International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Metagenomic Classification using Centrifuge

Tanaya Jadhav, Meghana Nagori

Abstract*: To assess the quality of food is a major challenge the food industry faces today. It is of utmost importance to test it for contaminates and non-edible material that may be present. To overcome these challenges metagenomic classification is majorly useful. Several researches involve various classification techniques and their studies. Difficulties in metagenomic classification include increasing number of genomes thereby requirement of computational methods to have high speed as well as high accuracy so as to compare DNA sequences to genomes. Centrifuge is a classification tool for quantification of species present in a sample so as to monitor the quality of the same. Given a food sample Centrifuge effectively classifies the species present in it enabling a timely and accurate analysis.

Index Terms: Burrows-Wheeler Transform, Centrifuge, FM index, Metagenomics, Metagenomic classification.

I. INTRODUCTION

Metagenomic classification has found potential applications as it is unbiased in giving the decision of "What is there?"- in a sample that may be clinically related to health of life forms on Earth. RefSeq database had 179 prokaryotic genomes in 2004 and it increased to 4278 in December 2015. With the ever increasing database of genomes there is a large history of Taxonomic classification algorithms. The requirement of the same is to have high speed and low index size. Machine learning based algorithms like the naïve bayes classifier[12] and PhymmBL[10] classify as few as 100 reads per minute. Another approach is the pseudo-alignment method employed in Kraken [8][9] that processes reads very rapidly with 1 million reads per minute but requires a very large index and a large RAM. Some algorithms like Bowtie [11] and BWA provide very fast alignment with a small memory size [1]. CLARK is a tool used to classify objects based upon reduced sets of k-mers [2]. It is a tool that can be used for taxonomic classification on desktop computers with low RAM(4 Gb RAM). Centrifuge is a metagenomic classification engine that allows fast and accurate labeling of reads and percentage calculation of species available on configuration as minimal as desktop computers.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Tanaya Jadhav*, ME student in Computer Science and Engineering Department of Government College of Engineering Aurangabad.
Meghana Nagori, Lecturer / Associate Professor / Data Science

The system uses an indexing scheme based upon the Burrows-Wheeler transform and the Ferragina-Manzini index, optimized specially for the metagenomic classification problem. Centrifuge builds a small index size as compared to other algorithms (5.8 GB for all entire bacterial and viral genomes and the human genome) and classifies at a high pace, thereby processing a millions of reads from a high throughput source of DNA sequencing run within a few minutes.
This allows classification to occur rapidly and accurately and in the specifications of conventional desktop computers.

II. METHOD AND MATERIAL

Centrifuge is designed in a manner such that the large set of genomes can be properly represented. Its indexing scheme is based on Burrows-Wheeler transform (BWT) and Ferragina-Manzini index (FM-Index).

1. Burrows-Wheeler Transform with FM-Index

BWT was a text compression algorithm which was later used for genomic classification applications [3]. Along with FM-index this algorithm has found applications in genomic industry. Following are the steps of the BWT algorithm:

1) *Rotate the string and write it down on a separate row*
2) *Sort the rotations alphabetically according to the first column.*
3) *The transformation result required is the last column of the strings obtained[3].*

Illustrating BWT with an example considering the string "BANANA$" ($ is the string terminator):

BANANA$
$BANANA
A$BANAN
NA$BANA
ANA$BAN
NANA$BA
ANANA$B

Table 1: Rotation

$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

Table 2: Sorting

Therefore, BWT(BANANA$)=ANNB$AA
BWT is particularly useful as it creates an array whose rows are formed by cyclic shifting of the target string, that may be a DNA sequence.
It is beneficial when the characters are clustered together, they are ordered which makes the string compressible. Also we can get back the original string without losing any data.
To construct the original string from the newly transformed string a table, known as FL table is maintained. The table consists of two columns F- Fist and L-Last. The L column stores the Burrows-wheeler Transformed string obtained and the F column has the lexicographically ordered BWT string.

1. FM-Index:

FM index is a technique to search the BWT string for matches. FM index has low memory requirements which makes it an efficient choice to use in genomic applications. FM index is a data structure for string searching and requires space which is a function of the entropy of the indexed data [4].
Considering a text T[1,n], the space occupied by FM index is atmost $5nH_k(T)+o(n)$ bits[5]. Wherein,
$H_k(T)$: kth order entropy of T.
It allows the search for *oc* occurrences of any pattern P[1,p] in text T[1,n] in time $O(p+oc*\log^{1+\varepsilon}n)$ where $\varepsilon<0$ [5].
All the above is validated and proven by experimental analysis in [5].
FM index is known to compress the data in equation with the best compressors and also search a string in hundreds of Megabytes in few milliseconds.

2. Centrifuge

Centrifuge uses an indexing scheme based on FM index. It considers the observation that a large number of closely related genomes are present for a particular organism. For example, Escherichia Coli has 131 genomes (as per the RefSeq dataset). Hence the algorithm further reduces the size by considering the identical parts of the genomes only ones.

