11-26-2011

Binning Metagenomic Data by CSSR

Rekha Bhaskarabhatla

BINNING METAGENOMIC DATA BY CSSR

By

Rekha Bhaskarabhatla

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Biology

The University of Memphis

December 2011

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Tit-Yee Wong for his continued and patient support throughout the work of this thesis. His guidance has been truly invaluable and has helped me to pursue my career aspirations. I would like to thank other members of the Advisory committee, Dr. King-Thom Chung and Dr. Lih-Yuan Deng, for their guidance and support. My sincere thanks to Dr. Max Garzon, for providing me with continuous feedback and also helping me with Perl and Matlab software.

I feel very grateful to my mom, Hemalatha and dad, Nagaraju for their love and support and also for standing by me through all the highs and lows of my life. Finally, I owe it to my wonderful brother, Rakesh Sharma, without whose emotional support completing my thesis looked nearly impossible.

ABSTRACT

Bhaskarabhatla, Rekha. MS. The University of Memphis. December 2011. Binning Metagenomic Data by CSSR. Major Professor: Dr. Tit-Yee Wong.

Metagenomics is the study of microbes in their natural environments without the need for isolation and lab cultivation. The DNA fragments obtained from sequencing of a sample of mixed species requires taxonomic characterization called binning. My research concerns binning of metagenomic data using a novel approach. Each genomic sequence was codified based on their Cistronic Stop Signal Ratio (CSSR) values. Since the genic CSSR values of phylogenetically related organisms often share a definable pattern, a neural network was trained to recognize the genic CSSR patterns of known species. The trained neural network was then used to cluster the CSSR values from the metagenomic data. To show the validity of this method, a total of 15,000 genic CSSR values were calculated from five different bacterial species. The data was randomly mixed and a neural network was used to recognize the originality of these genes, based on their unique CSSR values. Result showed that better than 95% of the genes were correctly binned to the rightful species. The metagenomic sequences from the fecal samples of 124 individuals were reanalyzed based on the CSSR - neural network method by training the genic values of a set of known enteric bacteria. The resulting clusters were discussed.

TABLE OF CONTENTS

LIST OF FIGURES

Page

1. BACKGROUND AND SIGNIFICANCE

Microbes are the oldest form of life on earth. They were the only life form on earth for about the last 2 billion years (Kapur & Jain, 2003). Although too small to be seen by the naked eye, they have remained the driving force behind all life forms on earth (New, T. H. E., 2007). Microorganisms form the critical components of the Earth's biosphere through recycling the various elements and making them available to all other forms of life. Without these recycling activities of microorganisms, all the elements would be fixed and life on earth would not continue (Staley, J. T., Castenholz, R. W., Colwell, R. R., Holt, J. G., Kane, M. D., Pace, N. R., Salyers, A. A., et al., 1997).

Microbes are the most abundant and diverse form of life. Their diversity exceeds the biodiversity of plants and animals by several folds and is largely unknown (Hoff, K. J., Tech, M., Lingner, T., Daniel, R., Morgenstern, B., & Meinicke, P., 2008). Over the last few centuries, a huge number of bacterial species in or on humans have been observed under the microscope. Only recently, the idea of interdependence of microbes and humans is now becoming clear. The term 'microbial community' refers to the complex microbial ecosystems that are ubiquitous. Although microbial community is often considered to play an important role in our daily function, our understanding in microbial communities is limited by the lack of suitable tools and methodologies. The number of microbial cells that colonize human body has been estimated to exceed our own cell number by tenfold (New, T. H. E., 2007). This indicates that the number of unique genes that these microbes encode outnumber the number of genes in our own genome by at least 100 fold (Ley, R. E., Peterson, D. A., & Gordon, J. I., 2006; Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., et al.,

2010). However, less than one percent of the estimated trillions of microbial species in their natural habitats are culturable (New, T. H. E., 2007). Thus, only a tiny pool of their genes is accessible using the culture dependent methods. The remaining 99 plus percent is left inaccessible. In his book "Evolution by Association", Jan Sapp provides numerous examples of interactions between microbes and their hosts (Sapp, J., 1994). It becomes clear that the survival of all living organisms of higher orders is highly dependent on the microbes they harbor. Some scientists even describe humans and other "higher animals" as 'superorganisms' in which our body and a large number of different organisms coexist as one (Goodacre, R. 2007; Proal, A. D., Albert, P. J., & Marshall, T., 2009). The complex and dynamic microbiota inside and around our body plays a crucial role in many basic bodily processes. The association of microbes with our body has shown to have profound influence on human development, physiology, nutrition and immunity (Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I., 2005; Hooper, L. V., Midtvedt, T., & Gordon, J. I., 2002; Qin et al., 2010). Changes, either in the number or composition of these microbial communities may lead to diseases like inflammatory periodontitis (Cobb, C. M., 2008) and gastrointestinal disturbance (Khachatryan, Z. A., Ktsoyan, Z. A., Manukyan, G. P., Kelly, D., Ghazaryan, K. A., & Aminov, R. I., 2008). Hence, studying the dynamic and variable nature of the microbial community in our body may contribute to better diagnosis, and treatment of diseases (Yang, B., Peng, Y., Leung, H. C.-M., Yiu, S.-M., Chen, J.-C., & Chin, F. Y.-L., 2010).

To fully understand how our body functions, it is important to understand the number and types of microbes associated with ouSr human body. Traditional methods for studying microbes focused on isolating and analyzing single species in pure culture.

However, analyzing individual cultures in their pure forms would not provide the information needed to understand the interactions and dynamics between these microbes at the various niches (organs/tissues, etc.) of their host. Recent advances in molecular sequencing techniques allow scientists to reveal the identities of many otherwise non-culturable microbes in our body. Metagenomics is one of the novel techniques that could revolutionize our understanding of the unexplored microbial communities. In other words, metagenomics is a science of microbial community. It is a way of looking simultaneously at all the genomes in a microbial community as a whole and study how these might interact and influence each other's activities in serving collective functions (New, T. H. E., 2007). Hence, Metagenomics can be defined as *"the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species."* (Chen, K., & Pachter, L., 2005).

The field of Metagenomics incorporates molecular biology and genetics to study the microbes directly in their natural environments (Havre, S. L., Webb-Robertson, B.-J., Shah, A., Posse, C., Gopalan, B., & Brockman, F. J., 2005). The knowledge from the study of microbial communities plays a key role in sustaining global life support systems. Exploration, evaluation and exploitation of microbial diversity are essential for scientific, industrial and social development (Kapur & Jain, 2003).

Using Metagenomics to explore the vast microbial communities residing in our body would provide an understanding of both beneficial and harmful microbes and thus helpful in diagnosing and treating new and emerging disease problems. Recent studies on human microbiota have revealed that many non-infectious diseases, such as obesity, are

associated with change in the microbial populations in the human gut (New, T. H. E., 2007; Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., et al., 2009). Metagenomic studies have also proclaimed that the metabolites released by some of the microbes in our body interfere with the gene expression of genes associated with autoimmune disease (Proal et al., 2009).

Currently, the study of microbial community by metagenomic analysis involves mainly four steps. (i) Direct isolation of genomic fragments from the natural environments by methods such as random PCR to generate many fragments. These fragments are then cloned, and gene libraries are constructed. (ii) Sequencing using the recent sequencing technologies like pyrosequencing, reversible termination reactions. (iii) Binning related fragments to their corresponding species; and (iv) Further analysis of these fragments.

Recent technological advances have allowed the processing of steps (i) and (ii) to leap forward rapidly and cheaply. A bacterial genome can now be sequenced in days with less than $3,000. A metagenomic sample containing hundreds, if not thousands of millions of gene fragments can now be obtained from an environmental sample inexpensively. However, grouping and annotating these fragments (steps (iii) and (iv)) are still labor intensive, inefficient, and costly.

Binning can be defined as *"the process of association between sequence data and contributing species (or higher level taxonomic groups)."* (Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P., 2008). In other words, binning is the classification of the contigs (generated from the first step in metagenomic process) into major groups of bacteria, then to subgroups and then to species sometimes. The analysis

of metagenomic sequences without taxonomic assignment will always provide a superficial and an incomplete view (Pignatelli, M., Aparicio, G., Blanquer, I., Hernández, V., Moya, A., & Tamames, J., 2008).

Binning plays a key role in a metagenomic process because of two main reasons. Firstly, sequencing of genetic material from the natural environments results in large highly fragmented datasets and hence, it would be wise to first classify them before any further analysis is made. Secondly, as bacterial communities are of high complexity with thousands of species present, binning makes analysis more smooth and straightforward (Kislyuk, A., Bhatnagar, S., Dushoff, J., & Weitz, J. S., 2009). The other reasons why binning would be essential are (i) for better understanding of the community dynamics like species-species interaction. (ii) to predict the outcome and the impact on the environment (such as obesity) (iii) to provide evidences in key issues in the evolution and changes of community, such as the extent of horizontal gene transfer or barriers shaping the species (Pignatelli et al., 2008).

The current methods to investigate the microbial community diversity can be broadly divided into three categories: the number of species in a community could be estimated based on the number of unique 16SrRNA sequences in that community, whereas the ratios of individual species within the community could be estimated base on the ratios of similar sequences, and the ratios of specific biomarkers or motifs found in the samples. The common drawback for all these methods is failure to account for phenomena like gene recombination that result in gene redundancy, and horizontal gene transfer. Additionally, each method has its own backlogs. In case of biomarker-based approaches, the major drawback is that the resolution of the biomarker genes is either too

high or too insensitive in reflecting the diversity. For example, a bacterium often has multiple 16SrRNA genes. The resolution of 16SrRNA sequence between phylogenetically similar organisms is poor. Besides, this monophasic approach could not truly represent the total genome diversity of the population.

Binning based on a certain short sequence of oligonucleotide markers is a powerful technique to bin phylogenetically related species. However, when the community becomes more complex, the commonly shared markers usually diminish. Also oligonucleotide based searches assume that the oligonucleotide distribution is uniform across the bacterial genome. This assumption is not satisfied biologically since gene-coding, RNA-coding and non-coding regions, leading and lagging strands of replication and genomic islands resulting from horizontal gene transfer can all exhibit distinct oligonucleotide distributions (Kislyuk et al., 2008).

In case of sequence similarity based approaches, multiple sequence alignment requires pervious knowledge of the samples for adjusting the weight factors. This often induces bias when originals in a sample, such as those in a metagenomic data, are unclear. Additionally, as the size of a set of metagenomic data is usually very large, large computational time is needed. In view of these problems, a powerful method that could overcome these problems is required.

My research concerns a novel approach using CSSR – neural network model in binning the metagenomic data. The following objectives briefly describe the materials, methods and the flow of my research.

2. STATEMENT OF OBJECTIVES

The objective of my research is to assess the possibility to codify individual cistrons based on their unique Cistronic Stop Signal Ratio (CSSR) values and use a conventional statistical method, such as neural network, to bin individual cistrons by their phylogenetic origins. Specifically,

1. A Perl script would be written to convert a cistron sequence in FASTA format into its corresponding CSSR values.
2. A set of five known bacterial species would be selected and the CSSR values of each of their cistrons will be calculated.
3. A neural network would be trained to recognize and distinguish the genic CSSR values of the above species.
4. The optimal conditions of the neural network in binning genic CSSR values would be selected and the percentage of confidence would be defined.
5. A suitable metagenomic data set will be selected. Criteria for suitability would be based on the quality of the data, the number of contigs, the number of confirmed ORFs, the published references.
6. Use the optimal model established from objective (4) for the binning the metagenomic sequence data.
7. The resulting binning of the metagenomic data would be compared with the published results.

3. MATERIALS AND METHODS

Bacterial Genome Downloaded

Most of the required bacterial genomes were downloaded from JCVI-CMR website (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi). A few others were downloaded from EMBL and BGI websites. The downloaded genomic files are in FASTA format (Table 1). Each FASTA file is a text-based format that begins with a single line description of each gene, followed by several lines of the nucleotide sequence. The description for every gene begins with a '>' symbol followed by the gene id, gene name and organism name. The organism name is enclosed within flower braces, '{''}'. The following table presents an example of a gene in FASTA format.

Table 1.

Example of a gene sequence in FASTA format downloaded from JCVI-CMR website.

