A Fuzzy Classifier to Taxonomically Group DNA Fragments within a Metagenome

Sara Nasser, Adrienne Breland, Frederick C. Harris Jr., Monica Nicolescu, University of Nevada Reno

Abstract— **Extracting microorganisms from their natural environment has become a popular technique. These metagenomic fragments lack enough information that can mark them into taxonomic groups. In this paper, we implement a fuzzy k-means classifier to separate fragments into taxonomic groups present in a metagenomic data set. The fuzzy classifier is used to group shotgun sequence fragments as small as 500 base pairs according to their DNA signatures, namely GC Content and oligonucleotide frequencies. A comparison of using different signatures is done and we analyze results and compare them. The classifier is also tested to classify Acid Mine Drainage metagenome into classes to represent the major Archea and Bacteria groups. The classification achieved an accuracy of 99% for Acid Mine Drainage a published environmental genome sample.**

I. Introduction

DNA is the building block of all life on this planet, from single cell microscopic bacteria to more advanced creatures such as humans. Microorganisms live in communities, and their structure and behavior is influenced by their habitat. Most microorganisms genomes are known from pure cultures of organisms isolated from the environment, be it a natural organism-associated (i.e, human) or artificial system. Cultivation-based approaches miss majority of the diversity that exists however, such that development of cultivation-free methods has been implied. In the past, microbial DNA was sequenced by culturing microorganisms in a controlled environment. Cultivating these organisms did not reveal enough information about these communities of organisms. Invitro cultivation methods allow the extraction of DNA from only a limited selection of microbial species that can grow in artificial environments. These methods do little to characterize the properties of globally distributed microbes, because the vast majority of them have not been cultured. New techniques in genomic sciences have emerged that allow an organism to be studied in its natural habitat as part of a community. Research has broadened from studying single species to understanding microbial systems and their adaptations to natural environments.

Metagenomics involves sampling of microbial DNA from natural environments rather than relying on traditional, single species cultivation techniques. In this approach DNA of multiple microorganisms is collected from its environment rather than culturing it. This, coupled with rapid developments in molecular biology is changing our understanding of bacterial evolution and naturally existing microbial systems. As an illustration, traditional culture and PCR-based techniques showed a bias of Firmicutes and Bacteroides as the most abundant microbial groups in the human gastrointestinal (GI) tract. Metagenomic sampling has revealed that Actinobacteria and Archaea are actually most prolific [9]. Research has broadened from the study of single species to understanding microbial systems and their adaptations to natural environments. This has been achieved by developing methods which can extract mixed DNA directly from environmental samples [2], [17]. In this way, metagenomics is the application of modern genomic techniques to the study of microbial communities in their natural environments, bypassing the need for isolation and lab cultivation of individual species [4].

The whole genome(DNA) or metagenome population cannot be sequenced all at once because available methods of DNA sequencing can handle only short stretches of DNA at a time. Although genomes vary in size from millions of nucleotides in bacteria to billions of nucleotides in humans, the chemical reactions researchers use to decode the DNA base pairs are accurate for only 600 to 700 nucleotides at a time [27]. Genomes are cut at random positions then cloned to obtain the smaller fragments, also known as shotgun sequences. Obtaining shotgun sequences has allowed sequencing projects to proceed at a much faster rate, thus expanding the scope of the realistic sequencing venture [26].

Even though the metagenomic approach makes the acquisition of genomic fragments easier, the approach has limitations. The diverse genomes fragments acquired together at the same time need to be assembled to make meaningful conclusions. Assembly is a computationally expensive process and can become slow for large data set. The metagenome approach of acquiring DNA fragments often lack suitable phylogenetic marker genes, rendering the identification of clones that are likely to originate from the same genome difficult or impossible [18]. Therefore taxonomical classification of genomic fragments can help assembling sequence.

Pre-assembly grouping of metagenomic fragments into classes can lead to faster and more robust assembly by reducing the search space. This is because DNA from different organisms could be seperated into groups, thus assembly of smaller groups can replace assembling the entire data set. Fuzzy logic is used for classification as sequences contain some errors, thus approximate results suit that data better than crisp results. In this paper we perform a taxonomical classification based on genomic DNA signatures. The DNA signatures chosen are GC content, tri- and tetra-nucleotide

Sara Nasser, Adrienne Breland, Frederick C. Harris Jr., and Monica Nicolescu are with the Department of Computer Science and Engineering, University of Nevada Reno, USA (email: {sara, adrienneb, fredh, monica}@cse.unr.edu).