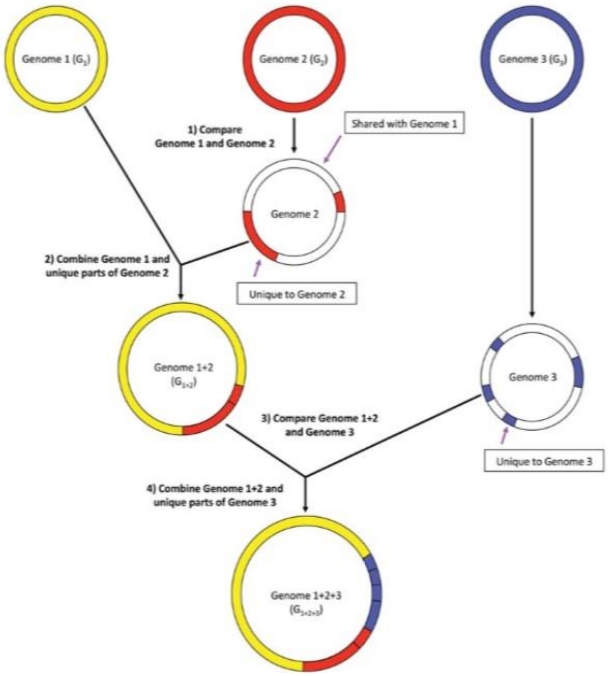

Figure 1: As Depicted in [1], this is the genome compression technique adopted by Centrifuge. All the similar genomes are combined into one single genome. This method runs on entire dataset of genomes used for metagenomic classification. Compression technique adopted by centrifuge helps reduction in the size of the genomes and this compression is known to have negligible impact on the results.

3. Database:

The sample files of reads is available from NCBI's SRR1745839 dataset and the set of genomes can be obtained from the RefSeq database[6]. Reference Sequence database is an integrated set of DNA sequences, transcripts and protiens.

III. ADVANTAGE OF FM-INDEX OVER K-MER BASED INDEXING SCHEMES

Many algorithms for metagenomic classification with k-mer based indexing schemes exist. The size that the k-mer table requires is huge, requiring large disk space. FM-index can run large and small k-mers by allowing very fast searches on k-mers, at speed as compared to that of k-mer table indexing schemes[1].

IV. CLASSIFICATION METHOD BASED ON FM INDEX

The algorithm searches both the read of given bp and the reverse complement of the same. It starts with a short matching of DNA read (a minimum of 16 base pairs). It stops wherever a mismatch occurs. Adopted from [1] is the example given in Figure 2 to demonstrate the concept. Matching happens starting from the right end.

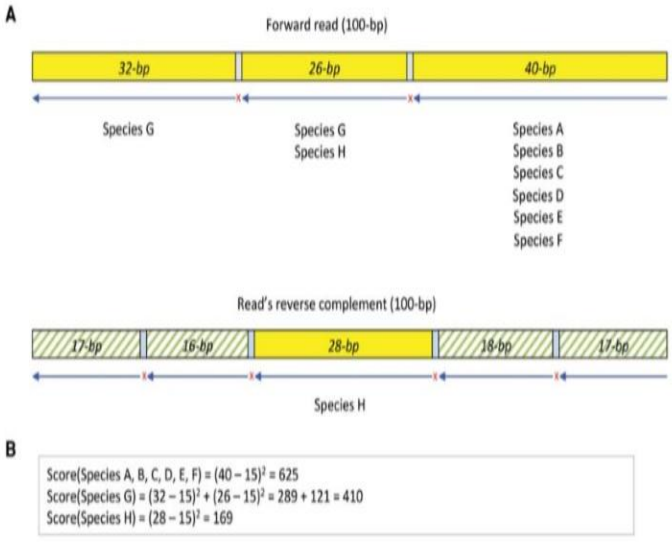

Figure 2

The example considers a 100bp read. The exact match happens till 40th base pair and there is a mismatch at 41st one. The 40 base pairs matched section matches with 6 species- A, B, C, D, E, and F, whose genomes are present in the database. The algorithm restarts at the 42nd read and continues matching until next mismatch is encountered which happens at number 68.

The second segment as shown is found in species G and H. The procedure goes on in the similar manner for the remaining base pairs. Similar algorithm is followed with the reverse complement of the string. (read).
Based on the exact matches of both the forward read and the reverse complement, classification takes place. As seen in the example, there are 3 exact matches in the forward read and one exact match in the reverse complement read. And the exact matching length of the read must be atleast 22 bp. There is a score given to each species in accordance with the following formula:

$$\text{Score(Species } X) = \sum_{\text{hit} \in \text{Species } X} (\text{length(hit)} - 15)^2.$$

Figure 3: Formula for calculation of score of a particular species.
The score varies proportionally to the matches. More are the matches of exact length, higher is the score, more likely is the presence of the particular species in the food sample. Figure 2 A shows the matching of the read and the read's reverse complement. Figure 2 B shows the score calculation. It is seen that species A, B, C, D, E, and F have high scores due to the presence of the long match of 40 bp.