>BF0001 putative SpoU rRNA methylase family protein {Bacteroides fragilis NCTC9343}

ATGCGAAAATTGAAAATAACCGAGCTGAACCGGATAAGTATAGAAGAGTTTA
AAGAAGCTGATAAATTGCCTTTAGTTGTAGTGTTGGACGATATACGGAGTTTG
CATAATATCGGTTCTGTGTTTCGTACGGCAGATGCTTTCCGGATTGAATGTAT
TTATCTGTGTGGAATTACGGCTACTCCTCCCCATCCCGAGATGCATAAGACAG
CTTTGGGAGCCGAGTTTACAGTGGATTGGAAGTATGTTAATAACGCAGTTGA
AACGGTTGATAACCTCCGGAGTGAAGGATATGTGGTATACTCTGTCGAACAG
GCGGAAGGGAGTATCATGTTGGATGAGTTAACACTGGACCGTTCGAAGAAAT
ATGCTGTAGTTATGGGAAATGAAGTAAAAGGAGTGCAGCAGGAGGTTATTGA
CCATTCGGATGGTTGTATTGAAATTCCCCAATATGGCACAAAACATTCATTGA
ATGTATCGGTAACAGCAGGAATTGTGATCTGGGATTTATTTAAAAAGTTGAA
ATAG

Cistronic Stop Signal Ratio (CSSR)

Cistrons are the protein-coding genes. In a genome, each cistron is represented by a series of codons, which in turn is a set of three nucleotides able to code for a specific amino acid. An open reading frame (ORF) is the protein-coding sequence of DNA that starts with a start codon and ends with a stop codon. According to the universal genetic code, the start codon is ATG, although alternate codons such as GTG, CTG, and TTG are occasionally used by bacteria as start codons. The stop codons are TAA, TAG, and TGA. There are three different reading frames in any string of DNA. The first reading frame is the real sequence that starts with a start codon and ends with a stop codon. The second and third reading frames are the series of triplets of nucleotides that start from the second and third nucleotide of a gene, respectively.

The stop codons occurring in the first reading frame are the Real Stop Codons (RSC) and they terminate protein synthesis. When the stop codons are found in the second and third reading frames of a protein-coding gene, they are considered as the Premature Stop Codon (PSC). These PSC do not terminate the protein synthesis instead, act as stop signals. They truncate protein synthesis only in case of frame shift mutations.

We classified the stop signals into nine groups based on their nucleotide sequences and by their locations on a cistron. The "Cistronic Stop Signals" (CSS) of a gene is defined as a series of nine scalars in which each scalar is the frequency of a particular stop signal in a particular reading frame of a cistron. For example, consider the following hypothetical cistron composed of 11 codons:

ATG, GTA, AGG, ATA, ATT, GAG, GTA, GCC, GGT, GAT, TAA

The CSS of this gene is represented by the following CSS series:

1, 0, 0, 2, 1, 0, 0, 0, 2.

The first, second, and third scalars of the above series represent the number of TAA (=1), TAG (=0) and TGA (=0) stop signals found on the first reading frame of this gene. The fourth, fifth and sixth scalars represent the number of TAA (=2), TAG (=1) and TGA (=0) stop signals (single-underline) found on the second reading frame of this gene. The seventh, eighth, and nineth scalars are the number of TAA (=0), TAG (=0), and TGA (=2) stop signals (double-underline) found on the third reading frame of this gene. Obviously, the value of each scalar of the CSS is directly influenced by the length of the cistron. In order to compare the stop signal profiles of genes of various lengths, the genic CSS value was normalized to generate a partition series, termed "Cistronic Stop Signals Ratio" (CSSR).

CSSR Calculator in Perl

The code for CSSR Calculator was initially developed in C language and was rewritten in Perl, an interpreted language. Perl was used as it has extremely powerful string handling features and sophisticated regular expressions that make handling and scanning the large amounts of sequence contained in gene files very easy. Perl programs used for CSSR calculation were developed on Perl Express 2.5, an integrated development environment containing multiple tools for writing running and debugging Perl programs. Perl Express is free and open source software when installed. The CSSR Calculator program requires only one input, the source folder directory. This contains the

bacterial genomic files in FASTA format. The program automatically creates a destination folder directory named ‘Output’.

The program picks up every gene from every file and scans for ‘>’ symbol. It then scans for and picks the content between‘{‘‘}’, which is the organism name and assigns this name to the output file. This is followed by the CSSR calculation. However sometimes due to errors in sequencing, certain ambiguous letters (other than ‘A’, ‘G’, ‘T’ and ‘C’) occur in the gene sequence. These ambiguous letters will be randomly replaced by the program by using criteria suggested by NCBI (Table2) (http://www.ncbi.nlm.nih.gov/blast/fasta.shtml).

Table 2.

Table to replace ambiguous letters that occur in the gene sequence.

Original	Replacement
R	G or A
Y	T or C
K	G or T
M	A or C
S	G or C
W	A or T
B	G or T or C
D	G or A or T
H	A or C or T
V	G or C or A
N	A or G or C or T

Once the gene is cleared off the ambiguous letters, the occurrence of each start codon is calculated in each reading frame. Finally, the ratio is calculated in all three reading frames and the output data is stored in four files, namely: 'X.txt - CSSRGene'; 'X.txt - CSSRNonGene'; 'X.txt - Gene'; and 'X.txt - Nongene', where 'X' represent the name of the species. The Gene files are genes with only one reading frames whereas the Non-Gene files contain genes with multiple reading frames. Non-gene data is not included in subsequent calculation. The CSSR files contain the converted CSSR values of each gene in the genome whereas the latter two files retain the DNA sequences of the genome in FASTA format. All the output files are stored in the destination folder.

Feed Forward Neural Network (FFNN) Classifier

Neural networks are computational models that mimic the functional aspects of a biological brain. A neural network in which the information always moves one direction, from input layer to the output layer through the hidden layers, is called a feed-forward neural network. Input layer is the layer where the inputs are given. The calculations are made in the hidden layer and the output layer represents the output from the neural network. Figure 1 is the diagrammatic representation of a simple feed forward neural network with one hidden layer.

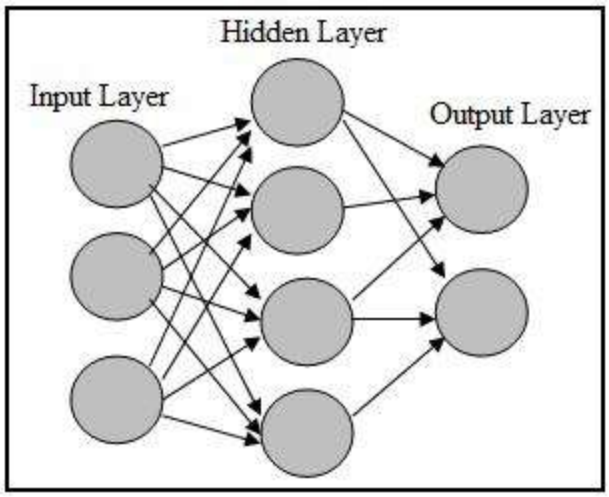

Figure 1. A simple feed forward neural network with one hidden layer.

One of the most important applications of feed forward neural networks is classification. Hence, this tool in conjunction with CSSR can be used for successfully binning gene fragments. The code for feed forward neural network model was implemented in Matlab. Matlab is a high-performance language that integrates computation, visualization and programming in an easy-to-use environment. It has several user-defined functions that can be used to create complex programs and can be easily applied to large datasets. The Matlab code was written on MATLAB 7.0. A feed forward neural network classifier was created with a specified number of hidden layers.

The type of learning used in the program was of 'supervised' type. In this type, the network is trained with a labeled set of training data which is then applied to the query data. For a set of training data made up of N input/output examples:

$T = \{(x_i, d_i)\}_{i=1}^{N}$

where x_i = input vector of the i^{th} example

d_i = desired (target) response of the i^{th} example

N = sample size

For a given set of training sample T, the parameters (like synaptic weights) of neural networks can be modified in such a way that the actual output y_i of a neural network for a given x_i is close enough to d_i (Haykin, S., 1999). The mean square error (E(n)) can be adjusted so as to achieve the performance goal (Haykin, 1999).

$$E(n) = \frac{1}{N} \sum_{i=1}^{N} (di - yi)^2$$

This classifier model requires two inputs – a file containing training data and another containing testing data. The training data comprises of the CSSR values of genes and the genomic sequences are labeled according to the species they originate from. The training set is divided into two groups – positive group and negative group. Half of the training set consists of positive group and the other half comprising of negative group. Each group has CSSR values of genes arising from different set of organisms and the genes are labeled accordingly. The testing data comprises of CSSR values from the metagenomic data set. The program then trains the model using the training set and tests

it for the testing set. The output is a text file reporting the number and percentage of positive as well as negative predictions.

To optimize this model, three different architectures were used, Architectures I (Figure 1), II (Figure 2) and III (Figure 3) with one, two, and three hidden layers respectively.

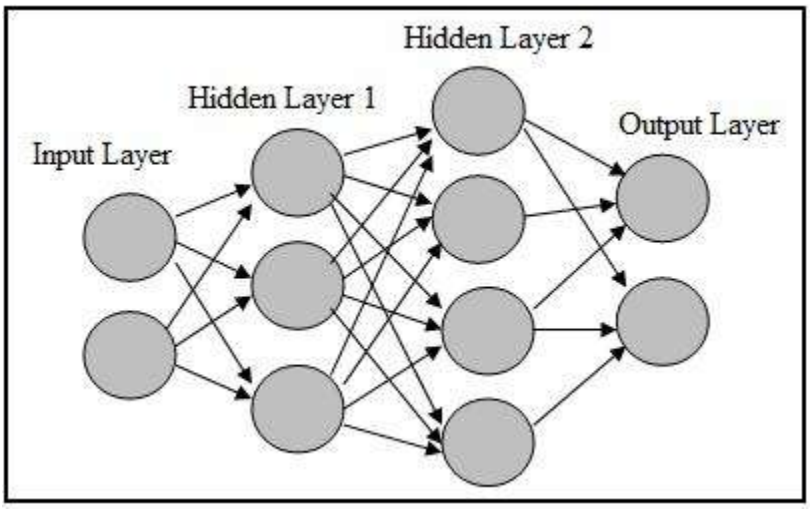

Figure 2. Architecture II with two hidden layers.

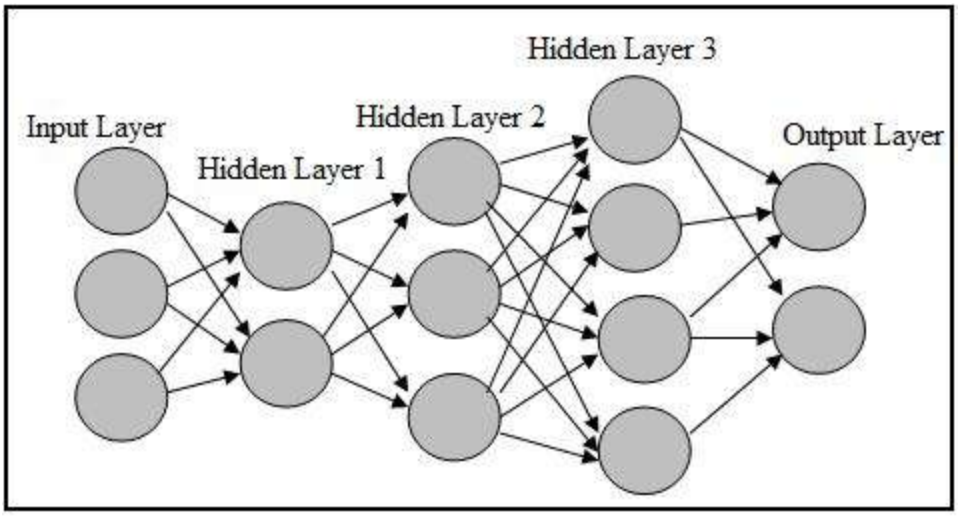

Figure 3. Architecture III with three hidden layers.

Random Gene Selector

To select the genes randomly from a genomic file, a code was written in Perl. The program requires two inputs, source folder directory and 'Gene Number'-a text file indicating the number of genes to be picked up. Each genome has a unique number of genes to be picked up and hence, a sequence of numbers indicating the number of genes to be picked is given as a text file after the folder name. The source folder directory contains all the genomic text files from the genes are to be picked up. These files are named after the organism name. These text files are in the FASTA format. The output of the program is in a destination folder directory.

The program picks up each file from the source folder directory and calculates the total number of genes in each one of these files. The program scans for '>' symbol through the whole file and increments the count every time it find this symbol. This is because the genes in FASTA format start with a '>' symbol. Once, the total number of

genes is calculated in each of these files, it picks up the respective number from the other input text file – 'Gene Number'. Then it randomly selects genes based on this number. The output is put into the destination folder directory in the form of text file (also named after the organism name).

Cluster Tool

Cluster analysis of the desired genomes was done using 'Cluster 2.11' software tool from Eisen Lab (http://rana.lbl.gov/EisenSoftware.htm). This software can be downloaded and used free of charge. It performs a variety of types of cluster analysis including hierarchical clustering, self-organizing maps (SOMs), k-means clustering and principal component analysis.

Proof of Concept

'Cross validation' was performed with known example sequences to find if the CSSR in combination with FFNN classifier model is able to distinguish between sequences of known organisms. 'Cross validation' is a technique in which a known example set of training data is fed to network in two subsets (Haykin, 1999). The first subset is used to train the network and the second subset, the validation subset is used to evaluate the performance of the model. Once the above step is completed, the network is trained with the actual training set and tested for the test data not seen before. This technique helps finding the optimal number of hidden layers in the network and the best number of epochs to be performed for efficient binning.

