DOI 10.29042/2018-4007-4009

Helix Vol. 8(5): 4007- 4009

k-mer Profiling for Bacterial Identification

[*1]Snehal V. Bhange, [2]Hitesh Tikariha, [3] S. S. Dongre, [4] H. J. Purohit

[1, 3] G.H. Raisoni College of Engineering, Nagpur 440016, India

[2, 4] CSIR-NEERI, Nagpur 440020, India

**Email: bhange_snehal.ghrcemtechcse@raisoni.net*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

In this paper, the bacterial identification is done using k-mer profiling. The idea is to create a kmer histogram using R programming and map the occurrences of kmers occurring in common within the same bacteria genera of different species and their respective strains. 5-mer is used in this study for proving the hypothesis. The kmer profile generated clearly distinguishes the two bacteria and look similar for closely related strains. Thus a deep screening and profile matching will help in rapid identification of two bacteria by their kmer profile.

Keywords: Bacterial Identification, Genera, k-mer Profiling

Introduction

Bacterial identification is the prominent task in research field related to biological sciences. Therefore researchers keeps on finding novel ideas in this area for bacterial identification. In the last two decades many researchers have found many ways which have evolved in identification of bacteria. Of all these method, statisticians have developed another area for bacterial identification using statistics and data analysis. Since the data in biological size is extremely huge and complicated, the task become more difficult. To overcome this difficulties, this paper proposes the k-mer profiling for bacterial identification and henceforth find the even and odds in the various species and respective strains of same bacteria genera. Of all the strategies in bacterial identification, K-mer counting helps to show more flexible results when they are counted in the spectra of $5<k<20$. k-mer counting is initial stage in algorithm related to bioinformatics. k-mer counting is very simple and effective means to study the nature of repetition of subsequences in genomic sequences. In this paper, the major inclination is given toward the use of R Programming[1]. Since R provides the feasible platform where manipulations are effectively handled also it provides with different packages which provides many different functions. R can import files of varies sizes and types. Also, package Biomartr helps to download genome [2] directly with less time consumption. K-mer profiling can be done in various forms like histograms, scatterplot etc. Histograms of k-mer frequencies can help in many ways and give valuable insights into the underlying distribution and indicate the error rate and genome size sampled in the sequencing experiments. k-mer counting helps in generalized statistics which may convey much information about the abundance of data. k-mer finds its way in many advance algorithms like KMC Tallymer, Jellyfish [3], BFCounter, DSK, KMC, Turtle and KAnalyze etc. k-mer generates the prominent characteristics of each genome and finds its way to the comparison with other related genome. Thus the subsequences of each genera is equal to the other genera except that the fluctuation is seen in the diverse numbers of each genera taken. The k-mer thus helps in beneficial statistics of occurrences of each genome.

Methodology:

The whole methodology follows three major steps as explained below:

Step1: The foremost step indulges with importing of whole genome file into R console using 'seqinr' [5] package.

Step2: Defining the 'k=5' and counting the occurrences of those subsequences in the complete genome.

Step3: Segregating the top 20 5-mers among all 5mers.

Step4: Visualizing the kmer profiling [4] using R package

Result & Discussion

The study was carried on the genome of 6 bacteria. To check the profiling of kmers, the results shown below is of the pseudomonas bacteria and Acidovorax bacteria. The studies was carried for two species i.e. syringae and putida of Pseudomonas and two different strains of Acidovorax bacteria. In each species the different strains were also studied. And the 'k' value was taken as 5. Therefore 5mers was carried for the whole statistical analysis

