matter. Thus, profiling the microbial community is one of the most important tasks for microbiologists to explore various ecosystems. However, our understanding of the kingdom Bacteria remains limited laboratory conditions (Ash et. al., 1991). In the past few decades, DGGE, Denaturing gradient gel electrophoresis,

(Audic, and. Claverie, 1997), T-RFLP, Terminal restriction fragment length polymorphism (Benson, et. al., 2000), FISH, fluorescent situ hybridization (Brown, 1999), and Genechips (Bruno, et. al., 2000) were used 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, Christensen, 1992) and Illumina (Cole, e., al., 2006) have facilitated a remarkable expansion of our knowledge regarding uncultured bacteria (Yang et., a., 2016).

A Brief History of Bacterial Classifications[1]

Ernst Haeckel, in the year 1866, in the Tree of Life in Generelle Morphologie der Organismen (Haeckel, 1867) first classified bacteria as plants, constituting the class Schizomycetes. 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.

Indeed a genus of comma shaped bacteria, Vibrio, first described in 1854 (Pacini, 1854). 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).

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

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, K. A. 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, Chlamydobacteriales, 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 tec technique to microbiology with his rRNA-based phylogenetic 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 genesequences consist of nine hypervariable regions that areseparated 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., 2012), 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

[1] https://en.wikipedia.org/wiki/Bacterial_taxonomy

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 inconsistent 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 taxonomies SILVA, RDP, Greengenes are visited.

2.1 Silva

From Latin silva, forest[2], the bacterial and archaeal classification in SILVA is based on Bergey's Taxonomic Outlines (Boone, et. al., 2001; Brenner, et. al., 2005; Vos, et. al., 2009; Krieg, et. al., 2010). It is a comprehensive resource for up-to-date quality-controlled databases of aligned ribosomal RNA (rRNA) gene sequences from the Bacteria, Archaea and Eukaryota domains and supplementary online services. SILVA provides a manually curated taxonomy for all three domains of life, based on representative phylogenetic trees for the small and large-subunit rRNA genes. The improvements of the SILVA taxonomy has undergone in the last five years.

A comparison of the SILVA taxonomy with Greengenes and RDP taxonomies reveales a reasonable overlap between the taxa names, and points to significant differences in both names and numbers of taxa between the three resources (Quast, et. al., 2013).

The SILVA database (Yilmaz et. al. 2014) bases primarily on phylogenies for small subunit rRNAs, 16S for prokaryotes and 18S for Eukarya. Taxonomic rank information for Archaea and Bacteria is obtained from Bergey's Taxonomic Outlines (Boone, et. al. 2013; Brenner, et. al. 2005; Vos, et. al. 2009; Krieg, et. al. 2010) and from the List of Prokaryotic Names with Standing in Nomenclature (LPSN) (Parte, 2014), whereas eukaryotic taxonomy is based on the consensus views of the International Society of Protistologists (Adl, et. al., 2005; Adl, et. al., 2012). Taxonomic rank assignments in the SILVA database are manually curated (Yilmaz et. al. 2014).

SILVA predominantly uses phylogenetic classification based on an SSU guide tree. Classification and clade names are informed by widely accepted sources, and discrepancies are resolved with the overall aim of making classification consistent with phylogeny. With release 100 in 2009, the SILVA full-length (>1200 bases for Bacteria/Eukaryota and >900 bases for Archaea) SSUgene guide tree went through a major manual curation effort to represent bacterial and archaeal taxa as groups in the tree. The core of this guide tree is based on the full length sequence tree of the ARB. 2004 release (curated and distributed by Wolfgang Ludwig), and is built by adding new sequences using the ARB parsimony tool in combination with filters to remove highly variable positions (Pruesse, et., al., 2006).

In the following releases, the curated classifications were extended to cover bacterial and archaeal full-length large subunit (LSU, 23S rRNA) and eukaryotic full-length SSU (18S rRNA) gene sequences. With the SILVA release 115 in August 2013, all quality-checked SSU and LSU rRNA gene sequences from all three domains of life were automatically classified based on the established SSU and LSU reference taxonomies.

Extensive effort is spent in every release to represent prominent clades known only from environmental sequences. The majority of these clades and groups are annotated inthe guide tree based on literature surveys, and occasionally based on personal communications; therefore, not all of these clades are available in publications. Some examples are OCS116 clade (Morris, et., al., 2005), SAGMC and SAGME groups (Takai, et., al., 2001), and termite clusters (Kohler, et., al., 2008). Supplementary Table S1provides a full list of all such clades and groups that are

[2] http://www.arb-silva.de

part of the current SILVA taxonomy. We chose to name phylogenetically coherent groups above the family rank, consisting of only sequences from uncultured organisms, after the clone name of the earliest submitted sequence.