For this, the sequences to be binned were taken from five known organisms. The organisms under study were *Escherichia coli K12-MG1655 (EC), Salmonella enterica Paratyphi ATCC9150 (SE), Shigella boydii Sb227 (SB), Pseudomonas aeruginosa PAO1*

(PA), Rickettsia typhi Wilmington (RT). EC, SE and SB are the enteric bacteria (closely related species) and therefore their CSSR values would be similar. PA is a soil bacterium and RT is an intracellular parasite. These are distantly related to the previous ones and their CSSR values would be quite different.

The method involves the following five main steps: (i) Random selection of genes from these organisms using partition function and calculation of the CSSR values. (ii) Division of the data into training and testing sets. (iii) Training the neural network with the training set. (iv)Testing it for the testing set (validation subset). (v) Analysis of the results.

The method is described in detail as follows. The FFNN classifier requires training and testing set. As mentioned above, training and testing sets comprise genes from known organisms and are prepared separately. The genomes of organisms under consideration are downloaded from JCVI-CMR website. From these genomic files, genes are randomly selected using partition function.

The number of genes for training and testing sets is determined separately as follows:

Number of genes to be selected for training set = [(number of genes in a particular organism) ÷ (total number of genes in all the all organisms in training set)] × (Number of genes considered for training set)]

Number of genes to be selected for testing set = [(number of genes in a particular organism) ÷ (total number of genes in all the all organisms in testing set)] × (Number of genes considered for testing set)

In this way, the number of genes to be selected for training and testing sets is calculated. Genes are then selected randomly using Random Gene Selector program, based on this number. Two experiments were designed to test if CSSR in combination with feed forward neural networks can successfully bin genes from known organisms.

Selection of a Metagenomic Dataset

The microbiota of the human intestine is composed of 100 trillion viable bacteria, representing 100 or more different species (Mitsuoka, T., 1992). They have a profound influence on human health and disease as mentioned earlier. The changes in the microbial populations in our gut may lead to bowel diseases, obesity and others unexpected effects (Proal et al., 2009; Turnbaugh et al., 2009). 16S ribosomal RNA gene (rRNA) sequence-based methods revealed that gut bacteria of mostly related to two bacterial divisions, the *Bacteroidetes* and the *Firmicutes*. Together, they constitute over 90% of the known phylogenetic categories and dominate the distal gut microbiota (Eckburg, P. B., 2006; Qin et al., 2010). My research concerns binning a selected metagenomic dataset using CSSR Calculator and FFNN Classifier Model. Several metagenomic datasets on gut bacteria are now available in the Genome Projects website. Out of many research papers published on gut microbiome, the following paper was chosen.

A human gut microbial gene catalogue established by metagenomic sequencing
(JJ Qin et al. Nature 464, 59-65 (2010) doi:10.1038/nature08821)

In this paper, a metagenomic dataset of 576.7 Gb was generated from the fecal samples of 124 European individuals. An average of 4.5 Gb of sequence was generated for each sample. The Illumina read assembly was performed for each sample independently, then all the unassembled reads were pooled for another round of assembly. ORF's were predicted in each of the contigs set, and were merged by removing redundancy. The non-redundant gene set was used in the further analysis. Essentially all (99.1%) of the genes are of bacterial origin, the remainder being mostly archaeal, with only 0.1% of eukaryotic and viral origins (Qin et al., 2010).

At the depth of sequencing, they found that around 40% of the gene pool from each individual is shared. Each individual harbors at least 160 bacterial species. Out of which 57 species were common to >90% of individuals. Among them, major portion of the bacterial genes belonged to members of *Bacteroidetes* and *Dorea/Eubacterium/Ruminococcus* groups and also *Bifidobacteria, Proteobacteria* and *streptococci/lactobacilli* groups (Qin et al., 2010).

The contigs and gene set generated in this paper are available to download from the EMBL (http:///www.bork.embl.de/~arumugam/Qin_et_al_2010/) and BGI (http://gutmeta.genomics.org.cn) websites.

This was chosen as reference paper because of the following reasons: (i) the sample (from 124 individuals) is sufficiently diverse to draw a reasonable conclusion. (ii) the ORFs generated are huge and long enough for the analysis to be accurate. Around 3.3 million ORF's were generated with an average length of 704 bp. This length is sufficient for getting a good CSSR profile. (iii) A list of major groups of bacteria were included that can be used to verify the binning result from CSSR-FFNN Classifier model.

Having selected a metagenomic dataset, the next step would be to bin this dataset using CSSR-FFNN Classifier model. This requires the preparation of both testing and training sets.

Training Set

Various studies on gut microbiota as well as the study currently chose indicate that a major portion of gut microbiota belongs to classes, Bacteroidia and Clostridia. Within Bacteroidia, most of them belonged to genus *Bacteroides*. Likewise, within Clostridia, most of them belonged to genera – *Clostridium, Eubacterium, Ruminococcus* and *Dorea* (Eckburg, 2006; Qin et al., 2010; Vaughan, E. E., Schut, F., Heilig, H., Zoetendal, E. G., De Vos, W. M., & Akkermans, A. D. L., 2005). Hence, the training set has been chosen to have CSSR values calculated from four genera namely, *Bacteroides, Eubacterium, Ruminococcus* and *Dorea*. The genus *Clostridium* was not taken into consideration since it was found the CSSR values of genes of this particular genus are highly diverse and hence, selection of genes might require further analysis.

Different species within each of these four genera were selected (Figure 4) and the complete reference genomes of these species were downloaded. Few of them were downloaded from JCVI-CMR site and the others from the EMBL and BGI sites. The detailed taxonomic information of these selected species was obtained from NCBI site (http://www.ncbi.nlm.nih.gov/guide/taxonomy/). Initially, the CSSR values of genomes for the selected species were calculated using the CSSR Calculator. These values were to cluster these organisms using the 'Cluster 2.11' tool. All the species of a particular genus form a cluster.

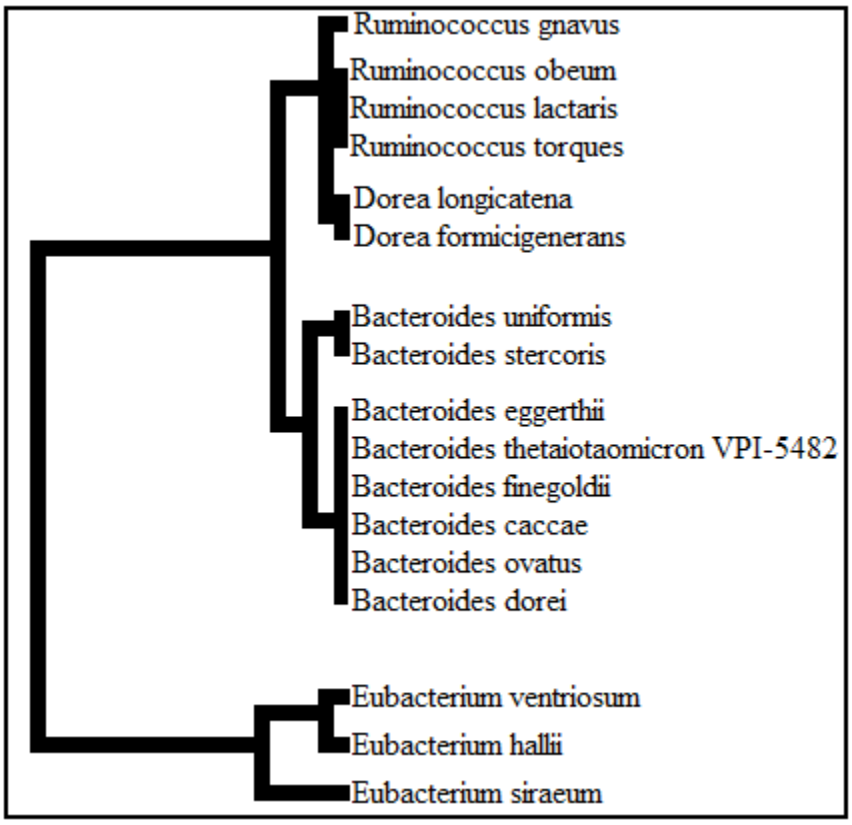

Figure 4. Cluster Analysis of species from four key intestinal genera.

The first step in preparation of the training set would be selection of genes from each of these species mentioned above. The genomic files are in FASTA format. The genes from these files were selected randomly by 'Random Gene Selector' program. The genomic files of all the species were placed in a source folder directory and the number of genes to be picked was given in the 'Gene Number.txt' file. The program then picks up genes from the respective files and places the output in an output folder directory. The number of genes picked from each genus was equal to ensure equal representation of all four genera in the training set.

The genes from all the output were mixed and randomized and then was given as input to the 'CSSR Calculator'. The output file consists of CSSR values of all the genes selected for the training set. Since the training is unsupervised, the genes were finally labeled according to the genus of their origin. This formed the training set.

Testing Set

The testing set comprised of the CSSR values of gene fragments from the metagenomic dataset. Firstly, the fragments of the metagenomic dataset were checked for ORF's. In other words, for each fragment the number of stop codons in the first reading frame was calculated. Only those fragments for which this number is equal to one were selected. Fragments for which this number is less than or greater than one were discarded. After filtering, testing data of about 1.38Gb was generated. The filtered fragments were then given as input to the 'CSSR Calculator'. The output is the CSSR values of gene fragments that formed the testing set.

Working Model

FFNN Classifier Model is based on a feed-forward type of network and hence has only two outputs as mentioned earlier. In other words, the testing set can be classified only either positive or negative. However, the training set consists of four groups labeled according to their genus of origin. Hence, the testing set (metagenomic fragments) should also be classified into four groups. With FFNN Classifier Model the work becomes more laborious and time consuming. Therefore, a search was made for an online tool that can classify the testing data into more than two groups in a much lesser time. Among the several tools available, 'Neural Tools 5.5' was selected as it suited our needs. This works on Microsoft Excel. This would make training easier for as the output from 'CSSR

Calculator', the CSSR value of a gene is a series of nine scalars. Also, this tool automatically gives the percent probability of a particular gene fragment being classified into a particular group. Moreover, the time required for classification is far less. Before applying directly to metagenomic dataset, an initial testing was done. It was tested on the data from experiment 2 (Figures 7 & 8) of the 'proof of concept' section. The result was similar to that got from 'FFNN Classifier Model' and was much better as it also indicates the percent probability or percent confidence level. With the same training and testing datasets, FFNN Classifier model as well as Neural Tools software gave almost the same result. Therefore, Neural Tools was used for binning the metagenomic data.

The first step in binning the metagenomic data would be to train the Neural Tools with the training data. For this, the CSSR values of the labeled training data will be loaded on a Microsoft Excel sheet. CSSR is a series of nine scalars as mentioned earlier and will be loaded in the first nine columns of the Excel sheet. The last column (10^{th} column) will consist of the label indicating the genus of origin of a particular gene fragment. This forms the training set and the Neural Tools model will be trained with this data.

Once the training is done, the model will be tested for the testing data prepared from metagenomic gene fragments. In a similar manner to training data, testing data is also CSSR values of gene fragments as a series of nine scalars. These nine scalars will be loaded on the Excel sheet and tested for predictions. The model then predicts and gives a percent confidence for each gene fragment binned. The results will then be obtained and analyzed.

4. RESULTS

Binning Genes from Known Bacterial Species

A: Training *Escherichia coli* vs. all others. The training set for training the FFNN Classifier consists of two groups – positive group and negative group, as mentioned earlier. In the first experiment, the CSSR value of genes from *Escherichia coli*, one of the three closely related bacteria (enteric group) were put into the positive group and the CSSR value of genes from all the others into the negative group (Figure 5).

When the FFNN Classifier was tested with the testing data, the binning efficiency of the model (with the three architectures) was about 80% (Figure 6).

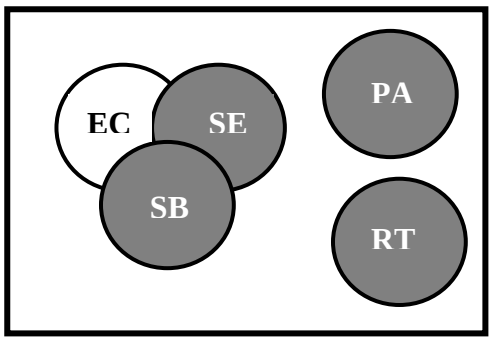

Figure 5. (Training A) Pie diagram showing five different bacteria. Uncolored as positive group and colored as negative group.