This work was supported in part by NSF EPSCoR RING III

frequencies. The proposed method uses a fuzzy classifier with the given signatures as feature set. The technique is verified with artificial shotgun sequences created using strains of *Escherichia coli* to measure correctness. It is also used to classify acid mine drainage environmental sequences.

Even though studies have successfully taxonomically differentiated fragments of sizes greater than 1000bp [19], [24] , there is a lack of availability of applications that classify shorter(500-900bp) shotgun fragments. Our approach is designed with goal of classifying these types of fragments. We present a tool that allows user to read fragments and obtain groups of classes that represent taxonomical groups. The remainder of this paper is laid out as follows: Section II presents background information on environmental genomics and DNA signatures. Section III gives an overview of K-means clustering algorithm. Section V contains the conclusions and future work.

II. Background

This section covers the literature review of environmental genomes, DNA signatures and sequence differentiation. We will discuss DNA signatures that are used in our current approach. The sections cover GC content and nucleotide frequencies.

A. Environmental Genomics

Cultivating microorganisms in isolation does not reveal much information about their environment. This culture independent approach was first used more than two decades ago [12]. Metagenomics or environmental genomics has impacted microbiology by shifting focus away from clonal isolates towards the estimated 99% of microbial species that cannot currently be cultivated [4], [7], [15]. An illustrative example of this is the Sargasso sea project [21]. Microbial samples collected through the filtering of sea water contained large amounts of novel genetic information including 148 new bacterial phylotypes, 1.2 million new genes, and 782 new rhodopsin-like photoreceptors. Bacteriorhodopsin enables the capture of light energy without chlorophyll, it was previously unknown that this type of phototrophy was quite abundant in marine waters. A similar metagenomic project giving new insight into natural existing bacterial systems was the sampling of an underground Acid Mine Drainage(AMD) biofilm [20]. Because this sample was from a system with low complexity, almost all DNA from species present were completely reconstructed. This allowed the examination of strain differences and naturally forming lineages. It also enabled access to the full gene complement for at least two species, providing detailed information such as metabolic pathways and heavy metal resistance.

Separation of specific genomic fragments and reconstruction is a complex process that involves identification of certain features exhibited by entire taxonomic groups. These features are used to group the metagenomic sample into classes.

B. DNA Signatures

DNA signatures are specific patterns that are observed within a DNA strand. These patterns can be observed in specific regions such as coding region or can be observed throughout a genome. There have been several studies on the patterns found in DNA sequences. These patterns can lead to certain information about the DNA or a region within the DNA. In context to this paper we will mention two kinds of signatures GC Content and oligonucleotide frequencies.

A DNA strand consists of four nucleotide Adenine(A), Cytosine(C), Guanine(G) and Thymine(T). These four nucleotide have hydrogen bonds that bond them with each other. A bonds specifically with T and C bonds with G. AT pairs have two hydrogen bonds whereas GC pairs have three. The three hydrogen bonds of CG pairs is the fact that makes the bonds thermostable.

GC content is found to be variable with different organisms, this is viewed to be contributed by variation in selection, bias in mutation, etc. [1]. Coding regions within a genome code for genes and are less divergent within populations. Genes represent characteristics of an organism and have a stronger selection process. Studies have shown that the length of the coding sequence is directly proportional to higher GC content [11]. Thus showing a strong correlation between GC content and gene properties. GC content is generally higher as organisms go higher in the taxonomy and becomes low as we go down the taxonomic groups. This property of GC content can be a useful feature to obtain a broad classification of sequences. The pre-assembly binning of Acid Mine Drainage data was performed by binning the fragments by their GC Content [20].

Another signature that has been used frequently for analysis of genome sequences are oligonucleotide frequencies. Nucleotide frequencies are a measure of occurrence of words of fixed sizes in the genomic sequence. The entire genome is scanned to determine the frequencies of each word. The reverse complement of the strand can be scanned n addition to the forward strand.