V. WORKFLOW OF THE CENTRIFUGE ALGORITHM

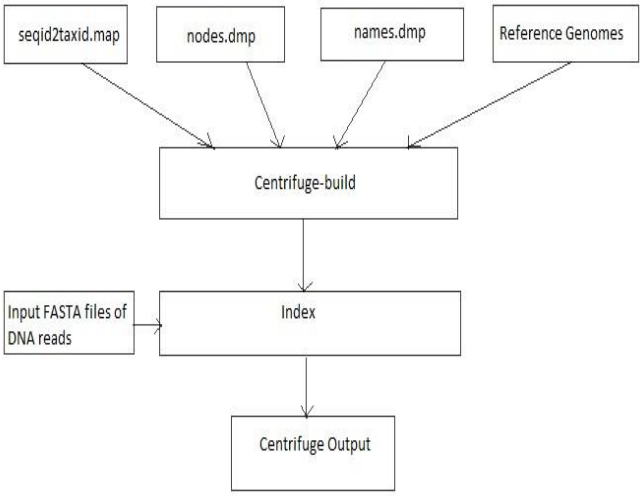

Figure 3: Workflow Of Centrifuge Algorithm

The input taken by the algorithm to build the index is:

1. Seq2taxid.map: Sequence ID to Taxonomic ID mappimg.
2. Nodes.dmp: Links parents to the taxonomic IDs.
3. Names.dmp: A file that is a link between taxonomic ID and scientific name [7].

Ones the index is built the reads are compared and output is produced in the form of scores.

VI. RESULTS

Classification Results of the sample taken for metagenomic classification:

Common name	Scientific name	Score
Cockroach	Blatella Germanica	15
Cow	Bos Taurus	686
Sheep	Ovis Aries	604
Pig	Sscrofa	76

Table 3: The results and scores allotted to the presence of respective organisms by Centrifuge. As per the scores obtained from Centrifuge it is seen that the sample contains Bos Taurus and Ovis Aries in larger quantity when compared to other species which are found in traces.

VII. APPLICATION

Metagenomic classification can be used as a qualitative measure of food hence can be very useful in the food industry. Presence of contaminants can be well assessed using the tools and techniques of metagenomic classification. Not only does it assess the presence of organisms like Cockroach , rat etc. that are non-edible but also can detect the presence of harmful microbes that may degrade the food quality. Therefore, for maintaining the food standard and quality metagenomic classification turns out to be an important tool.

REFERENCES

1. Daehwan Kim, Li Song, Florian P.reitwieser, Steven L. Salzberg,, "Centrifuge: Rapid and sensitive classification of metagenomic sequences", *Genome Research 2016*
2. Rachid Ounit, Steve Wanamaker, Timithy J Close, Stefano Lonardi, "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers", BMC Genomics, 2015
3. Ahmad Al Kawam, Sunil Khatri, Aniruddha Datta, "A Survey of Software and Hardware Approaches to Performing Read Alignment in Next Generation Sequencing", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, vol. 14, No. 6, 2016
4. Paolo Ferragina, Giovanni Manzini, "The FM-index: A compressed Full-Text Index Based on the BWT", DIMACAS, Aug. 19-20, 2004, 15-16
5. P. Ferragina and G. Manzini. "An experimental study of an opportunistic index", Proc. 12th ACM-SIAM Symposium on Discrete Algorithms, pages 269–278, 2001.
6. NCBI, (2017) "Refseq", [Online] Available: https://www.ncbi.nlm.nih.gov/refseq/
7. John Hopkins University, "Centrifuge" [Online] Available: https://ccb.jhu.edu/software/centrifuge
8. Derrick E Wood, Steven L Salzberg, "Kraken: Ultrafast metagenomic sequence Classification using exact alignments", BMC Genome Biology, 2014
9. NCBI, (2017) "Refseq", [Online] Available: http://ccb.jhu.edu/software/kraken/
10. Arthur Brady, Steven L. Salzberg," Phymm and PhymmBL: Metagenomic Phylogenetic Classification with Interpolated Markov Models", PMC US National Library of Medicine, 2009
11. John Hopkins University, "Bowtie", [Online] Available: http://bowtie-bio.sourceforge.net
12. Gail L Rosen, Erin R. Reichenberger and Aaron M. Rosenfeld," NBC: Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads", PMC US National Library of Medicine, 2010

AUTHORS PROFILR

Tanaya Jadhav is a ME student in Computer Science and Engineering Department of Government College of Engineering Aurangabad.

Dr. Meghana Nagori is a resourceful and dedicated Lecturer / Associate Professor / Data Science Specialist, with diverse experience and an exceptional record of sustained success; heading academic departments and driving forward best practice. A strategically focused leader and committed academic, passionate about ensuring the highest levels of teaching methodologies, who genuinely values opportunities to learn, improve and grow in pursuit of positive change. An accomplished manager, able to assess, define and execute the needs of teams; developing and training colleagues whilst heading academic projects and publishing research