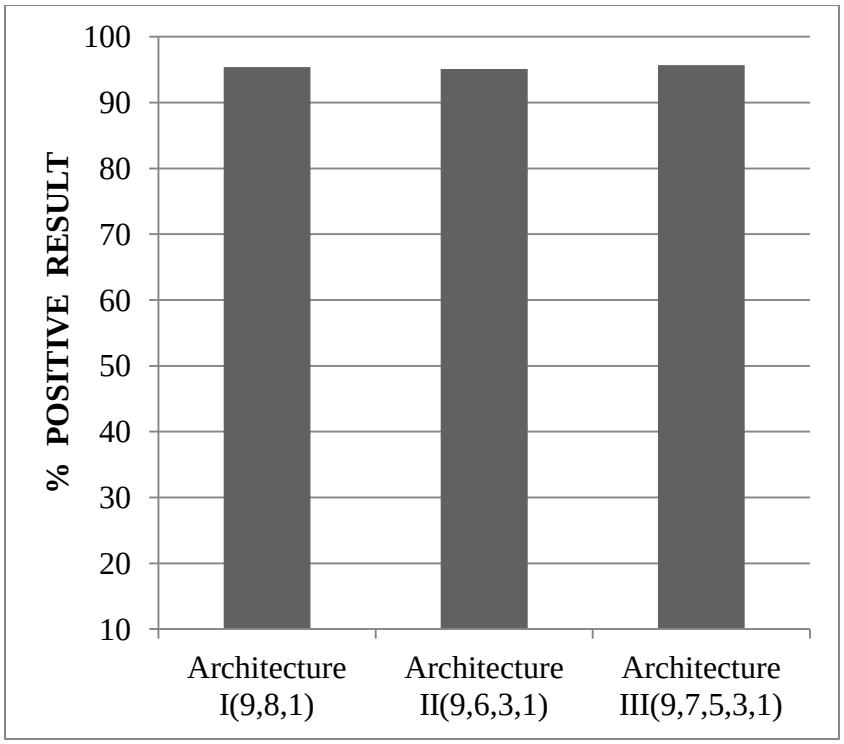

Figure 6. (Training A) Result from FFNN Classifier for three different architectures.

B : Training Enterics (*Escherichia/ Shigella/ Salmonella*) vs. Non-Enterics (*Pseudomonas/ Rickettsia*). When the CSSR value of genes from enteric group - EC, SE, SB were pooled and put into the positive group and the CSSR value of genes from the other bacteria (PA and RT) were put into the negative group (Figure 7), the binning efficiency increased to as high as 96% (Figure 8).

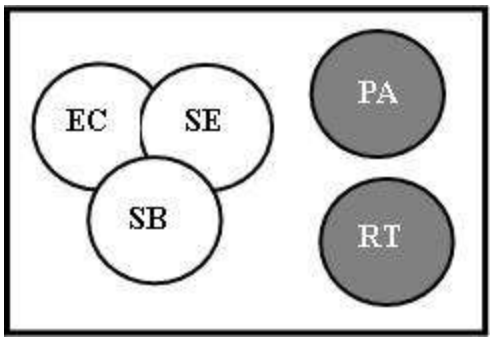

Figure 7. (Training B) Pie diagram showing five different bacteria. Uncolored as positive group and colored as negative group.

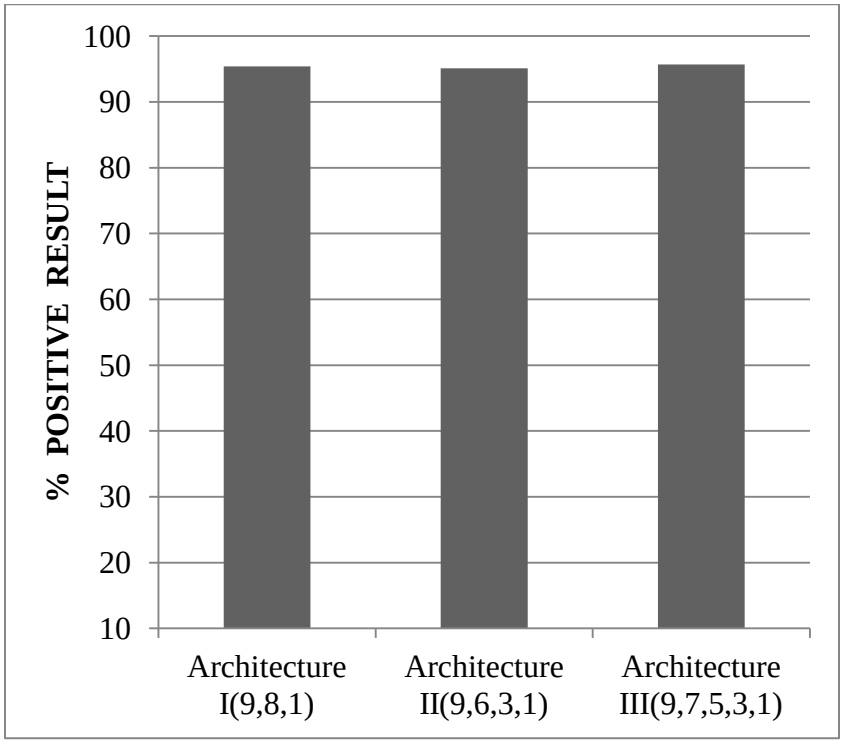

Figure 8. (Training B) Result from FFNN Classifier for three different architectures.

Initial Testing With Neural Tools Software

The result from the initial testing of the Neural Tools software for enteric group (EC, SE, SB) and the other group (PA, RT) is indicated in Figure 9 and Figure 10 respectively. When tested for around 1100 genes from enteric group, the Neural Tools software could classify all of them correctly into enteric group with a confidence level greater than 50%. Most of them were classified with a confidence level greater than 90%. From the graph, out of 1100 genes tested, 970 genes were classified with a confidence level greater than 90%. Also, more than 1050 (96% of total testing genes for enteric group) were classified with greater than 70% confidence level.

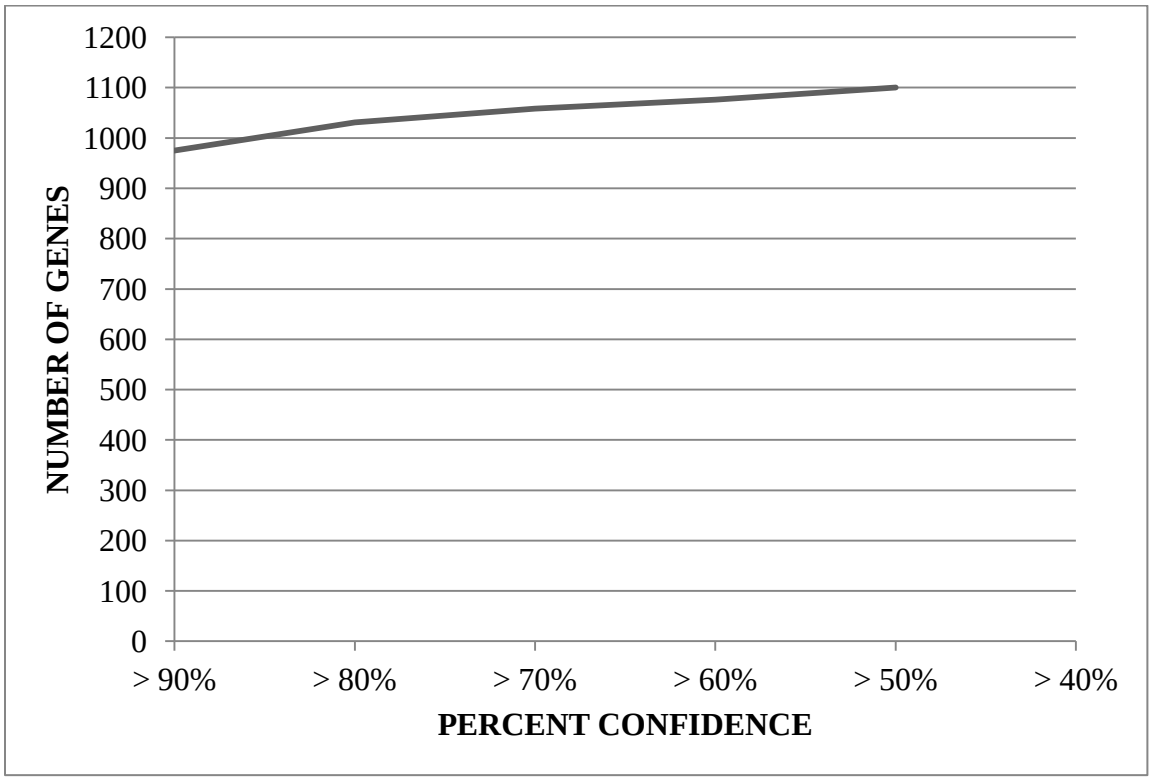

Figure 9. Number of genes classified as enteric group by Neural Tools software.

From this, it is clear that the Neural Tools software is very efficient in binning genes from known genomes. Also, high percentage of genes (with confidence level > 90%) form the core genes of the enteric group and hence, the software could classify them with such high confidence level. Very few of them (genes with confidence level >60%) might be genes acquired from other genomes or outside through phenomenon like horizontal gene transfer and are not representative of the core genes of the enteric group.

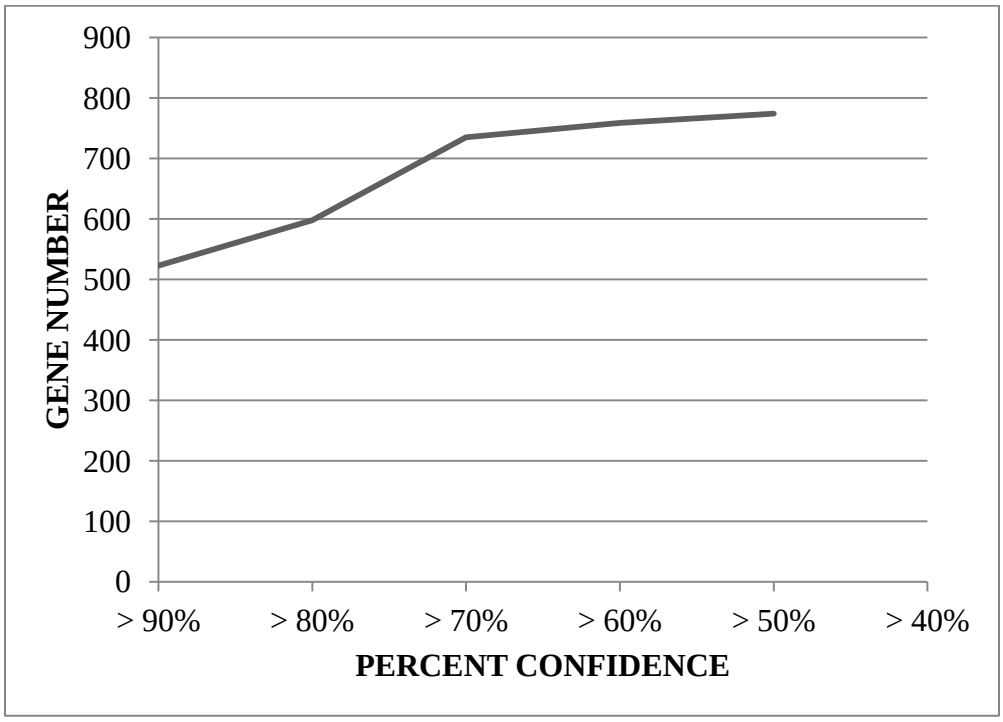

Figure 10. Number of genes classified as belonging to PA, RT groups by Neural Tools software.

Similar result was observed when tested on the dataset containing genes from PA and RT. Around 770 genes were tested using the Neural Tools and all of them were

predicted correctly with a confidence level greater than 50%. Out of 770, around 520 of them were classified into that group with a confidence level greater than 90% indicating that they form the core genes representing group PA, RT. Around 735 genes (94.9% of the total testing genes for group PA and RT) were classified with a confidence of greater than 70%. The remaining genes might be the non-representative or the genes from horizontal gene transfer. However, the number of genes with a confidence level greater than 90% has decreased when compared to the above graph (Figure 9). This is because the bacteria in the first group (enteric) belong to the family and are very closely related and hence have many core genes in common. In the second case, group containing genes PA and RT do not belong to same family and hence might have less core genes in common. Therefore, the Neural Tools software binned high percentage of genes with a confidence level of greater than 90% into the enteric group (Figure 9) when compared to those to group containing PA and RT (Figure 10). The same testing data with FFNN Model had 95% correct predictions (Figure 8) with a confidence level of 50%.

Binning Metagenomic Dataset

After training the Neural Tools with the labeled training data representing the four different bacterial genera, the model was tested with the testing data generated from the metagenomic dataset. The testing data was binned into four groups, namely *Bacteroides* group, *Dorea* group, *Ruminococcus* group and *Eubacterium* group. Percent confidence level was predicted for each gene fragment. There were more than 3,00,000 gene fragments in the testing dataset. Out of them, 27.35% were binned to *Bacteroides* group; 8.31% were binned to *Dorea* group; 29.31% were to *Ruminococcus* group; 35.01% were

binned to *Eubacterium* group (Figure 11) with confidence levels ranging from 20 – 100%.