Finally with the release 132 appeared in July 2017, the SILVA alignment is 50,000 columns long so that it can be compatible with 18S rRNA sequences as well as archaeal 16S rRNA sequences. In a shift from previous version of the SILVA references, it provides now the SEED database, the full-length sequences available from the NR SILVA database, and a SILVA aligned version of the gold database that is used for reference-based chimera detection.

Table 1. Levels and number of sublevels in SILVA

Levels	# Sublevels
Phylum	81
Class	424
Order	844
Family	2118
Genus	5318
Species	183284

2.2. Ribosomal Database Project (RDP)

The RDP database (Cole, et., al., 2014) is based on 16S rRNA sequences from Bacteria, Archaea and Fungi (Eukarya). It contains 16S rRNA sequences available from the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane, et., al., 2016) databases. Names of the organisms associated with the sequences are obtained as the most recently published synonym from Bacterial Nomenclature Up-to-Date. Information on taxonomic classification for Bacteria and Archaea is based on the taxonomic roadmaps by Bergey’s Trust and LPSN (Parte, 2014). Taxonomic information for fungi is obtained from a hand-made classification dedicated to fungal taxonomy (Cole, et., al., 2014).

2.2.1 History of Rdp

The RDP arose out of research conducted by two University of Illinois at Urbana-Champaign (UIUC) faculty members, Carl R. Woese and Gary J. Olsen. Woese recognized that, due to rRNA's conserved sequence, ribosomal RNA could be used to elicit phylogenetic relationships between organisms. They foresaw that making a collection of rRNA gene sequences available would be useful to the research community and stimulate research in this area. Initial funding for the RDP was awarded in 1989 by the Biological Instrumentation and Resources Program of the National Science Foundation. Argonne National Laboratory first hosted the RDP ftp and public sites and on January 5, 1992, 471 16S rRNA sequences, many of which were generated in Woese's laboratory, were made available to the public in the first release of the RDP. The public sites were moved to UIUC for Release 3.0 in August 1993. NSF predominantly supported the RDP to 1997. As data were originally stored as flat files, additional funding to migrate to a commercial database management system was awarded jointly to Michigan State University (MSU) and UIUC in 1995. During the last 18 months of core NSF funding, discussions with MSU faculty at the Center for Microbial Ecology led to the relocation of the RDP to MSU.

The first data release and official announcement of the RDP-II WWW site occurred on July 31, 1998. For Relases 7.1 and 8.0, RDP-II staff members at MSU included Bonnie Maidak, responsible for curation and user support, and Jim Cole, who oversaw, and continues to oversee, the website, database and development.

Release 9.0 marked a substantial change to the RDP. Due to an explosion of sequence data being made available by sequence repositories,

Since the first published article describing the RDP in 1991, eight additional articles describing the RDP have been published in the annual databases issue of Nucleic Acids Research. The ribosomal RNA sequences in the RDP alignments are drawn from major sequence repositories, GenBank (Benson, et., al., 1993) and EMBL (Rice, et., al., 1993), and direct submissions to the RDP. They are organized and presented in aligned and phylogenetically ordered form. Each sequence is annotated with its organismal source, for cultured organisms: the genus, species, culture collection numbers, etc., cellular compartment, origin of sequence data, and other relevant information

As of September 2006 (Release 9.42), the RDP maintained 262 030 aligned and annotated public rRNA sequences. Of these, 84 442 were from cultivated bacterial strains, while177 588 were derived from environmental samples. A totalof 101 877 sequences were near-full-length (>1200 bases)and 5543 sequences were from bacterial type strains; these sequences are of special importance as they help to link taxonomy and phylogeny.

As a major quality improvement, all sequences are now tested for sequence anomalies, including chimeric sequence anomalies, using Pintail from the Cardiff Bioinformatics Toolkit (Ashelford, et al., 2005). Using Pintail on a subset of the RDP public sequences, those authors reported that at least 5% of rRNA records contain some type of anomaly. Cole et al. employed a similar strategy to detect anomalous sequences (Cole, et al., 2007). Each sequence is compared with at least two sequences from different publications and those reported as anomalous in both comparisons are marked as suspect. (For a small percentage of sequences, results of the first two tests are not consistent and additional comparisons are necessary to establish a pattern.) Of the 262 030 sequences in release 9.42, 21 771 are deemed anomalous by this criterion. When the sequences are subdivided based on source (isolate versus environmental) and short versus long, we find the anomalies are greatest in the environmental and short sequences.

As of September 2008 (release 10.3), the Ribosomal Database Project (RDP) maintained 33 082 archaeal and 643 916 bacterial small subunit rRNA sequences. Of these, 142 511 came from cultured organisms while 534 487 were sequences obtained from environmental samples.

