**Title, Center, Times New Roman, 14, Bold**

1\*First Author, 1Second Author and 1Third Author

1Faculty of Engineering and Natural Sciences,

International University of Sarajevo International University of Sarajevo,

Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,

Bosnia and Herzegovina

\**Corresponding Author*: name@ius.edu.ba

Article Info

*Article history:*

Article received on dd mm 20xx

Received in revised form dd mm 20xx

*Keywords:*

16S rRNA gene; Longest Common Subsequence; Taxonomic clustering; Snowball

**ABSTRACT:** To analyze complex biodiversity in microbial communities, 16S rRNA marker gene sequences are often assigned to operational taxonomic units (OTUs). The abundance of methods that have been used to assign 16S rRNA marker gene sequences into OTUs brings discussions in which one is better. Suggestions on having clustering methods should be stable in which generated OTU assignments do not change as additional sequences are added to the dataset is contradicting some other researches contend that the methods should properly present the distances of sequences is more important. We add one more de novo clustering algorithm, Rolling Snowball to existing ones including the single linkage, complete linkage, average linkage, abundance-based greedy clustering, distance-based greedy clustering, and Swarm and the open and closed-reference methods. We use GreenGenes, RDP, and SILVA 16S rRNA gene databases to show the success of the method. The highest accuracy is obtained with SILVA library.

# INTRODUCTION

Metagenomics is a recently-born and highly popular field that studies the genomic contents of microbial communities living in certain environments and tries to understand the structure and function of these microbial communities by sequencing genomic fragments from environmental samples without the need of cultivating them in a laboratory [1], [2]. The microbiome is considered to be the "dark matter of the

biological universe" as most of the microorganisms are very difficult to culture and are still unknown [5]. Reconstructing the taxonomic composition of a bacterial community has a critical role in understanding that such a community might play an important role in affecting change in that environment and in creating different types of medicinal drugs and determining different types of functions both in plant and animal kingdoms.

In the early metagenomic studies, the sequencing of a complete 16S rRNA gene was a common approach using the traditional Sanger sequencing methodology [6], [7]. Although this approach was informative, it is expensive and provides a limited depth of sequencing in discovering the complete bacterial diversity that exists in a complex environment.

# PROBLEM STATEMENT

The main problem with existing methods in taxonomy prediction, OTU clustering, and denoising is the tradeoff between computational time and accuracy. The length of short reads has a huge impact on this challenge. Furthermore, the best performing tools often may not be open-sourced and free.

NGS technologies provide short reads and huge sequencing depth at a much lower cost. Hence, recent metagenomic projects shift to focus on the sequencing of only a single or combination of two or more hypervariable regions. Therefore, specialized tools are needed for highly accurate taxonomic classification of species using these short length sequences.

# BACKGROUND

Many algorithms have been developed for taxonomy prediction such as RDP Naive Bayesian Classifier (NBC) [8], GAST, 16Sclassifier [9], SPINGO, Metaxa2, SINTAX, PROTAX [10], microclass. There are also implemented methods in MOTHUR, QIIME v1 and QIIME v2.

RDP-Classifier which uses a Naive Bayesian Classifier is one of the most commonly used tools. It is highly accurate on complete 16S rRNA sequences but suffers in accuracy for targeted HVRs which are short in length.

Table 1 Some Encoding Schemes [11]

|  |  |
| --- | --- |
| **Scheme Name** | **Discrete numeric values** |
| **Atomic Number** | C=58, T=66,A=70, G=78 |
| **EIIP** | C=0.1340, T=0.1335, A=0.1260,G=0.0806 |
| **Molecular Mass** | C=111.1, T=112.1, A=135.13,G=151.13 or C=110, T=125, A=134, G=150 |
| **Thermodynamics** | TC=5.6, GA=5.6, CA=5.8, TG=5.8, TA=6.0, AC=6.5, GT=6.5, CT=7.8,AG=7.8, AT=8.6, TT=9.1, AA=9.1, CC=11.0, GG=11.0, GC=11.1, CG=11.9 |
| **Three-group** | (1) R={A, G}, Y={C, T} , (2) M={A, C}, K={G, T}, (3) W={A, T}, S={G, C} |
| **Dinucleotide** | Sixteen dinucleotides are mapped to a unit circle. |
| **Ring Structure** | AG: (0, 1.5), CT: (0, -1.5), CA:(1, 1), TG: (-1, -1), CG: (1, -1), TA: (-1, 1), GA: (1,0), GT(0.5, -1.25), GC: (-0.5, -1.25), TC:(-1, 0), AC: (-0.5, 1.25), AT: (0.5, 1.25),AA: (0, 1), TT: (0.5, 0), GG: (0, -1), CC:(-0.5, 0). |

As presented in Figure 1, the models shows different behaviours at the beginning of the traning, while the rest of the performace is similar for all models.

(1)