Studies have shown that oligonucleotide composition within a genome contains bias. These oligonucleotide usage patterns are known to be species-specific [8]. Phylogenetically related groups of sequences may show similar nucleotide frequencies either because of convergence or because they were inherited from a common ancestor [5]. For example a study conducted on *E-coli* revealed a non random utilization of codon pairs [6]. Some of the most frequent codon pairs found were: CTGGCG, CTGGCC, CTGGCA, CTGGAC, AACCCG, CTGGAA. This study and others reveal that there is a non random over representation and under representation of certain codon pairs within a species. Nucleotide frequencies are generally taken from a group of two, three, four, five, or six nucleotides. These are known as di-, tri-(codon), tetra-, penta-, hexa- (dicodon) nucleotide frequencies respectively. Frequencies of larger word size such as penta, dicodon frequencies are considered more reliable as they are robust to insertions and deletions. Obtaining frequencies of larger groups of nucleotides depends on the size of data set. If we were to calculate dicodon frequencies there are a total of 4096 dicodons, a small sample of genomic sequence may not be able to cover the 4096 dicodons. The same sample can easily include all 256 tri nucleotide frequencies. Most commonly used frequencies are tri- and tetra-nucleotides.

Nucleotide frequencies have been extensively used for grouping genomes or for differentiation of genomes. Evaluation of frequencies for separation of fragments based on taxonomy was performed by Teeling et al [19]. In this paper Teeling et al showed that GC content is not sufficient for separating fragments. Tetra nucleotide frequencies showed better differentiation. Fragments of size 40kb were used for the analysis. TETRA, a web based application uses tetra nucleotide frequencies for genome separation [18]. Another grouping based on tetra nucleotide frequencies resembles the phylogenetic grouping of the representative organisms [14]. In another approach differentiation of bacterial genomes was performed using statistical approaches to perform structural analysis of nucleotide sequences [16].

C. Clustering

Clustering of fragments may be viewed as a more structured form of fragment thinning before alignment comparisons are made. Clustering is a process of grouping objects into like groups based on some measure of similarity. Clustering or classification can be achieved by several techniques such as K-means, Bayesian networks and artificial neural networks. A divide-and-conquer strategy for sequence assembly is described in [13]. A K-means clustering scheme was applied to fragments based on their Average Mutual Information (AMI) measures.

K-means is an unsupervised learning algorithm to group objects into categories. It has been widely used in pattern recognition problems. The simplest K-means algorithm places *N* objects into *K* classes by using the minimum distance from the center of *K* to each object. In the simple K-means approach, *K* is fixed a priori. Clustering problems generally derive some kind of similarity between groups of objects. K-means clustering is a simple and fast approach to achieve a grouping for data. A well-known approach to fuzzy classification is the fuzzy C-means algorithm [25]. An improvement of K-means using the fuzzy logic theory was presented [3], in which the concept of fuzziness was used to improve the original K-means algorithm. A K-means algorithm starts with a large number of seeds (initial samples) for the potential clusters. It uses a set of unlabeled feature vectors and classifies them into *K* classes, where *K* is given by the user. From the set of feature vectors *K* of them are randomly selected as initial seeds. Remaining samples are then assigned to a cluster based on their distance from the seed. The feature vectors are assigned to the closest seeds depending on their distance from it. The centroid is recomputed for each cluster and the data points are reassigned. The algorithm runs until it converges or until the desired number of clusters is obtained.

Due to its simple method of using feature vectors as seeds and the arithmetic mean as the center for the clusters, the K-means algorithm suffers from drawbacks. An improvement to this approach was to start with a huge random population of seeds [22]. This method has been shown to find better seeds, since the initial seeds are more than *K* and are distributed in the data set. Even though this was an improvement on the simple K-means, it was limited in its ability to find better centers, since the mean does not always represent the center of a given data. A modified K-means was developed that uses a weighted fuzzy average instead of the mean to get new cluster centers. Using a fuzzy weighted average instead of a simple mean improved K-means and also leads to convergence [3]. In this paper, a modification of the fuzzy K-means algorithm with fuzzy weighted averages is used for fragment clustering.