RDP Release 11.1 consists of 2,809,406 aligned and annotated 16S rRNA sequences and 62,860 Fungal 28S rRNA sequences. RDP Release 11.2 consists of 2,929,433 aligned and annotated 16S rRNA sequences and 95,365 Fungal 28S rRNA sequences.

RDP Release 11.3 consists of 3,019,928 aligned and annotated 16S rRNA sequences and 102,901 Fungal 28S rRNA sequences. Release 11.5 consists of 3,356,809 aligned and annotated 16S rRNA sequences and 125,525 Fungal 28S rRNA sequences. Release 11.4 consists of 3,224,600 aligned and annotated 16S rRNA sequences and 108,901 Fungal 28S rRNA sequences.

Table 2. Levels and number of sublevels in RDP

Levels	# Sublevels
Phylum	51
Classs	125
Subclass	225
Order	390
Suborder	2040
Family	109
Genus	353
Species	No species data

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 rankmapping 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 3. Levels and number of sublevels in Greengenes

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

2.4 NCBI

The NCBI taxonomy (Federhen, 2012) contains the names of all organisms associated with submissions to the NCBI sequence databases. It is manually curated based on current systematic literature, and uses over 150 sources, for example, the Catalog of Life , the Encyclopedia of Life, Name-Bank and WikiSpecies as well as some specific databases dedicated to particular groups of organisms. It contains some duplicate names that represent different organisms. Each node has a scientific name and may have some synonyms assigned to it (Federhen, 2012). NCBI taxonomic classification files are updated on a daily basis; in this paper we use the version as of 05/10/2016 (Balvocilute, and Huson, 2017).

Table 4. Levels and number of sublevels in NCBI

Levels	# Sublevels
Phylum	51
Classs	125
Subclass	225
Order	390
Suborder	2040
Family	109
Genus	353
Species	No species data

2.5 Open Tree Of Life Taxonomy (OTT)

The Open Tree of life Taxonomy (Hinchliff, et. al., 2015) aims at providinga comprehensive tree spanning as many taxa as possible.OTT is an automated synthesis of published phylogenetic trees and reference taxonomies. Phylogenetic trees have been ranked, aligned and merged together; taxonomies have been used to fill in the sparse regions and gaps left by phylogenies. Phylogenetic trees for the synthesis are obtained from Tree BASE (Sanderson, et. al., 1994), Dryad (Dryad, 2016) and in some cases directly from contributing authors. Taxonomies are sourced from Index Fungorum, SILVA, NCBI, Global Biodiversity Information Facility, Interim Register of Marine and Nonmarine Genera and some clade specific resources (Hinchliff, et. al., 2015).

Table 5. Levels and number of sublevels in OTT

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

3. MATERIALS AND METHODS; LONGEST COMMON SUBSEQUENCE

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:

Similarity of the two bacteria

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.

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

$$TC = \{B_1, B_2, B_3, \dots, B_n\}. \tag{1}$$

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

$$A = \{S_1, S_2, S_3, \dots, S_n\}. \tag{2}$$

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

$$F(Q, TC) = Max(A). \tag{3}$$

3.1 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

$$C = \{C_1, C_2, C_3, \dots, C_m\}. \tag{4}$$

and the affinity of Q to those classes be

$$F = \{F_1, F_2, F_3, \dots, F_m\}. \tag{5}$$

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.

4. RESULTS

At 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 following accuracies are achieved.

4.1 Annotation Accuracies for SILVA

The taxonomy SILVA has phylum, class, order, family, genus, and species levels. Accuracies obtained in re annotations are as in Table 6.

Table 6. Accuracies obtained in re annotations in SILVA

Level	#Sublevels	% ccuracy
Phylum	81	98.77
Class	424	79.46
Order	844	80.04
Family	2118	83.76
Genus	*5318	89.69
Species	**183284	82.00

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

4.2 Annotation Accuracies for RDP

The taxonomy RDP has phylum, class, subclass, order, suborder, family, and genus levels. Accuracies obtained in re annotations are as in Table 7.

Table 7. Accuracies obtained in re annotations in RDP

Level	#Sublevels	% ccuracy
Phylum	51	100.00
Class	125	91.24
Subclass	225	92.04
Order	390	92.78
Suborder	*2040	86.73
Family	109	93.58
Genus	353	83.07

*Sublevels with only 1 and 2 bacteria are disregarded.

4.3 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 8.

Table 8. 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.

4.4 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 9. 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

5. CONCLUSION

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

Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MF. (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 52(5):399–451.