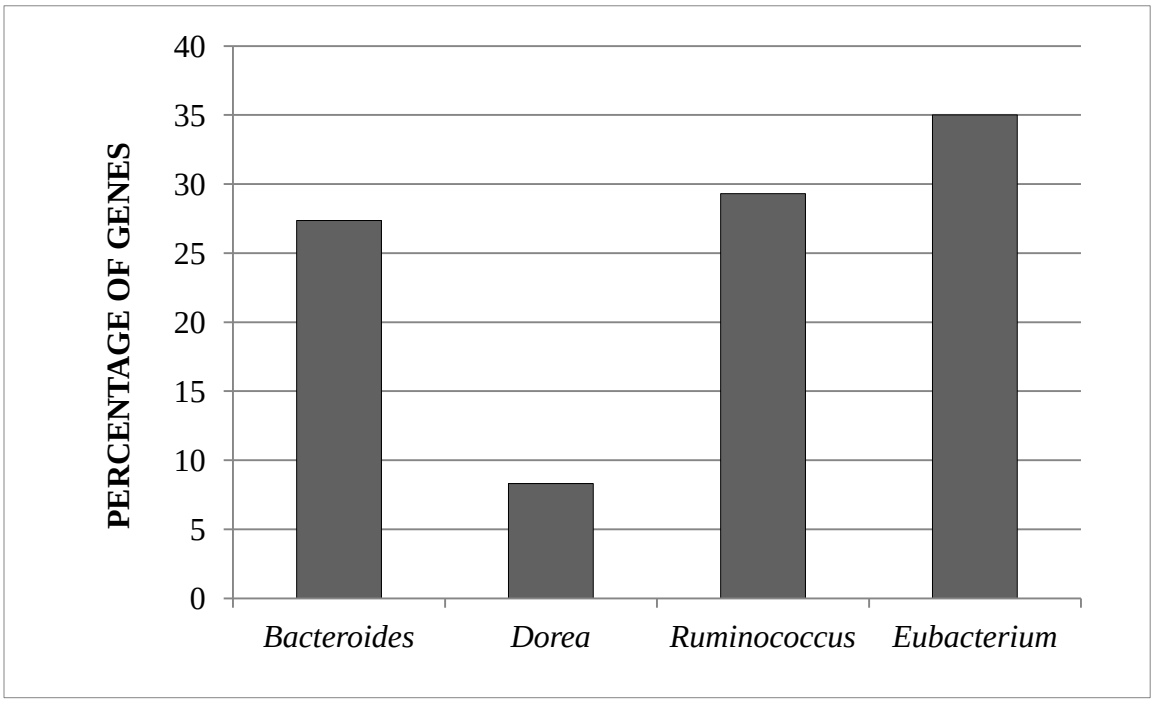

Figure 11. Percentage of genes belonging to each group.

Bacteroides **Group**

Out of total metagenomic gene fragments, 240386 gene fragments (27.35% of the total metagenomic dataset) were binned into this group with confidence levels ranging from 20-100%. The results are summarized below. The number of genes within a particular range of confidence levels is shown in Figure 12.

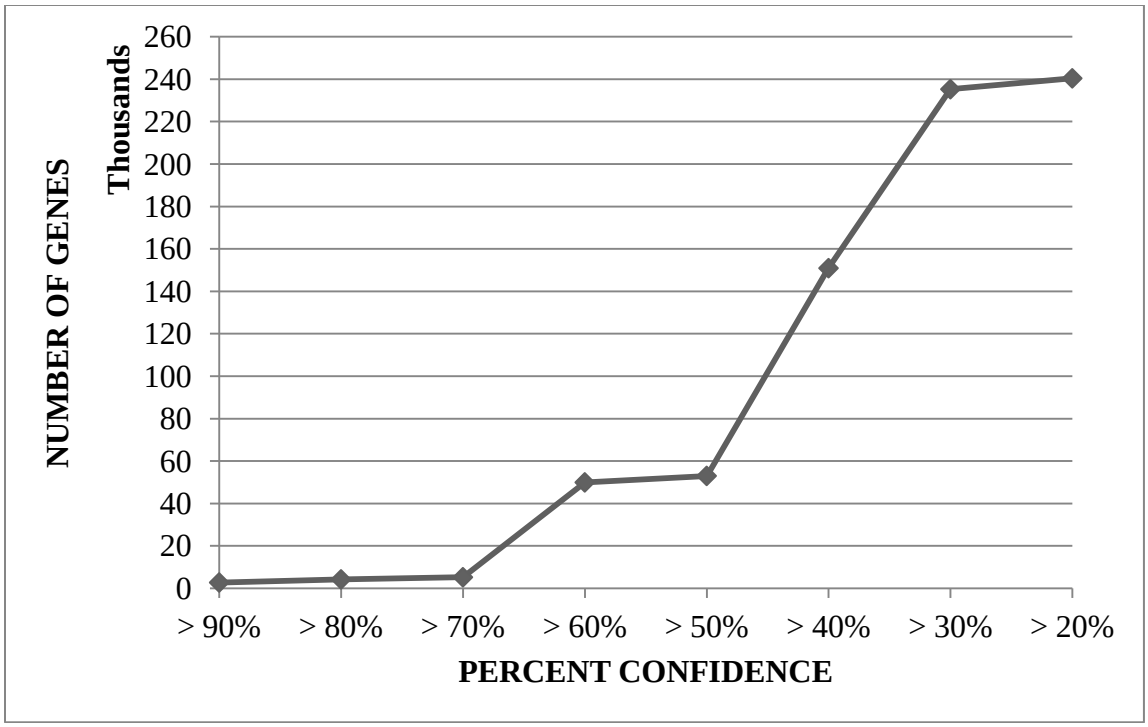

Figure 12. Number of genes binned into *Bacteroides* group with varied confidence levels.

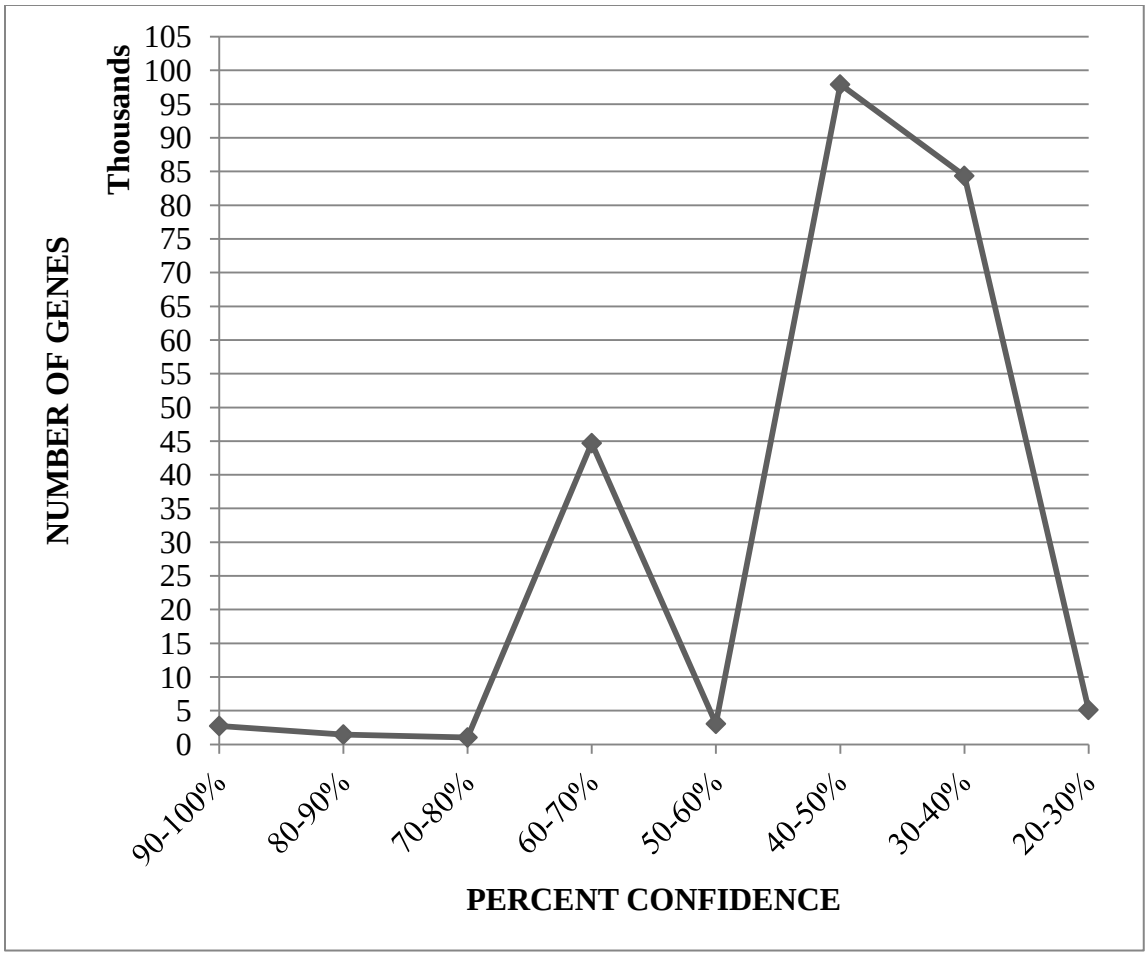

Figure 13. Number of genes binned into *Bacteroides* group within different narrow ranges of confidence levels.

Dorea Group

Out of total metagenomic gene fragments, 73057 gene fragments (8.31% of the total metagenomic dataset) were binned into this group with confidence levels ranging from 20-100%. Compared to other groups, the number of genes binned into this group was low. The results are summarized below.

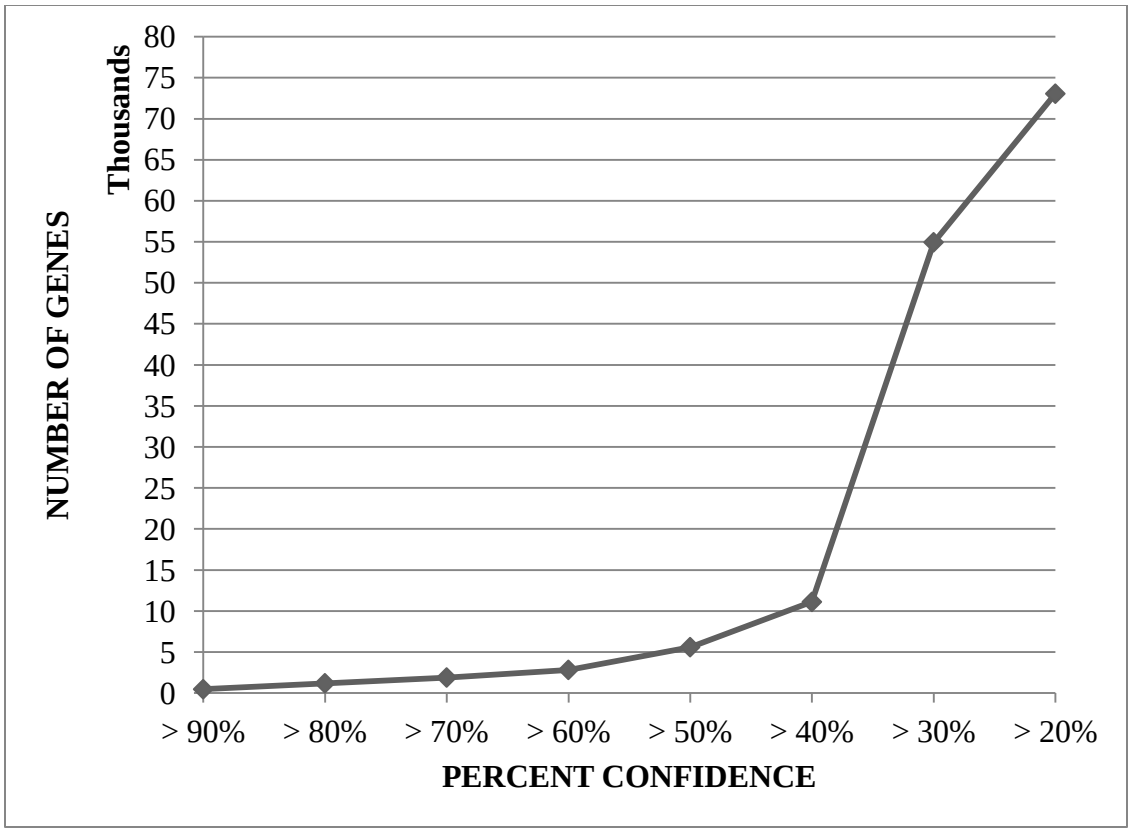

Figure 14. Number of genes binned into *Dorea* group with varied confidence levels.

