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

Artificial Neural Networks 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, the face of the microbiology dramatically changed with the rRNA-based phylogenetic classifications, by Carl Woese. He delineated the three main branches of life. He used the technique 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 they obtained high accuracies using available datasets of their time. Recently the number of known bacteria increased enormously. In this article we used ANN's to annotate bacterial 16S rRNA gene sequences from five selected phylums in Greengenes database taxonomy: Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria, and Chloroflexi. 93% average accuracy is obtained in classifications. When we used the bundle testing technique, the average accuracy easily raised to 100%.

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 identified as the causes of human and animal diseases. They also 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, and other living creatures.

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 because most bacteria cannot be cultured or isolated under laboratory conditions (Ash et. al., 1991). In the past few decades, DGGE, Denaturing gradient gel electrophoresis,

Until the development of high-throughput sequencing technology, (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. 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).

Early Formal Classifications[1]

In the year 1866, Ernst Haeckel, in the Tree of Life in Generelle Morphologie der Organismen (Haeckel, 1867) Bacteria are first classified as plants constituting the class Schizomycetes, which along with the Schizophyceae formed the phylum Schizophyta. Haeckel 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 Vibrio, a genus of comma shaped bacteria first described in 1854 (Pacini, 1854). The genus Bacterium was a taxon described in 1828 by Christian Gottfried Ehrenberg (Ehrenberg, 1828). Ehrenberg also described Spirillum, spiral shaped bacteria in 1832 (Ehrenberg, 1832), Bacillus, a genus of spore-forming rod shaped bacteria in 1835, and Spirochaeta, thin spiral shaped bacteria in 1835 (Ehrenberg, 1835).

The classification of Cohn (1872) was influential in the nineteenth century, and recognized six genera: Micrococcus, Bacterium, Bacillus, Vibrio, Spirillum, and Spirochaeta (Murray, and Holt, 2005).

In 1905 Erwin F. Smith accepted 33 valid different names of bacterial genera and over 150 invalid names, (Smith 1905) and in 1913 Paul Vuillemin (Vuillemin, 1913) in a study 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.

Ferdinand Cohn (Cohn, 1875) recognized 4 tribes: Sphero-bacteria, Microbacteria, Desmobacteria, and Spirobacteria. Stanier, and van Niel, (Stanier, and van Niel, 1941) recognized the Kingdom Monera with 2 phyla, Myxophyta and Schizomycetae, the latter comprising classes Eubacteriae, 3 orders, Myxobacteriae, 1 order, and Spirochetae, 1 order. Bisset (Bisset, K. A. 1962) distinguished 1 class and 4 orders: Eubacteriales, Actinomycetales, Streptomycetales, and Flexibacteriales. Migula, (Migula, 1897), which was the most widely accepted system of its time and 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 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.

Different authors often reclassified the genera due to the lack of visible traits to go by, in different ways throughout the bacteria classification history. A poor state is resulted as summarized by Robert Earle Buchanan in 1915 (Buchanan, 1916). By then, the whole group continuing to receive different ranks and names by different authors.

Relatively recently, in 1980s, Carl Woese dramatically changed the face of the microbiology with his rRNA-based phylogenetic classifications. It delineated the three main branches of life (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. Phylogenetic methods are mostly clustering techniques, while 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 classification structure (taxonomy). Several recent events have helped change this situation (Wang, et. al., 2007).

For phylogenetic analysis, the 16S rRNA gene sequence was first used in 1985 (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 TODAY

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), sequencing 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:

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

Taxonomic Classifications

Each of the five taxonomies that we compare 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
Greengenes	Automatic	3,093	Species	2013
SILVA	Manual	12,117	Genus	2017
RDP	Semi	6,128	Genus	2016
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 and SILVA only go down to the genus level, whereas NCBI and OTT go down to the species level and below. The two latter taxonomies also have a number of intermediate ranks and contain many intermediate nodes.

Because of the known inconsistencies of microbial classifications (Beiko, 2016), the choice of reference taxonomy is important. For our purposes we have chosen the Greengenes taxonomy.

2.1 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 gene 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. 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 five years.

Table 3. Levels and number of sublevels in Greengenes

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

3. A BRIEF NOTE ON ANNS

This brief presentation of artificial neural networks will focus on a particular structure of ANNs, multi-layer feedforward networks, which is the most popular and widely-used network paradigm in many applications including forecasting volatilities and prices in markets. For a general introductory account of ANNs, readers are referred to Wasserman (1989); Hertz et al. (1991); Smith (1993). Rumelhart et al. (1986a), (1986b), (1994), (1995); Lippmann (1987); Hinton (1992); Hammerstrom (1993); Haykin 1999 illustrate the basic ideas in ANNs.