Adl SM, Simpson AG, Lane CE, Lukeš J, Bass D, Bowser SS, Brown M, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, leGall L, Lynn DH, McManus H, Mitchell EAD, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Schoch C, Smirnov A, Spiegel FW. (2012) The revised classification of eukaryotes. J Eukaryot Microbiol. 2012;59(5): 429–93. doi:10.1111/ j.1550-7408.2012.00644.x.

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.

Ashelford,K.E., Chuzhanova,N., Fry,J.C., Jones,A.J. and Weightman,A.J. (2005) At least one in twenty 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol., 71, 7724–7736.

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) 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

Boone,D.R., Castenholz,R.W., Garrity,G.M. and Stanley,J.T. (2001) The Archaea and the Deeply Branching and Phototrophic Bacteria. Springer, New York.

Brenner DJ, Krieg NR, Garrity GM, Staley JT, (eds). (2005) Bergey’s Manual of Systematic Bacteriology, Volume 2: The Proteobacteria, 2nd edn. US: Springer-Verlag.

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.

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

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.

Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. (2010) Qiime allows analysis of high-throughput community sequencing data. Nat Methods. 7(5):335–6.

Cochrane G, Karsch-Mizrachi I, Takagi T (2016) International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. Nucleic Acids Res. 44 (Database issue): 48–50. doi:10.1093/nar/gkv1323.

Cohn, F. (1872a) Organismen in der Pockenlymphe. Virchow's Archiv, 55, 229-238. p. 237

Cohn, F. (1872b) Untersuchungen ilber Bakterien. Beitrage zur Biologie der Pflanzen, 1 (Heft 1), 127-224. p. 136

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., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 33: D294–D296.

Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam,S.A., McGarrell,D.M., Garrity,G.M. and Tiedje,J.M. (2005) The Ribosomal Database Project(RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res., 33, D294–D296.

Cole, J. R. Chai, B., Farris, RJ, Wang, Q., Kulam-Syed-Mohideen, AS. McGarrell, D. M, . Bandela, AM., Cardenas, E., Garrity, GM., and Tiedje, J. M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data, Nucleic Acids Research, Vol. 35, Database issue D169–D172

D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. Nucleic Acids Research, 2007, Vol. 35, Database issue D171 et al. (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.

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. (1833-1835) Dritter Beitrag zur Erkenntniss grosser Organisation in der Richtung des kleinsten Raumes. Physikalische Abhandlungen der Koeniglichen Akademie der Wissenschaften zu Berlin aus den Jahren 1833-1835, 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].

Erwin F. Smith (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.

Felsenstein,J. (1981) J. Mol. Evol., 17, 368-376.

Gutell RR, Larsen N, Woese CR. (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective, Microbiol Rev. Mar;58(1):10-26.

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.

Kohler,T., Sting, U., Meuser,K. and Brune,A. (2008) Novel lineages of Planctomycetes densely colonize the alkaline gut of soil-feeding termites (Cubitermes spp.). Environ. Microbiol., 10, 1260–1270.

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

Morris,R.M., Vergin,K.L., Cho,J.C., Rappe,M.S., Carlson,C.A. and Giovannoni,S.J. (2005) Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the bermuda atlantic time-series study site. Limnol. Oceanogr., 50, 1687–1696.

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.]

Olsen, G.J., Matsuda,H., Hagstrom,R., and Overbeek,R. (1994) Comput. Appl. Biosci., 10, 41-48.

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.

Parte AC. (2014) LPSN—list of prokaryotic names with standing in nomenclature. Nucleic Acids Res. 42(Database issue):613–6.

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.

Pruesse E, Peplies J, Glöckner FO. Sina (2012) Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28(14):1823–9.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig,W., Peplies, J. and 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.

Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole1 (2007) Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy, Applied And Environmental Microbiology, Vol. 73, No. 16, p. 5261–5267

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J.,Schweer, T., Yarza, P., Peplies J., and Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Research, Vol. 41, pp. 590-596.

Rice, C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) Nucleic Acids Res., 21, 2967-2971.

Sanderson MJ, Donoghue MJ, Piel W, Eriksson T. (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am J Bot. 81(6):183. 28. Dryad. Dryad Digital Repository. http://datadryad.org. Accessed 11 Oct 2016.

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

Takai,K., Moser,D.P., DeFlaun,M., Onstott,T.C. and Fredrickson,J.K. (2001) Archaeal diversity in waters from deep South African gold mines. Appl. Environ. Microbiol., 67, 5750–5760.

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

Vos,P.D., Garrity,G.M., Jones,D., Krieg,N.R., Ludwig,W., Rainey,F.A., Schleifer,K.H. and Whitman,W.B. (2009) The Firmicutes. Springer, New York.

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.

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.