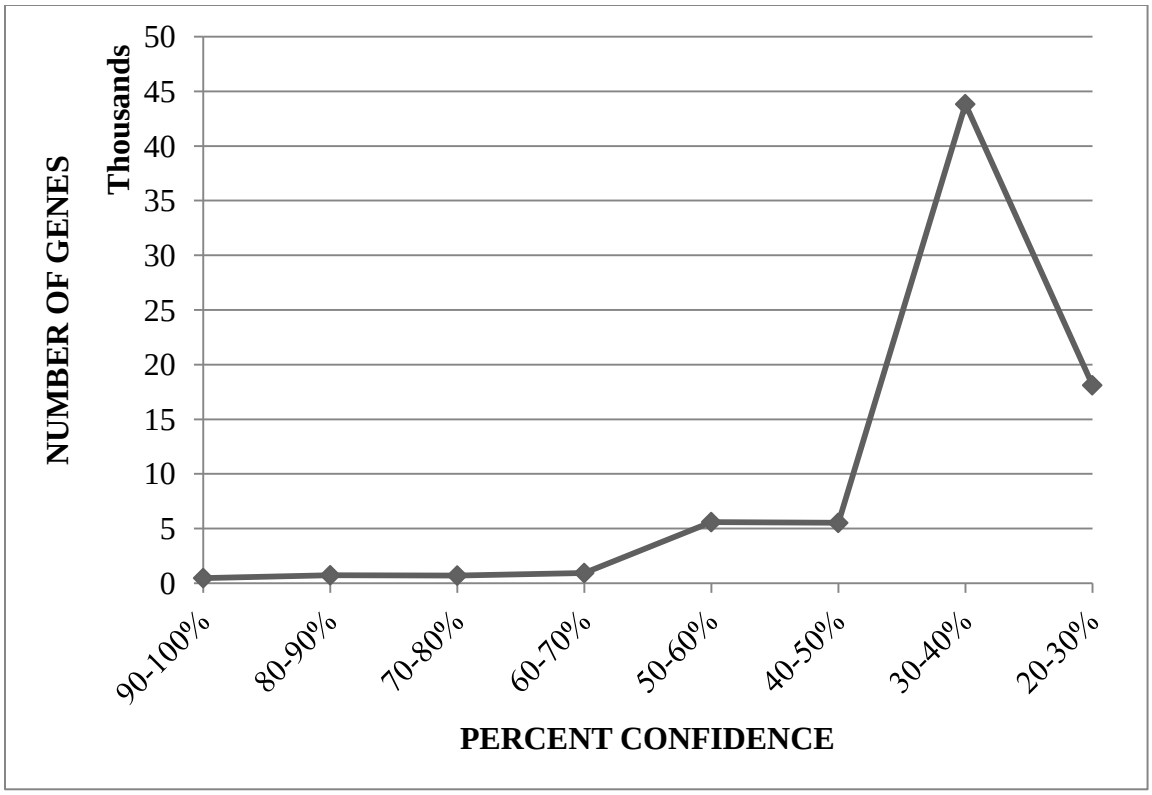

Figure 15. Number of genes binned into Dorea group within different narrow ranges of confidence levels.

Ruminococcus Group

Out of total metagenomic gene fragments, 257564 gene fragments (29.31% of the total metagenomic dataset) were binned into this group with confidence levels ranging from 20-100%. The results are summarized in Figure 16.

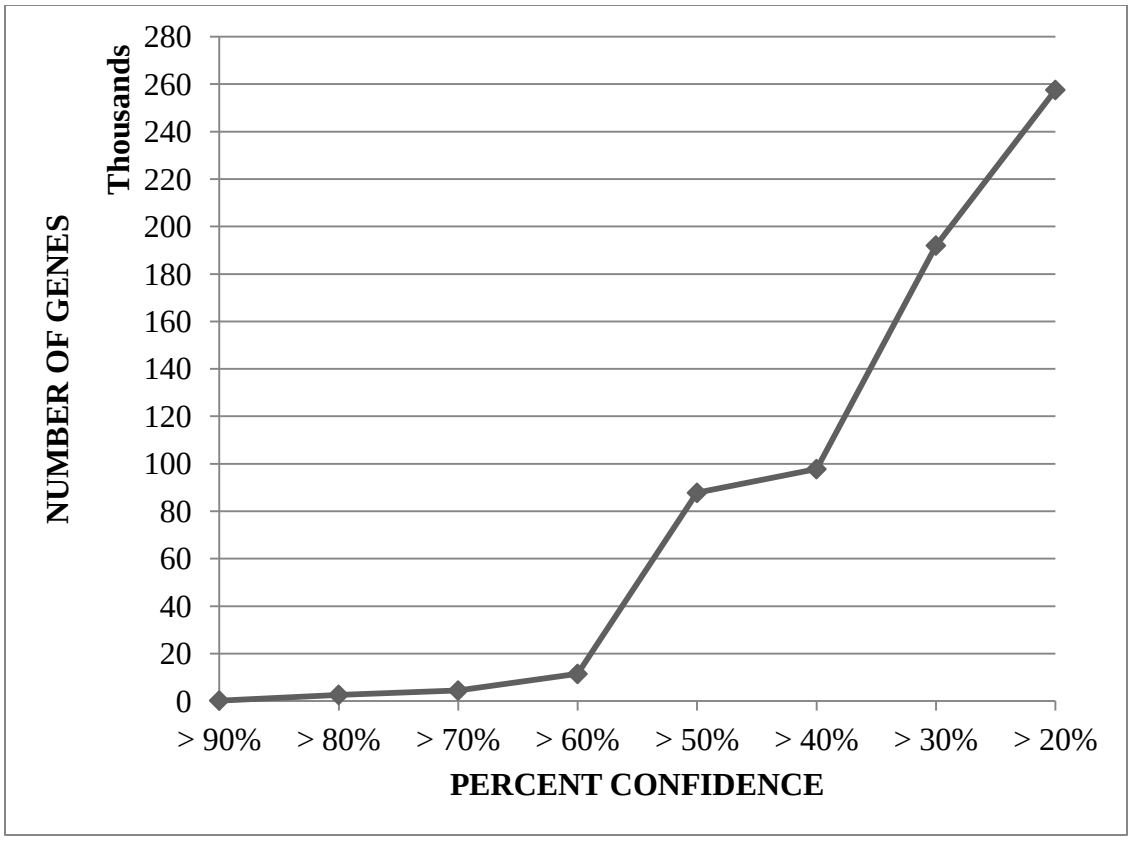

Figure 16. Number of genes binned into *Ruminococcus* group with varied confidence levels.

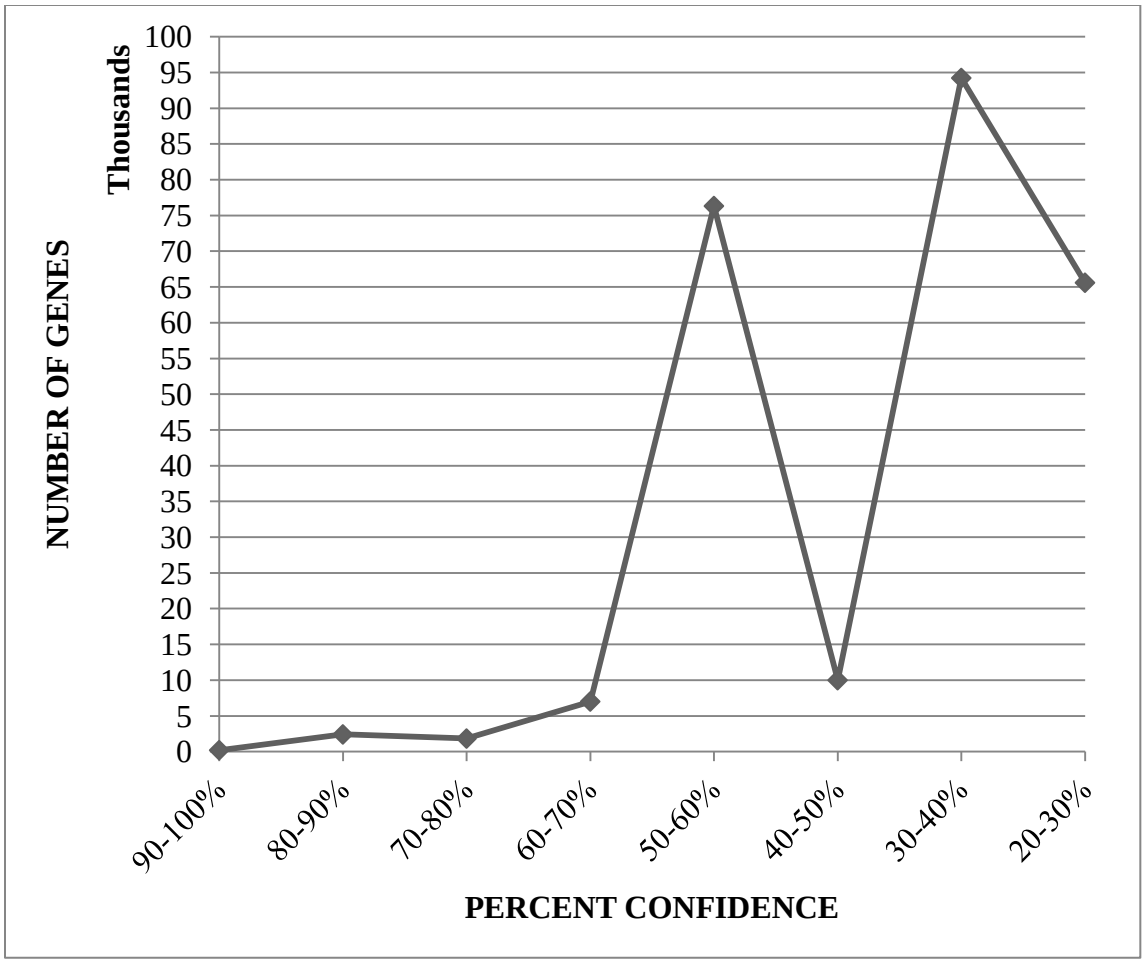

Figure 17. Number of genes binned into *Ruminococcus* group within different narrow ranges of confidence levels.

Eubacterium Group

Out of total metagenomic gene fragments, 307692 gene fragments (35.01% of the total metagenomic dataset) were binned into this group with confidence levels ranging from 20-100%. Compared to other groups, the number of genes binned into this group was the highest. The results are summarized in Figure 18.

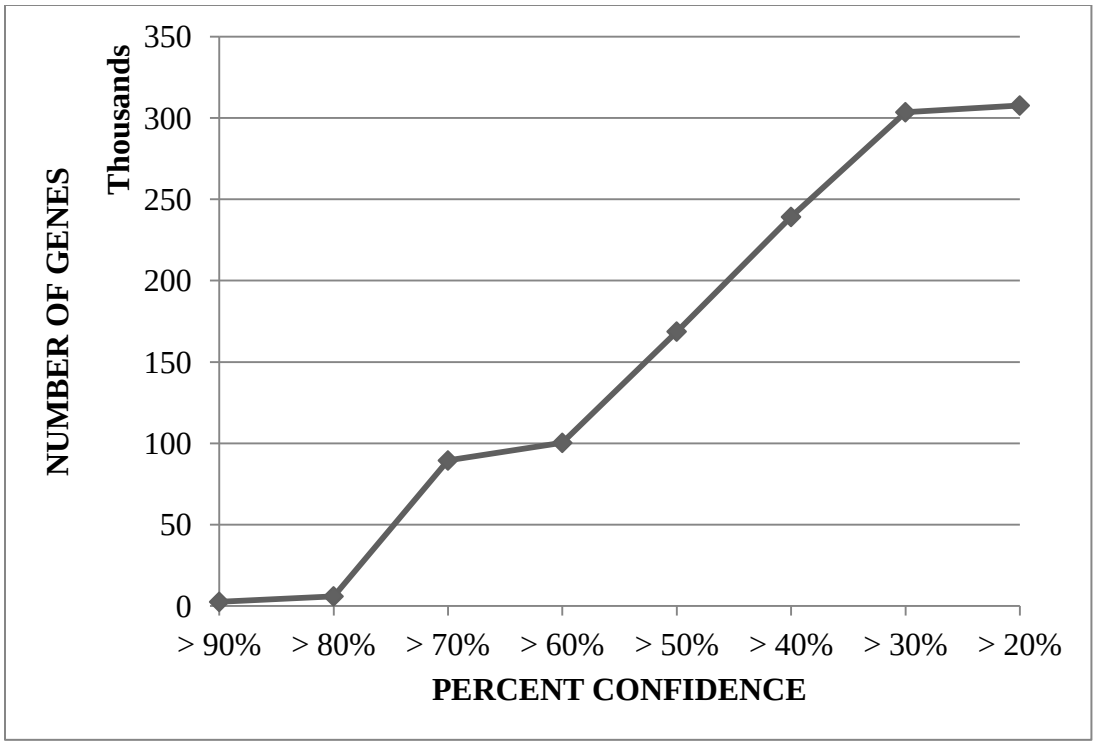

Figure 18. Number of genes binned into *Eubacterium* group with varied confidence levels.