3.1 Recurrent Neural Networks (RNN)

Financial time series mostly dependent nonlinearly on time and hence recurrent neural networks (RNN) are particularly useful (Szkoła, et al, 2011; Lipton, 2015). They are constructed by taking a feedforward network and adding feedback connections from output and/or hidden layers to input layers. The standard backpropagation algorithm also trains these networks conditional that patterns must always be presented in time sequential order. The one difference in the structure is that there are extra neurons in the input layer that is connected to the hidden layer and/or output layer just like the other input neurons. These extra neurons hold the contents of one of the layers as it existed when the previous pattern was trained. In this way, the network takes into account previous knowledge it has about previous inputs. These extra neurons are called the context unit and it represents the network's long-term memory (Balkin 1997).

There are three types of RNNs: Jordan, Elman, and Jordan/Elman recurrent networks. A Jordan neural network (JNN) has additional neurons in the input layer, which are fed back from output layer (Carcanoa, et al, 2011). While an Elman neural network (ENN) has additional neurons in the input layer, which is fed back from hidden layer (Elman, 1990). The mixture of the two, Jordan/Elman recurrent networks (JENN) has additional neurons in the input layer, which is fed back from hidden layer, and output layer Demir, and Can, 2018).

3.2 Jordan Recurrent Neural Networks (JNN)

A Jordan neural network (JNN) has several feedback connections from the output layer to the input layer. The input layer has additional neurons, which are fed back from the output layer (Carcanoa, et al, 2011).

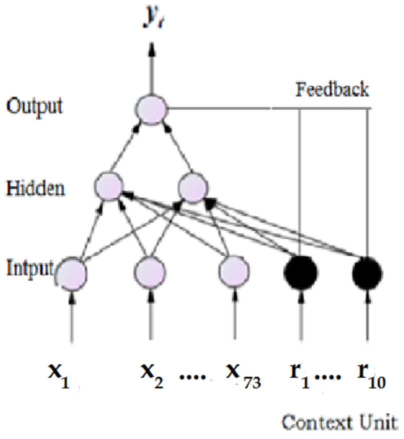

Figure 1. JNN with a single hidden layer representing a nonlinear regression model

4. MATERIALS AND METHODS

To train and test RNN's, from the 86 phylum level class of the Greengenes taxonomy, five phylum classes are chosen.

Table 1. Names of the Phylums chosen, and numbers of bacteria in those phylums

	Name of the Phylum	#Bacteria
1.	Proteobacteria	60.827
2.	Firmicutes	55.677
3.	Bacteroidetes	25.811
4.	Actinobacteria	15.711
5.	Chloroflexi	4.682

From each of these phylums, 300 bacteria are chosen randomly. 100 for training, 100 for validation and 100 for testing.

4.1 Coding Data

To transform the bacteria data into numerical values a coding method is adapted. First we decided about a word length 4^k. First disregarding our computational limitations, we have chosen k=2. The number of all possible words of length $4^2=16$, which are written by a four letter alphabet {A,T,C,G} is $4^{16} = 256^4 = 4.294.967.296$. Data vectors of that length are out of the range of our computers. Therefore we have chosen k=1. In that case the number of all possible words of length 4, which are written by a four letter alphabet {A,T,C,G} is $4^4 = 256$ which is reasonable.

For each bacterium, we prepared an empty grid with 256 pockets. We start by the first possible four letter word AAAA, count the appearances of this word in the sequence of the chosen bacteria, and write this number as the first component of the 256 dimensional vector. Repeating this procedure for all other 255 possible words, the bacteria sequence is transformed into a 256 dimensional vector of integers. Finally these vectors are divided by the maximum ingredient in the corpus to normalize.

1	0	16	2	...	35	9	23	1	7

Figure 2. 256 dimensional vector of integers representing the sequence of a bacterium.

4.2 Training RNN's

Artificial Neural Networks are the best when they are used to distinguish two groups. To distinguish these five bacteria phylums, we train 20 neural networks which will distinguish bacteria from phylum i, and phylum j; i,j=1,2,3,4,5 i≠j.

During the training alongside the training error, error on the validation set is also plotted. To avoid over learning, training is early stopped when error on validation set starts getting bigger than the error on the training data.

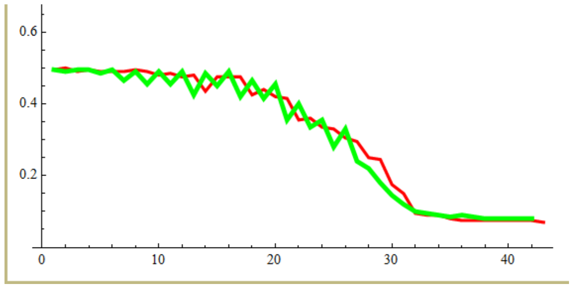

Figure 3. To avoid over learning, training is early stopped when error on validation set (Green) starts getting bigger than the error on the training data (red)

4.3 Aggregating the Decisions of RNN's

When a bacteria from one of the five phylums is presented the twenty recurrent neural networks, assuming the RNN $a_{i,j}$ is trained to distinguish the two bacteria one from phylum i, and one phylum j; i, j=1,2,3,4,5 i≠j, his vote is interpreted as follows: If he votes +1, it means that his vote is in the direction that the bacterium is from phylum i, while If he votes -1, then his vote is that, the bacterium is from phylum j.