III. An Overview of the Algorithm

The first step to classification is the identification of the DNA signatures. After the signatures are extracted the feature vector is initialized. The k means algorithm is run to create initial classes. The DNA signatures extracted using the Markov chains are used as feature set for a modified fuzzy k-means algorithm. The fragment classification divides entire data sets into smaller categories.

The operations carried will be described in the following subsections. In Subsections III-A and III-B we describe the method to obtain the nucleotide frequencies and the modified k-means approach for classification in Subsection III-C.

A. GC Content

GC content can be expressed as the percentage of C and G present in the fragment. It is calculated as follows

$$\frac{C+G}{A+C+G+T} \times 100 \tag{1}$$

GC content of two genomes from Acid Mine Drainage (AMD) environmental sample is illustrated in Figure 1. The histogram of GC content is taken over a window size of 700 nucleotides (nt). The classes shown belong to archea and bacteria groups, indicating that the two groups do not have close features. The histograms obtained for the two classes of AMD sample indicate that there is small region of overlap between these datasets. But most of the fragments have non-overlapping GC content.

Figure 2 shows that the GC content of the two *e-coli* genomes. The two plasmids sequences belong to the *Escherichia coli Proteobacteria* obtained from NCBI [10]. The first sample is *Escherichia coli* plasmid pCoo, RefSeq: NC_007635. This complete sequence contains 98,396 base pairs. The second sequence is *Escherichia coli* plasmid, RefSeq: NC_008460, 120,730 base pairs. In both Figure 1 and Figure 2 the histogram of the genomes have overlap one another. The histogram for two Ecoli strains has a larger overlap region indicating that these two sequence groups are closer in phylogeny. These strains belong to the bacteria group and are close in phylogeny and thus their GC content is similar. This indicates that sequences belonging to similar subtypes have close GC content values.

GC Content can prove to be a good parameter to separate fragments. Even though, GC content can separate fragments certain factors need to be considered when using GC content. Fragment size is a factor that needs to be considered, local biases from fragments of smaller size can influence the GC Content value. GC content is known to be more influential in coding regions. Shotgun fragments of metagenomes do not contain information that reveals directly whether certain fragment contains coding regions or the percentage of fragment region that can code for a gene.

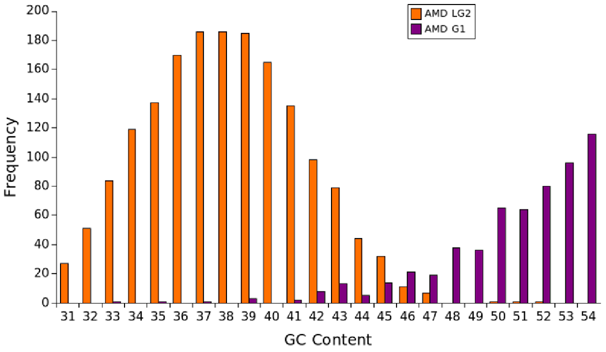

Fig. 1. Histogram of Acid Mine Drainage classes: AMD LG2 refers to *Leptospirillum sp. Group II* environmental sequence of length:960,150 nt. AMD G1 is *Ferroplasma sp. Type II* environmental sequence of length: 1,317,076 nt.

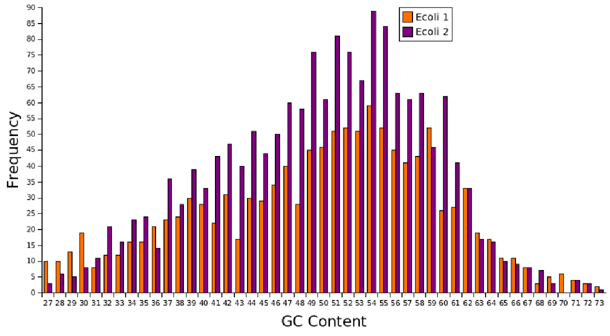

Fig. 2. Histogram of Ecoli strains: Ecoli 1 corresponds to *Escherichia coli* plasmid pCoo, RefSeq: NC_007635. Ecoli 2 corresponds to *Escherichia coli* plasmid, RefSeq: NC_008460.