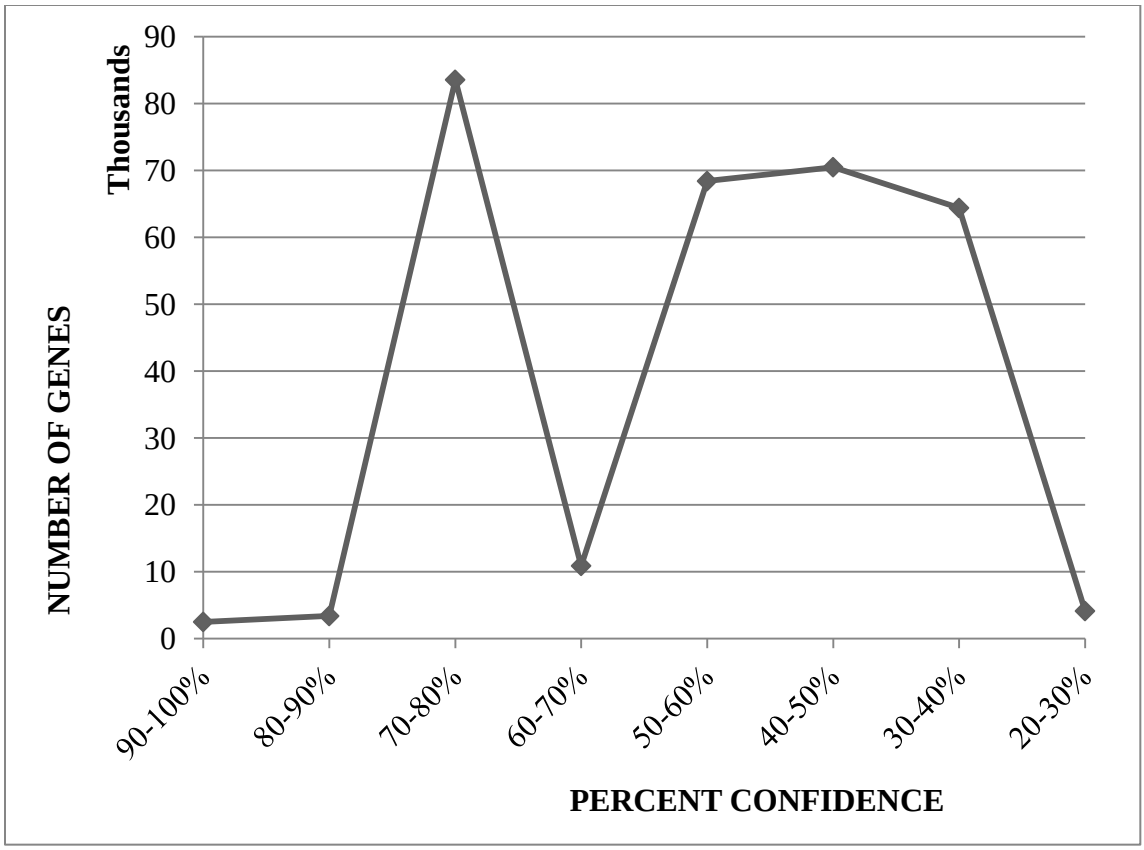

Figure 19. Number of genes binned into *Eubacterium* group within different narrow ranges of confidence levels.

5. DISCUSSION

The intestinal microbiota plays a key role in the maintenance of host health, providing energy, nutrients and protection against invading organisms (Woodmansey, E. J., 2007). Hence, novel molecular technologies have been increasingly used to study these complex communities and their interaction with their host (Vaughan et al., 2005). In the present study, attempts were made to see if CSSR can be used as a biomarker in binning metagenomic data.

The two experiments initially conducted to evaluate the binning efficiency of the model proved that the CSSR can be used as a biomarker in binning genes from known organisms. In the first experiment (Training Escherichia coli vs all others - Figures 5 & 6)**,** as the CSSR values of the closely related organisms are similar and since one of the closely related bacteria was put into one group, it was difficult for the model to differentiate between them. However, the result yielded an acceptable efficiency. In the second case (Training Enterics vs Non-enterics - Figures 7 & 8), the efficiency increased because CSSR-FFNN model could readily differentiate between the groups as the differences of their CSSR values are large. This suggested that a this model could be used to differentiate distantly related bacteria. Hence, this model was further used to work on the metagenomic data.

The initial success in using CSSR as binning tool for known genes from known organism lead us to develop a more complex scheme to example some available metagenomic data of human guts. Among the body sites colonized by the community of microbes, the human gut harbors the greatest number and highly diversified bacteria (Sears, C. L., 2005). Several studies were conducted by many researchers to explore this

complex microbiota and all of them have certain common conclusions indicating the key intestinal genera. At birth, humans become colonized with facultative aerobes including *Streptococci spp.* and *Escherichia coli* but, at the critical juncture of weaning, there is a dramatic shift in the flora with obligate anaerobes, particularly *Bacteroides* species, becoming significant (Hooper, L. V. 2004; Sears, 2005). The dominant fecal flora of healthy adults consists of mainly *Bacteroides* and *Eubacterium spp.* (Mitsuoka, 1992).

Zeotendal et al. (1998) observed that the most dominant bands in the TGGE profile comprised of sequences from undescribed bacterial species and found three species with greatest similarities to *Ruminococcus obeum*, *Eubacterium halii* and *Fusobacterium prausnitzii* were dominant in all the individuals investigated. Simmering et al. (1999) reported that each human individual tested had his or her specific *Eubacterium ramulus* strain, reinforcing that fact the *Eubacterium spp*. is dominant in healthy individuals . In another study with fecal samples from three healthy adults, Eckburg (2006) indicated that phyla - Firmicutes and Bacteroidetes dominated the gut flora. Moreover, most of them were novel and 95 % of the Firmicutes sequences were members of Clostridia class (Eckburg, 2006). A recent study of distal human intestine also revealed that three genera – *Bacteroides*, *Clostridium* and *Eubacterium*, each comprise nearly 30% of bacteria in fecus and the mucus overlying the intestinal epithelium (Backhed et al., 2005). The reference paper used for the current study indicates that the prominent gut species were the members of Bacteroidetes and *Dorea/Eubacterium/Ruminococcus* groups and also *bifidobacteria, proteobacteria* and *streptococci/lactobacilli* groups (Qin et al., 2010).

Bacteroides species utilize a wide variety of carbon sources and are known to play a key role in the digestion of majority of polysaccharides occurring in the human colon (Gibson, G. R. 1991; MacFarlane, G. T., & Salyers, A. A., 1984). *Eubacterium* and *Ruminococci* species are known to be involved in fermentative metabolism (Zoetendal et al., 1998). Another investigation on human fecal flora identified a novel uncultured bacterium which was found to be a nearest relative of *Eubacterium formicigenerans* and was designated a species of a novel genus namely, *Dorea longicatena gen. nov.* (Taras, D., Simmering, R., Collins, M. D., Lawson, P. A., & Blaut, M., 2002).

Based on the above mentioned studies it is clear that members of Firmicutes and Bacteroidetes are dominant in the intestinal microbiota. Within these phyla, the Clostridia and Bacteroides classes dominate the gut flora. Within these classes, four key genera – *Bacteroides spp.*, *Eubacterium spp.*, *Ruminococcus spp.*, *Dorea spp*. were picked for further analysis. The genera *Eubacterium, Ruminococcus* and *Dorea* are within Clostridales. Another genus, *Clostridium*, was also found to be dominant by many studies (Woodmansey, 2007). However, this was not taken into consideration in the present study as it was found that this genus is highly diversified and their genome CSSR values were also very diverse. This shouldn't be surprising as the criteria for grouping of clostridial species are often loosely defined. *Clostridium* is often considered as paraphyletic, with species appearing in multiple locations within a phylogenetic tree (Sneath, P., 1984). This indicates that many *clostridia* appearing in clades apart from the type strain are clearly misclassified. It should be noted that, a 'species' is often defined based on few morphological characteristics, pathogenicity and source of isolation.

16SrRNA sequence analysis is based on sequence similarity of one particular gene. Hence, certain genera are usually very diverse as they are based on only certain characteristics. In the current study, the CSSR of a genome is the average of CSSR values of all genes in the genome. Hence, there is no bias in selection of genes and all genes are considered in defining a species. Hence, the current method will be more effective in binning gene fragments from a metagenomic dataset. However, further investigation is required to study and understand the diversity of genera like *Clostridium*, in order to prepare the training set required for binning. The following table indicates the relative percentage of each genus binned using CSSR – FFNN Model (Table 3).

Table 3.

Relative percentage of each genus in the selected metagenomic dataset.

Percent confidence level	***Bacteroides***	***Dorea***	***Ruminococcus***	***Eubacterium***
>90	0.31	0.05	0.02	0.28
>80	0.48	0.14	0.3	0.67
>70	0.6	0.22	0.5	10.18
>60	5.6	0.32	1.3	11.41
>50	6.03	0.64	9.99	19.2
>40	17.17	1.26	11.12	27.22
>30	26.77	6.25	21.85	34.55
>20	27.36	8.31	29.31	35.01

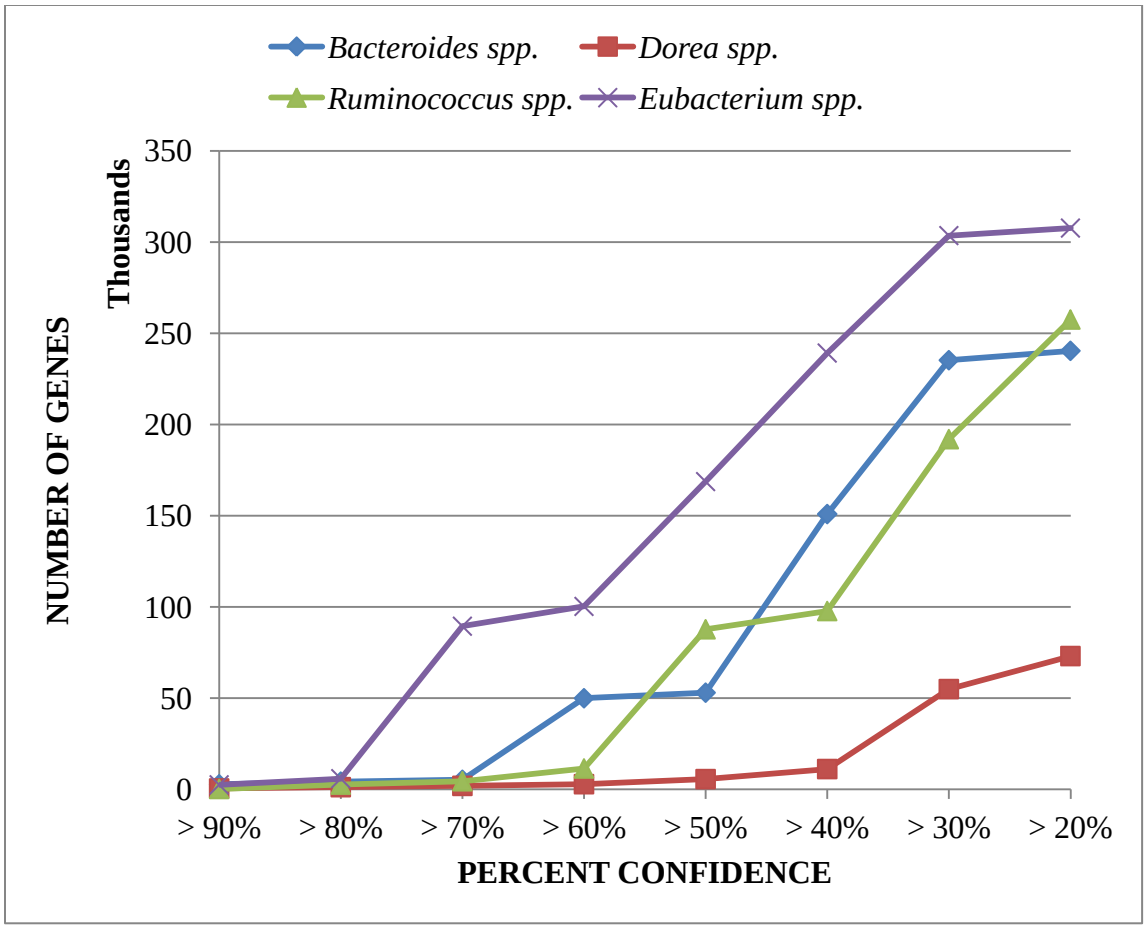

Figure 20. Relative abundance of each genus in the selected metagenomic dataset.

With reference to the confidence level, as the percent confidence decreases the number of genes assigned to a particular genus was found to increase. Several conclusions can be made from the above graph (Table 3 & Figure 20).

One of them is that the genes above a confidence level of 90% or 80% might be the genes from the same species considered in the training set. The following statements explain the reasons for the above conclusion. The confidence level >90% or >80% indicates that the CSSR values of these genes (under >90% or >80%) is very similar to those of the genes in the training set. Hence, probably these genes are from the same

species that are considered for the training set. For further illustration, the graph of the *Eubacterium* group is almost constant for a range of confidence from >90% to >80%. This indicates that these genes in the metagenomic dataset have very similar CSSR values to those of the *Eubacterium spp.* genes considered in the training set. Hence, the genes from metagenomic dataset above confidence 90% or 80% might be genes from the same species of *Eubacterium* used in the training set. The CSSR values of genes in the training set representing *Eubacterium* genus, are from three species namely - *Eubacterium halii, Eubacterium siraeum* and *Eubacterium ventriosum.* Hence, these genes above confidence level 90% or 80% might be from these three species.