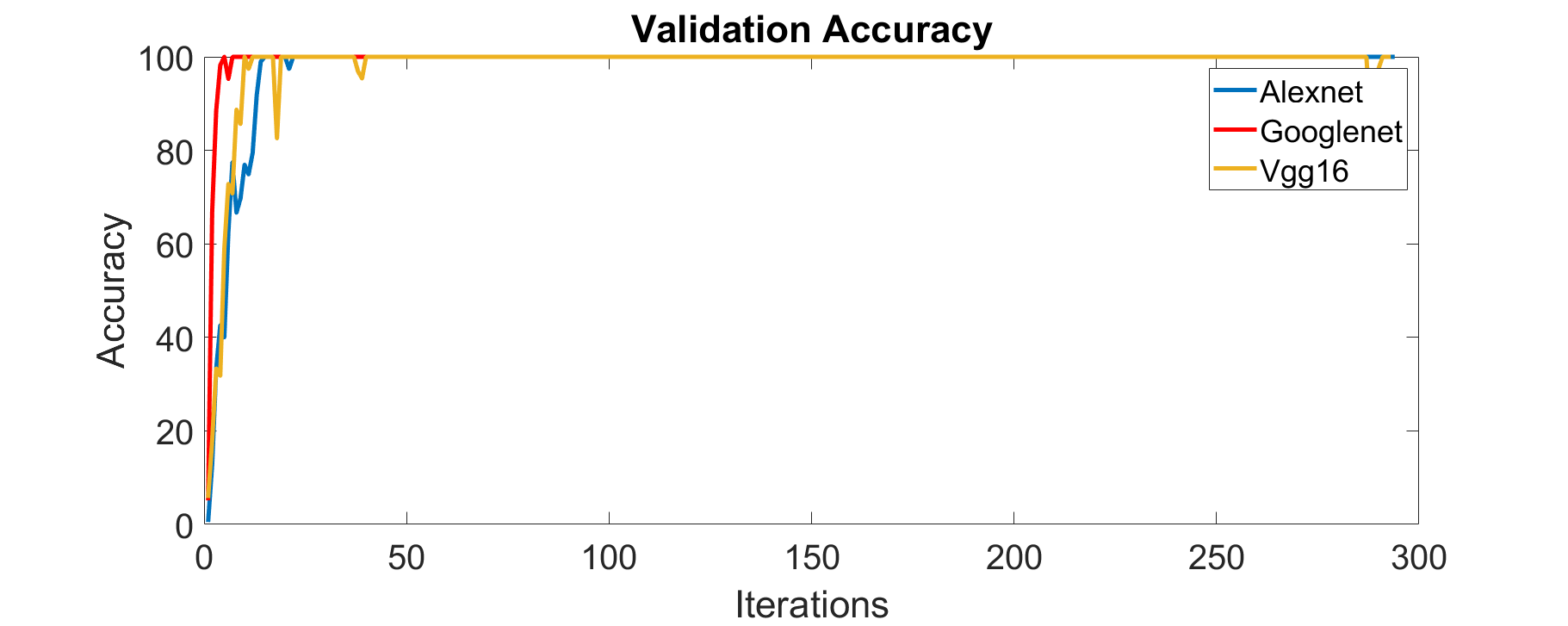


Figure 1: Validation accuracy

# MATERIALS AND METHODS

Existing 16S rRNA Reference databases Greengenes, SILVA, and RDP are used.

Sequences are converted to genomic signals with complex numbers encoding scheme [i,-i,1,-1] and randomly selected 50 taxa each having 50 sequences from the genus level are used to compute in-class and inter-class similarities. The average similarities are 69.94% and 25.37% respectively. The difference between in-class / inter-class similarities is very promising, and such a similarity measure results in good taxonomy prediction accuracy and specificity in OTUs clustering.

Preliminary results for the SILVA database at the genus level, show distinguishable in-class, inter-class similarities

## Data

Gene databases independently get updated and have a different approach to taxonomy construction. Taxonomy is ranked as kingdom/domain, phylum, class, order, family, genus, and species levels. RDP has no species level and has additional subclass and suborder levels.

# REFERENCES, *IEEE*

[1] J. Qin *et al.*, “A human gut microbial gene catalogue established by metagenomic sequencing,” *Nature*, vol. 464, no. 7285, pp. 59–65, 2010, doi: 10.1038/nature08821.

[2] C. Huttenhower *et al.*, “Structure, function and diversity of the healthy human microbiome,” *Nature*, vol. 486, no. 7402, pp. 207–214, Jun. 2012, doi: 10.1038/nature11234.

[3] G. Bernard, J. S. Pathmanathan, R. Lannes, P. Lopez, and E. Bapteste, “Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery,” *Genome Biol. Evol.*, vol. 10, no. 3, pp. 707–715, Mar. 2018, doi: 10.1093/gbe/evy031.

[4] C. Lok, “Mining the microbial dark matter,” *Nature*, vol. 522, no. 7556, pp. 270–273, 2015, doi: 10.1038/522270a.

[5] E. Kellenberger, “Exploring the unknown,” *EMBO Rep.*, vol. 2, no. 1, pp. 5–7, Jan. 2001, doi: 10.1093/embo-reports/kve014.

[6] J. F. Petrosino, S. Highlander, R. A. Luna, R. A. Gibbs, and J. Versalovic, “Metagenomic pyrosequencing and microbial identification,” *Clin. Chem.*, vol. 55, no. 5, pp. 856–866, 2009, doi: 10.1373/clinchem.2008.107565.

[7] L. Dethlefsen, S. Huse, M. L. Sogin, and D. A. Relman, “The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing,” *PLoS Biol.*, vol. 6, no. 11, pp. 2383–2400, 2008, doi: 10.1371/journal.pbio.0060280.

[8] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, “Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy,” *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, Aug. 2007, doi: 10.1128/AEM.00062-07.

[9] N. Chaudhary, A. K. Sharma, P. Agarwal, A. Gupta, and V. K. Sharma, “16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets,” *PLoS One*, vol. 10, no. 2, pp. 1–13, 2015, doi: 10.1371/journal.pone.0116106.

[10] P. Somervuo, S. Koskela, J. Pennanen, R. Henrik Nilsson, and O. Ovaskainen, “Unbiased probabilistic taxonomic classification for DNA barcoding,” *Bioinformatics*, vol. 32, no. 19, pp. 2920–2927, 2016, doi: 10.1093/bioinformatics/btw346.

[11] and Z. Y. Ning Yu, Zhihua Li, “Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning,” *Big Data Min. Anal.*, vol. 1, no. 3, pp. 191–210, Sep. 2018, doi: 10.26599/BDMA.2018.9020018.