Artificial Neural Networks in Bacteria Taxonomic Classification

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

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

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 Monerans 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, Metaprimerista, and Orla-Jensen, (Orla-Jensen, 1909) which divided Class Schizomycetes into 4 orders: 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 summarized as those that were acceptable 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. RNA-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.

**Table 1** Overview of five taxonomic classifications

<table>
<thead>
<tr>
<th>Taxonomy</th>
<th>Type</th>
<th>modes</th>
<th>Lowest</th>
<th>Latest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greengenes</td>
<td>Automatic</td>
<td>3,093</td>
<td>Species</td>
<td>2013</td>
</tr>
<tr>
<td>SILVA</td>
<td>Manual</td>
<td>12,117</td>
<td>Genus</td>
<td>2017</td>
</tr>
<tr>
<td>RDP</td>
<td>Semi</td>
<td>6,128</td>
<td>Genus</td>
<td>2016</td>
</tr>
<tr>
<td>NCBI</td>
<td>Manual</td>
<td>1,522,150</td>
<td>Species</td>
<td>2017</td>
</tr>
<tr>
<td>OTT</td>
<td>Automatic</td>
<td>2,627,066</td>
<td>Species</td>
<td>2016</td>
</tr>
</tbody>
</table>

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 CyanobDB (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**

<table>
<thead>
<tr>
<th>Levels</th>
<th># Sublevels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>86</td>
</tr>
<tr>
<td>Class</td>
<td>232</td>
</tr>
<tr>
<td>Order</td>
<td>366</td>
</tr>
<tr>
<td>Family</td>
<td>466</td>
</tr>
<tr>
<td>Genus</td>
<td>1949</td>
</tr>
<tr>
<td>Species</td>
<td>2389</td>
</tr>
</tbody>
</table>

**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 (Szkola, 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).

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.

<table>
<thead>
<tr>
<th>Name of the Phylum</th>
<th>#Bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Proteobacteria</td>
<td>60.827</td>
</tr>
<tr>
<td>2. Firmicutes</td>
<td>55.677</td>
</tr>
<tr>
<td>3. Bacteroidetes</td>
<td>25.811</td>
</tr>
<tr>
<td>4. Actinobacteria</td>
<td>15.711</td>
</tr>
<tr>
<td>5. Chloroflexi</td>
<td>4.682</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Name of the Phylum</th>
<th>#Bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Proteobacteria</td>
<td>60.827</td>
</tr>
<tr>
<td>2. Firmicutes</td>
<td>55.677</td>
</tr>
<tr>
<td>3. Bacteroidetes</td>
<td>25.811</td>
</tr>
<tr>
<td>4. Actinobacteria</td>
<td>15.711</td>
</tr>
<tr>
<td>5. Chloroflexi</td>
<td>4.682</td>
</tr>
</tbody>
</table>

4.1 Coding Data

To transform the bacteria data into numerical values a coding method is adapted. First we decided about a word length \( k \). First disregarding our computational limitations, we have chosen \( k=2 \). The number of all possible words of length \( 2^4=16 \), which are written by a four letter alphabet \( \{A,T,C,G\} \) is \( 16^4 = 4,294,967,296 \). 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.

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 \( j \); \( i,j=1,2,3,4,5 \).

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.

4.3 Aggregating the Decisions of RNN’s

When a bacteria from one of the five phylums is presented to twenty recurrent neural networks, assuming the RNN \( a_{i,j} \) is trained to distinguish the two bacteria one from phylum \( i \), and one phylum \( j \); \( i,j=1,2,3,4,5 \) \( i\neq j \), 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>
<thead>
<tr>
<th>j</th>
<th>a1j</th>
<th>a2j</th>
<th>a3j</th>
<th>a4j</th>
<th>a5j</th>
<th>aj1</th>
<th>aj2</th>
<th>aj3</th>
<th>aj4</th>
<th>aj5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>91</td>
<td>94</td>
<td>94</td>
<td>89</td>
<td>0</td>
<td>86</td>
<td>100</td>
<td>99</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>0</td>
<td>93</td>
<td>99</td>
<td>89</td>
<td>95</td>
<td>0</td>
<td>91</td>
<td>96</td>
<td>94</td>
</tr>
<tr>
<td>3</td>
<td>96</td>
<td>92</td>
<td>0</td>
<td>92</td>
<td>96</td>
<td>95</td>
<td>90</td>
<td>0</td>
<td>88</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>92</td>
<td>100</td>
<td>0</td>
<td>89</td>
<td>92</td>
<td>100</td>
<td>89</td>
<td>0</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>95</td>
<td>96</td>
<td>100</td>
<td>89</td>
<td>0</td>
<td>93</td>
<td>100</td>
<td>76</td>
<td>96</td>
<td>0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Name of the Phylum</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteobacteria</td>
<td>97</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>91</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>93</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>88</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>96</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>B</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>R</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

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


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


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

Cohn, F. (1872b) Untersuchungen über Bakterien. Beiträge zur Biologie der Pflanzen, 1 (Heft 1), 127-224. p. 136
Cohn, F. (1875) Untersuchungen über Bakterien. II. Beiträge zur Biologie der Pflanzen 1: 141-207


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


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

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


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