B. *Nucleotide frequencies using Markov Chain Model*

Markovian models have been used in several fields such as statistics, physics, queuing theory etc. Hidden Markov models are used in pattern recognition to represent unknown probabilities. Markov chain predictors have been used to predict coding regions thus find genes. They have been used for both prokaryotic and eukaryotic genomes. The simplest chain is the zeroth order Markov chain which can be estimated from the frequencies of individual nucleotides A, C, G, T. The approach used to estimate 0th order Markov chain rules is shown below. Consider the sequence GGATCCC the nucleotide frequency is given by:

$$p(GGATCC) = p(G)p(G)p(A)p(T)p(C)p(C) \tag{2}$$

Higher order oligonucleotides can also be determined using zero-order Markov method. It was shown that zero-order Markov method yields greater inter-species distinction [14]. Zero-order Markov method removes biases only from mono nucleotide frequencies thus includes all other oligonucleotide frequencies. Given the dinucleotide frequencies the tri nucleotide frequencies can be estimated by the product of the constituting overlapping dinucleotide frequencies being divided by overlapping single nucleotide frequencies.

$$p(GGATCC) = \frac{p(GG)p(GA)p(AT)p(TC)p(CC)}{p(G)p(A)p(T)p(C)} \tag{3}$$

Higher order Markov chains can also be constructed using only the previous state frequencies. Maximal order Markov chain removes biases from all the previous states and is dependent on only the past state. In our approach we use tri-nucleotide frequencies and tetra-nucleotide frequencies. These can be calculate using a maximal-order Markov chain. Expected values are directly calculated from the observed values as shown in equation 4. In equation 4 and equation 5, O refers to the observed values and E is the expected value.

$$E(N_1N_2N_3) = \frac{O(N_1N_2)O(N_2N_3)}{O(N_2)} \tag{4}$$

$$E(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3)O(N_2N_3N_4)}{O(N_2N_3)} \tag{5}$$

C. *Fuzzy K-means Clustering*

Clustering for a meta-genome assembly problem has a two fold purpose, the first to divide the space for performance improvement. The second purpose of clustering is to group fragments into classes such that each class has fragments from one group. The K-means algorithm uses a set of unlabeled feature vectors and classifies them into k classes. From the set of feature vectors k of them are randomly selected as initial seeds. The feature vectors are assigned to the closest seed. The mean of features belonging to a class is taken as the new center.

The approach used in this paper is described as follows. Given N sequences, such that S=$\{$C$\}$ ^{i},where C =$\{$A, C, G, T$\}$. We randomly select *"K"* sequences as the initial seeds, where *K* is less than the number of sequences *N*. The nucleotide frequencies and GC Content for all sequences are calculated using the above equations. These frequencies form the *p* features to be used in classification.

The sequence is assigned to the class which has the highest fuzzy similarity. The fuzzy similarity is calculated as given below. Let $x_1, ..., x_P$ be a set of *P* real numbers. The number of iteration is given as *r*. The weighted fuzzy average (WFA) is given by

$$\mu^r = \sum_{p=1,P} w_p^{(r)} x_p, \; r \; = 0, 1, 2, \ldots \qquad (6)$$

Here *x* is the parameter or feature and *p* the number of features. Given *i=0,...,N* and *j=0,...,k,* the distance d $_{i,j}$ for each cluster can be calculated as follows:

$$d_{i,j} = max(\mu_j^r), for\ all\ j = 0, \ldots, k \qquad (7)$$

An initial mean is taken and a Gaussian is centered over the mean and weight w_p is obtained for x_P. Feature vectors are assigned to each seed. Empty or small classes are eliminated. Classes that are close to each other are merged to form one class. Cluster centers are replaced with weighted fuzzy averages and feature vectors are reassigned. This process is repeated until convergence.

IV. Clustering via Feature extraction

This section shows the results obtained and describes the genomes used to test the approach.