For each bacteria introduced to the committee of RNN's, twenty votes are observed. The commonest of these votes is the aggregated vote of the committee.

The success of RNN's in distinguishing the bacteria phylums that they are trained for is seen in Table 4.

Table 4. Successes of RNNs a_{ij} j; i, j=1,2,3,4,5 i≠j trained to distinguish bacteria of phylum i, from bacteria of phylum j.

j	a1j	a2j	a3j	a4j	a5j	aj1	aj2	aj3	aj4	aj5
1	0	91	94	94	89	0	86	100	99	87
2	90	0	93	99	89	95	0	91	96	94
3	96	92	0	92	96	95	90	0	88	92
4	78	92	100	0	89	92	100	89	0	92
5	95	96	100	89	0	93	100	76	96	0

When votes are aggregated bacteria of phylum i, are correctly classified with the accuracy as in Table 5.

Table 5. Aggregated votes classifies bacteria of phylum classes with accuracies in the below.

Name of the Phylum	Accuracy %
Proteobacteria	97
Firmicutes	91
Bacteroidetes	93
Actinobacteria	88
Chloroflexi	96

4.4 Bundle Decision

When bacteria are collected from nature or from patient samples, first DNA and genes responsible from the coding of 16S rRNA are isolated. Then these isolates are clustered for annotation. Therefore it is natural that we may have several bacteria (bundle) are recruited from the same cluster. Then the question becomes: this bundle of bacteria are from the same class. Which class is it?

To aggregate the decisions of experts, let us send a bundle of nine bacteria from one of the phylums to the classifier. If the decision for these nine is like

B	1	2	3	4	5	6	7	8	9	MC
D	1	5	3	4	5	5	3	5	5	5

Then we decide that the bundle comes from the phylum of class five, since the most common vote is five.

We can repeat the same aggregation technique several times, we aggregate votes again by the "most common vote" understanding. Let us repeat the above bundle decision seven times; we may get a voting like:

R	1	2	3	4	5	6	7	MC
D	4	5	5	4	5	5	1	5

Then we decide that the bundle comes from the phylum of class five, since the most common vote is five.

In our experiment, we have got always 100% accuracy almost at all cases.

5. RESULTS

The coded data for bacteria is a very successful feature set for classification of bacteria into Greengenes taxonomy phylum classes. We have got the accuracy as in Table 5. At the cases where bundle decision is applicable, the success is almost 100%.

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.

Balkin, S.D. (1997). Using recurrent neural networks for time series forecasting, Working Paper Series number 97–11, International Symposium on Forecasting, Barbados 2–54.

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

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.

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

Carcanoa, E.C., Bartolinia, P., Muselli, ., and Piroddi, L., (2008) Jordan recurrent neural network versus IHACRES in modelling daily streamflows, Journal of Hydrology, Volume 362, Issues 3–4, 5 December, Pages 291-307

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

Demir, N. M., and Can, M. (2018) Authorship Authentication of Short Messages from Social Networks Using Recurrent Artificial Neural Networks, Southeast Europe Journal of Soft Computing Vol.7 No.1 March 2018 (25-30)

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.

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.

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].

Elman, J. L. (1990), 'Finding structure in time', Cognitive Science, 14, 179-211.

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.

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.

Hammerstrom, D. (1993) Neural networks at work, IEEE Spectrum, June, 26–32.

Haykin, S. (1999) Neural Networks, a Comprehensive Foundation, Prentice Hall.

Hertz, J., Krogh, A., Palmer, R.G. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley, Reading, MA.

Hinton, G. E., (1992) How Neural NetworksLearn from Experience, Scientific American, September 1992, pp. 145-151.

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.

Lippmann, R.P., (1987) An introduction to computing with neural nets, IEEE ASSP Magazine, April, 4–22.

Lipton, Z.C. (2015) A Critical Review of Recurrent Neural Networks for Sequence Learning, University of California, San Diego, June 5th.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 Balvociute, M., and Huson, DH. (2017) SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?, BMC Genomics 2017, 18 (Suppl 2):114

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.

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

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986) Learning internal representation by back-propagating errors. In: Rumelhart, D.E., McCleland, J.L., the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, MA.

Rumelhart, D.E., Widrow, B., Lehr, M.A. (1994) The basic ideas in neural networks. Communications of the ACM 37 (3), 87–92.

Rumelhart, D.E., Durbin, R., Golden, R., Chauvin, Y. (1995) Backpropagation: the basic theory. In: Chauvin, Y., Rumelhart, D.E. (Eds.), Backpropagation: Theory, Architectures, and Applications. Lawrence Erlbaum Associates, New Jersey, pp. 1–34.

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.

Smith, M. (1993) Neural Networks for Statistical Modeling, Thomson Learning ©1993

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

Szkoła, J., Pancerz, K., and Warchoł J., (2011) Recurrent Neural Networks in Computer-Based Clinical Decision Support for Laryngopathies: An Experimental Study, Comput Intell Neurosci.

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.

Wasserman, P.D. (1989) Neural Computing: Theory and Practice. Van Nostrand, Reinhold, New York.

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