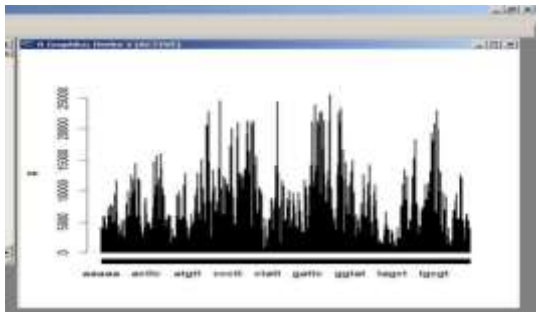

Fig 1a: 5mer Plot of *Pseudoputida*_KT2440

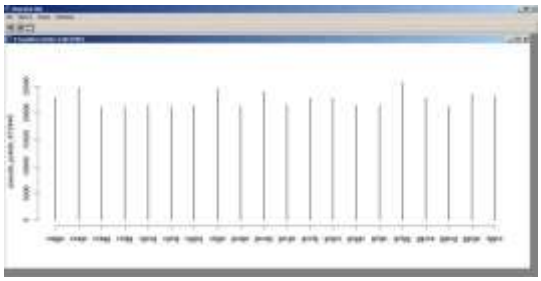

Fig 1b: 5mer Plot of Top 20 Occurrences of Pseudoputida_KT2440

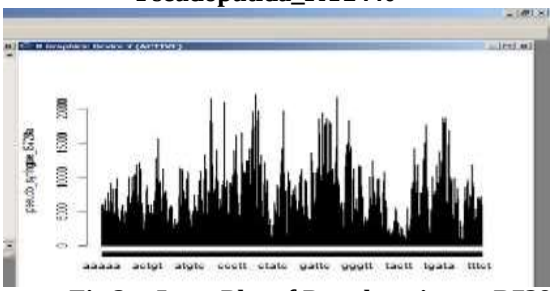

Fig 2a: 5mer Plot of Pseudosyrinage_B728a

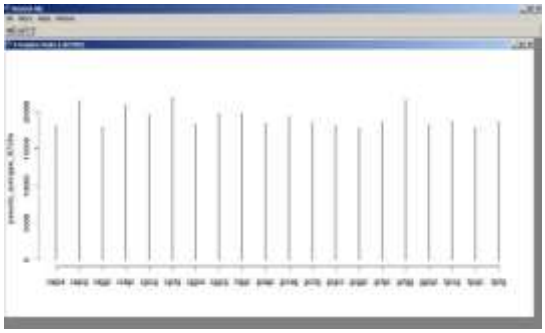

Fig 2b: 5mer Plot of Top 20 Occurrences of Pseudosyrinage_B728a

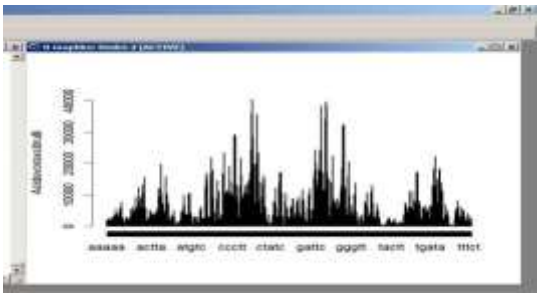

Fig 3a: 5mer Plot of *Acidovorax citrulli*_ KACC17005

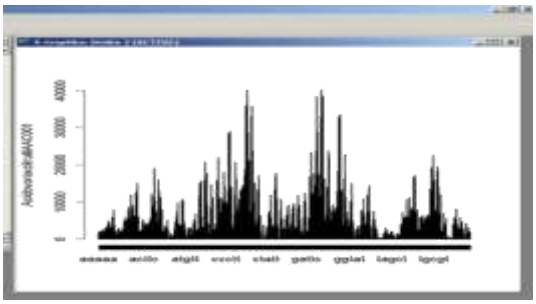

Fig 3b: 5mer Plot of *Acidovorax citrulli*_AAC001

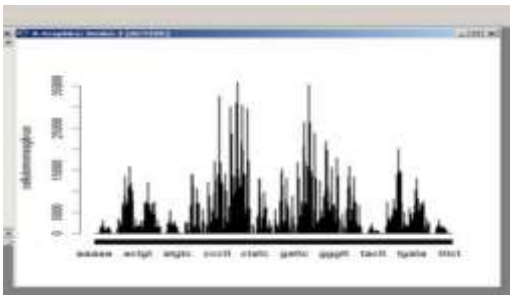

Fig 4a: 5mer Plot of *Cellulomonas gilvus _ATCC13127*

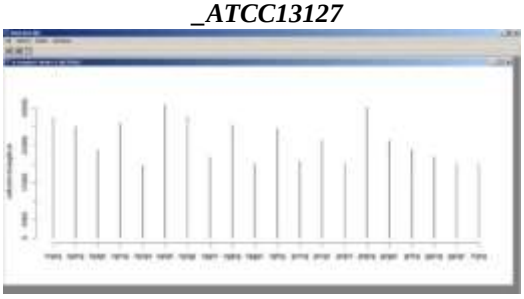

Fig 4b: 5mer Plot of Top 20 Occurrences of *Cellulomonas gilvus _ATCC13127*

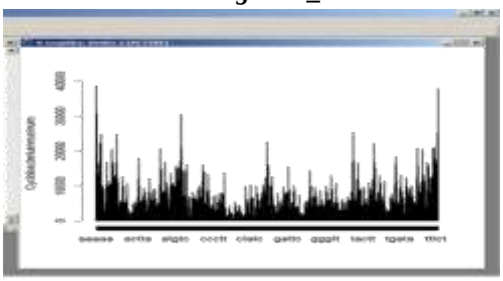

Fig 5a: 5mer Plot of *Cyclobacterium marinum*_ DSM 745

Fig 5b: 5mer Plot of Top 20 Occurrences of *Cyclobacterium marinum*_DSM 745

Fig 6: Color Coding Chart of Similar 5mers

Discussion

The Fig. 1a), 2a) are histograms generated for two species of Pseudomonas bacteria complete genome. Fig 1b), 2b), are the 20 highest occurring 5mers in the genome. Whereas fig 3a) and 3b) depicts the kmer profile of Acidovorax bacteria. Therefore, from the above results, it is clear that the occurrence of 5mers of same bacteria genera has then resemblance with each other in their species. And therefore, when compared with two different genera of bacteria, kmer profiling is entirely different from each other. As the sequence imparts specific properties to a bacteria, there variation result in emergence of a new bacteria genera and species. The kmer profiling based on this simple concept will surely help in identification of bacteria and R- programming will clearly help in carrying out statistical analysis for generating kmer profile and their comparative analysis for distinguishes two or different bacteria.

Conclusion

Hence, the above study shows the statistical analysis of the complete genome and helps to analyse certain properties of respective bacteria based on the k-mer counting. This study on a large scale will surely assist in bacterial identification and categorizing as per their phylogeny.

References:

[1] http://www.bioconductor.org
[2] https://www.ncbi.nlm.nih.gov
[3]Abdullah-Al Madman,Soumitra Pal, Sanguthevar Rajasekaran, “Efficient Techniques for k-mer Counting” 2015.
[4] Dapeng Wang, “KGCAK: a K-mer based database for genome-wide phylogeny and complexity evaluation.” 2015.
[5] Ugo Bastolla; SeqinR 1.0-2: “A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis “.