A. *Artificial Shotgun Sequence Testing*

To assess the performance of fuzzy clustering on genomic sequences, two genomes that are significantly different are used for the first test case. Fragments are generated from genome belonging to plants and a virus genome sample. These fragments are mixed with each other. The classifier first extracts the GC content and the nucleotide frequencies and classifies the fragments into categories. The classifier is run until compact classes are generated. These 'k' classes are greater than or equal to the actual number of genomes classes used. Fragments of average size 750 base pairs (between 500-900) are used in this test case. Table I shows the results obtained after classifying these two samples. In Tables I and II, GC refers to clustering with GC content, T_z refers to tri nucleotide and TR_z tetra nucleotide frequencies using zero order Markov chains. TR_m refers to tetra nucleotide frequencies using maximal order Markov chain, T_m is the tri nucleotide frequencies using maximal order Markov chain. Combinations of different signatures are shown by hyphenating individual frequencies. A value of NA represents that the signatures could not separate the fragments into groups and all the data was placed in one class.

B. *Acid-Mine Drainage Metagenome*

Acid-Mine Drainage Metagenome (AMD) was obtained from Richmond Mine at Iron Mountain, CA. Acid mine drainage environmental genome was shown to contain 5 genomes. We use shotgun sequences of two genomes of AMD, namely Leptospirillum sp. Group II (Lepto) environmental sequence and Ferroplasma sp. Type II(Ferrop. Type II) environmental sequence. These sets are 960,150 and 1,317,076 nucleotide base pairs respectively. The first group belongs to the bacterial genus Leptospirillum; the second one is an archea from the genus ferroplasma. These genomes are almost complete and are available at NCBI. Shotgun sequences of average size 700 base pairs (sizes between 500-900) were created from these genomes. These shotgun fragments are randomly combined with each other to create a environmental genome sample of AMD metagenome. Figure 3 depicts the classification results on AMD data. A smaller set of 3000 sequences was used for the display. Figure 3 indicates that combination of GC content and GC skew were able to separate the two classes. There are few cases where fragments were marked incorrectly.

	# Fragments classified incorrectly		
Sequence	Genome 1	Genome 2	Total %
GC	39	1	0.08
T_z	NA	NA	NA
TR_z	NA	NA	NA
$GC - T_z - TR_z$	18	32	0.1
T_m	7	0	0.02
TR_m	15	0	0.03
T_m- TR_m	11	3	0.028
$GC - T_m - TR_m$	5	1	0.012

TABLE I

Separating 500 Fragments Belonging to Two Organisms Using Different Signatures

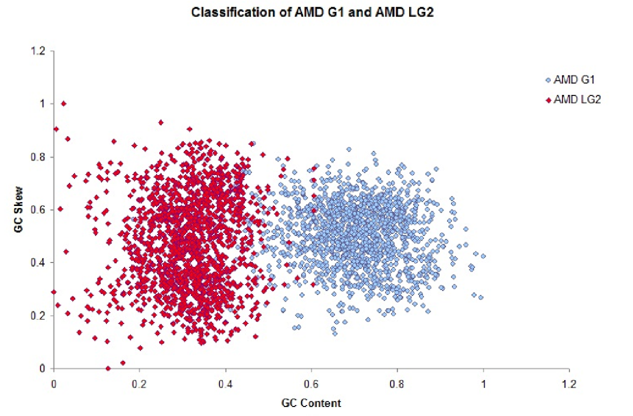

Fig. 3. Classification using GC Content for 3000 Shotgun Sequence Fragments obtained from AMD G1 and AMD LG2

The results of classification using the modified k-means approach using DNA signatures is given in Table II.It compares the classification results for the two AMD genomes. Fragments with average length of 700 base pairs are used. Classification was performed using different combinations of signatures and the results are displayed in Table II. The final classification resulted in two groups one with fragments from Lepto and another with Ferrop. Type II fragments respectively. The results indicate that frequencies obtained using maximal order Markov chain created the better classification than zero order. Combination of different signatures also resulted in lesser misclassifications.