Second conclusion with regard to the confidence level is that as the confidence level decreases, the genes from the metagenomic dataset under that particular confidence level might be genes from related species or genera or family. The following statements support this. There is a steep increase in the number of genes from confidence level >80% to >70% and remains almost constant till >60% incorporating lot many genes from the metagenomic dataset. Since the CSSR values of genes from species under the same genus are very similar and are seen to cluster together, (Figure 4) these genes from >80% to >60% might be from other species under the same genus not considered in the training set. The number of genes from confidence level >60% to >30% exponentially increase and becomes constant later. This indicates that these genes might be from other genera under the same family like *Anaerofustis* (Eubacteriaceae) or from genera closely related to *Eubacterium* genus like *Clostridium* and *Lactobacilli* from other families. For example, some strains of *Eubacterium aerofaciens* are very similar to strains of *Streptococcus intermedius*, which is also one of the major genera found in human gut

(Moore, W. E. C. & Holdeman, L. V., 1974). Hence, these might be genes from *Streptococcus* genus. However, further analysis is required to confirm this. Another round or rounds of training and binning with different training sets is needed to explore further.

Third conclusion is that these genes from confidence level >60% to >30% might be completely novel, uncultured and having close characteristics to *Eubacterium spp.* They might also be genes that might have been acquired by unknown bacteria from *Eubacterium spp.* by phenomenon like horizontal gene transfer and hence, have a 30% or so similarity in CSSR values of the genes or gene sequence. Hence, a major conclusion is that CSSR can also be used a biomarker to predict genes from novel bacteria having certain characteristics similar to known genomes. Similar conclusions can be made upon observing graphs for other genera namely, *Ruminococcus, Bacteroides,* and *Dorea.*

Fourth conclusion is that the shape of the graph with regard to confidence level indicates the diversity of a particular species. Unlike *Eubacterium* group, the graph for genera *Ruminococcus* and *Bacteroides* groups is almost constant till about >70% confidence level. This indicates that the CSSR values or the stop signal ratio of these genes is conserved. Hence, the species under these genera are more conserved than those under genus, *Eubacterium*. This forms the fourth major conclusion that the *Eubacterium* genus is more diverse than other genera. The graph of genus *Dorea* is more conserved as the graph is almost constant till a confidence level above 40%.

Dorea genus as mentioned above was derived from *Eubacterium* genus and is very closely related to *Eubacterium formicigenerans*. Also, the number of genes associated with *Dorea* group is very less when compared to the number of genes

associated with other genera. As the genes associated with its closest relative are already separated, less number of genes is associated with genus *Dorea*. This leads to another major conclusion that CSSR can also be used as a biomarker in binning phylogenetically closely related organisms.

On the whole, genera – *Eubacterium, Bacteroides* and *Ruminococcus* are more abundant which is in concordant with many studies mentioned above (Backhed, 2005; Mitsuoka, 1992; Qin et al., 2010; Simmering et al., 1999; Zoetendal et al., 1998). The overall numbers may not exactly match with those of the previous studies. This is because several factors, both intrinsic such as GI tract location or genetic background and extrinsic factors such as diet and health influence the overall numbers of microbes in the gut intestinal flora (Vaughan, 2005). The following example explains this.

According to the result obtained from binning, *Eubacterium spp.* was found to be the most abundant among other genera. Previous research reported *Eubacterium spp*. to be the second most abundant genus after the *Bacteroides spp.* (Woodmansey, 2007). The difference in numbers depends upon the age of the individuals from which the fecal samples are collected. It was observed previously that the *Eubacterium spp*. increase in elderly volunteers compared to their younger counterparts (Woodmansey, E. J., McMurdo, M. E. T., Macfarlane, G. T., & Macfarlane, S., 2004). The age of the volunteers from which the samples were collected was not mentioned in the selected reference paper. Hence, it can be assumed that there might be more elderly volunteers than younger ones and this accounts for the difference in the exact overall increase in *Eubacterium spp*. compared to *Bacteroides spp.*

Also in the selected reference paper, it was mentioned that the fecal samples were collected from healthy, over-weight and obese individual adults. This might also be another reason because, previous studies propose that the number of Bacteroidetes in obese people is far less when compared to lean people (Turnbaugh et al., 2009). The converse is true in case of Firmicutes. Furthermore, no microbial community on the biosphere has been sampled to completion. The biases in the current sampling method and its inability to distinguish live from dead organisms may also contribute to the difference in the overall populations of different species.

At a confidence level greater than 80%, very less percentage of total metagenomic data was binned to these four genera indicating the presence of many novel species. This further concludes that there is a very vast and highly diverse gut microbiota to be explored with thousands of bacterial species present which is in concordant with the previous studies (Mitsuoka, 1992).

With the above mentioned evidences, it can be concluded that CSSR is an efficient biomarker in binning metagenomic data with a much lesser effort and time when compared to other binning methods. However, as the number of types of microbes or bacteria increase in the dataset, the binning accuracy decreases. Further rounds of binning and analysis are required in order to bin to a specific species level and to find the exact overall numbers of each species in a particular metagenomic dataset. Moreover, there is no particular tool that can successfully solve the metagenomic binning problem. Always, a proper combination of two or more effective tools would help to come to a definitive conclusion.

REFERENCES

Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science, 307(5717),* 1915-1920.

Chen, K., & Pachter, L. (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLoS Computational Biology, 1*(2), 106-112.

Cobb, C. M. (2008). Microbes, inflammation, scaling and root planing, and the periodontal condition. *Journal of dental hygiene JDH American Dental Hygienists Association, 82 Suppl 3*, 4-9.

Eckburg, P. B. (2006). Diversity of the human intestinal microbial flora. *Science, 308(5728),* 1635-1638.

Eisen, J. A. (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biology, 5*(3), e82.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America, 95*(25), 14863-14868. The National Academy of Sciences.

Goodacre, R. (2007). Metabolomics of a superorganism. *The Journal of nutrition, 137*(1 Suppl), 259S-266S.

Havre, S. L., Webb-Robertson, B.-J., Shah, A., Posse, C., Gopalan, B., & Brockman, F. J. (2005). Bioinformatic insights from metagenomics through visualization. *Proceedings IEEE Computational Systems Bioinformatics Conference CSB IEEE Computational Systems Bioinformatics Conference, 4*, 341-350. Ieee.

Haykin, S. (1999). *Feedforward neural networks: an introduction. Nonlinear dynamical systems feedforward neural network perspectives* (Vol. 13, pp. 1-16). Wiley Publishing, Inc.

Hoff, K. J., Tech, M., Lingner, T., Daniel, R., Morgenstern, B., & Meinicke, P. (2008). Gene prediction in metagenomic fragments: A large scale machine learning approach. *BMC Bioinformatics, 9*(1), 217. BioMed Central.

Hooper, L. V. (2001). Commensal Host-Bacterial Relationships in the Gut. *Science, 292(5519),* 1115-1118. American Association for the Advancement of Science.

Hooper, L. V. (2004). Bacterial contributions to mammalian gut development.*Trends in Microbiology*, *12*(3), 129-134.

Hooper, L. V., Midtvedt, T., & Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine.*Annual Review of Nutrition*, *22*(1), 283-307. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA.

Hsieh, L.-C., Luo, L., Ji, F., & Lee, H. C. (2003). Minimal Model for Genome Evolution and Growth. *Physical Review Letters*, *1*(January), 1-4.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era.*Genome Biology*, *3*(2), reviews0003.1-reviews0003.8. BioMed Central.

Kapur, M., & Jain, R. K. (2003). Microbial Diversity : Exploring the Unexplored.*World Federation of Culture Collection WFCC.*

Khachatryan, Z. A., Ktsoyan, Z. A., Manukyan, G. P., Kelly, D., Ghazaryan, K. A., & Aminov, R. I. (2008). Predominant Role of Host Genetics in Controlling the Composition of Gut Microbiota. (J. A. Fraser, Ed.)*PLoS ONE*, *3*(8), 16. Public Library of Science.

Kislyuk, A., Bhatnagar, S., Dushoff, J., & Weitz, J. S. (2009). Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, *10*(1), 316. BioMed Central.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. *Microbiology and molecular biology reviews MMBR*, *72*(4), 557-578, Table of Contents. American Society for Microbiology (ASM).

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, *124*(4), 837-848.

MacFarlane, G. T., & Gibson, G. R. (1991). Co-utilization of polymerized carbon sources by Bacteroides ovatus grown in a two-stage continuous culture system. *Applied and Environmental Microbiology*, *57*(1), 1-6.

Mitsuoka, T. (1992). Intestinal flora and aging. *Nutr. Rev.50*: 438-446.

Moore, W. E. C., & Holdeman, L. V. (1974). Human fecal flora: the normal flora of 20 Japanese-Hawaiians. *Applied Microbiology*, *27*(5), 961-979.

New, T. H. E. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Design* (p.170). The National Academies Press.

Pignatelli, M., Aparicio, G., Blanquer, I., Hernández, V., Moya, A., & Tamames, J. (2008). Metagenomics reveals our incomplete knowledge of global diversity.*Bioinformatics, 24*(18), 2124-2125. Oxford University Press.

Proal, A. D., Albert, P. J., & Marshall, T. (2009). Autoimmune disease in the era of the metagenome. *Autoimmunity Reviews, 8*(8), 677–681. Elsevier.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59-65. Macmillan Publishers Limited.

Salyers, A. A. (1984). Bacteroides of the human lower intestinal tract. *Annual Review of Microbiology*, *38*, 293-313.

Sapp, J. (1994). Evolution by association: a history of symbiosis (p. 272). Oxford University Press, USA.

Sears, C. L. (2005). A dynamic partnership: Celebrating our gut flora.*Anaerobe, 11*(5), 247-251.

Simmering, R., Kleessen, B., & Blaut, M. (1999). Quantification of the flavonoid-degrading bacterium Eubacterium ramulus in human fecal samples with a species-specific oligonucleotide hybridization probe. *Applied and Environmental Microbiology*, *65*(8), 3705-3709. American Society for Microbiology.

Sneath, P. (1984). *Bergey's manual of systematic bacteriology.(Vol II*, p.2648). Baltimore, MD: Williams Wilkins.

Staley, J. T., Castenholz, R. W., Colwell, R. R., Holt, J. G., Kane, M. D., Pace, N. R., Salyers, A. A., et al. (1997). *The microbial world: Foundation of the biosphere*, 1-32.

Taras, D., Simmering, R., Collins, M. D., Lawson, P. A., & Blaut, M. (2002). Reclassification of Eubacterium formicigenerans Holdeman and Moore 1974 as Dorea formicigenerans gen. nov., comb. nov., and description of Dorea longicatena sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology, 52*(Pt 2), 423-428. Soc General Microbiol.

Torsvik, V., & Øvreas, L. (2002). Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology, 5*(3), 240-245. Elsevier.

Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, *6*(11), 805-814. Nature Publishing Group.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., et al. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480-484. Nature Publishing Group.

Vaughan, E. E., Schut, F., Heilig, H., Zoetendal, E. G., De Vos, W. M., & Akkermans, A. D. L. (2005). A molecular view of the intestinal ecosystem. *Current Issues in Intestinal Microbiology*, *1*(1), 1–12. Horizon Scientific Press.

Ward, B. B. (2007). How many species of prokaryotes are there ? *PNAS,99* (16), 10234-10236.

Wong, T.-Y., Fernandes, S., Sankhon, N., Leong, P. P., Kuo, J., & Liu, J.-K. (2008). Role of premature stop codons in bacterial evolution. *Journal Of Bacteriology*, *190*(20), 6718-6725. American Society for Microbiology (ASM).

Woodmansey, E. J. (2007). Intestinal bacteria and ageing. *Journal of Applied Microbiology*, *102*(5), 1178-1186. Wiley Online Library.

Woodmansey, E. J., McMurdo, M. E. T., Macfarlane, G. T., & Macfarlane, S. (2004). Comparison of Compositions and Metabolic Activities of Fecal Microbiotas in Young Adults and in Antibiotic-Treated and Non-Antibiotic-Treated Elderly Subjects. *Applied and Environmental Microbiology*, *70*(10), 6113-6122. American Society for Microbiology.

Yang, B., Peng, Y., Leung, H. C.-M., Yiu, S.-M., Chen, J.-C., & Chin, F. Y.-L. (2010). Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics*, *11*(Suppl 2), S5. ACM.

Zoetendal, E. G., Akkermans, A. D. L., & De Vos, W. M. (1998). Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Applied and Environmental Microbiology*, *64*(10), 3854-3859. American Society for Microbiology.