	# Fragments classified incorrectly		
Sequence	Lepto.	Ferrop. Type II	Total %
GC	500	27	0.026
T_z	NA	NA	NA
TR_z	NA	NA	NA
$GC - T_z - TR_z$	640	27	0.033
T_m	147	16	0.0081
TR_m	170	6	0.0088
T_m- TR_m	127	10	0.0068
$GC - T_m - TR_m$	129	11	0.007

TABLE II

Separating 20K fragments from AMD into two classes using different signatures

V. Conclusions and Future Work

A. Conclusions

The paper proposes a fuzzy clustering algorithm to classify shotgun genome fragments into taxonomical classes using combination of DNA signatures. In this paper we proposed few DNA signatures as feature set for genome classification. Results were obtained for an artificial set that was constructed using two different genomes. We classified this set using different signatures, and combination of signatures. Our approach could successfully classify sequences of lengths smaller than 1000 base pairs. We also tested AMD metagenome and classified it into two groups. The results obtained indicated that maximal order Markov chains were the best separators and obtained the best classification. Zero order Markov chain could not classify the data. This could be due to the fact the zero-order chain does not remove dependencies from previous frequencies.

B. Future Work

The results indicate that we were able group shotgun sequences by their frequencies and GC Content. We would like to add higher level frequencies such as penta nucleotide, dicodons, etc. We would like to perform analysis of DNA signatures to find the best discriminatory oligonucleotide pairs. This will enable selection of features that suit the data set rather than using all available frequencies. Improvement to this technique would be to measure the clustering validity and determine the number of classes. We would also like to build a taxonomical pyramid using frequencies that indicates the level in taxonomy that can be classified using DNA signatures.

References

[1] JA. Birdsell. Integrating genomics, bioinformatics and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*, **19**:1181–1197, 2002.

[2] O. Bj, M.T. Suzuki, E.V. Koonin, L. Aravind, A. Hadd, L.P. Nguyen, and et al. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology*, **2**:516529, 2000.

[3] Looney C.G. Interactive clustering and merging with a new fuzzy expected value. *Pattern Recognition*, 35:2413–2423(11), November 2002.

[4] K Chen and L Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology*, **1**:106112, 2005.

[5] G. C. Conant and P. O Lewis. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. *Mol. Biol.Evol.*, **18**:1024–1033, 2001.

[6] GA Gutman and GW Hatfield. Nonrandom utilization of codon pairs in escherichia coli. *Proc Natl Acad Sci U S A*, **86**:36993703, 1989.

[7] P Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biology*, **3**:REVIEWS0003, 2002.

[8] S Karlin, I Ladunga, and BE Blaisdell. Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci U S A*, **91**:1283712841, 1994.

[9] Emmanuel Mongodin, Joanne Emerson, and Karen Nelson. Microbial metagenomics. *Genome Biology*, 6(10):347, 2005.

[10] NCBI. National center for biotechnology information. http://www.ncbi.nlm.nih.gov/, NIH, 2007.

[11] JL. Oliver and A. Marn. A relationship between gc content and coding-sequence length. *Journal of Molecular Evolution*, **43(3)**:216–223, 2004.

[12] GJ Olsen, DJ Lane, SJ Giovannoni, NR Pace, and DA Stahl. Microbial ecology and evolution: A ribosomal rna approach. *Annu Rev Microbiol*, **40** :337365, 1986.

[13] Hasan H. Otu and Khalid Sayood. A divide-and-conquer approach to fragment assembly. *Bioinformatics*, **19(1)**:22–29, 2003.

[14] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.*, **13(2)** :145 – 158, February 1, 2003.

[15] M Rappe and S Giovannoni. The uncultured microbial majority. *Annu Rev Microbiol*, **57**:369394, 2003.

[16] Oleg Reva and Burkhard Tmmler. Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics*, 6(1):251, 2005.

[17] M.R. Rondon, P.R. August, A.D. Bettermann, S.F. Bradly, T.H. Grossman, M.R. Liles, and et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorgansims. *Applications Environmental Microbiology*, **66**:25412547, 2000.

[18] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatic*, **5**:163, 2004.

[19] Hanno Teeling, Anke Meyerdierks, Maragrete Bauer, Rudolf Amann, and Frank Glockner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, **6**:938–947, 2004.

[20] Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37 – 43, 2004.

[21] J. C. et al Venter. Environmental genome shotgun sequencing of the sargasso sea. *Science*, **304**:66–74, 2004.

[22] N. Watanabe and T. Imaizumi. Fuzzy k-means clustering with crisp regions. *The 10th IEEE International Conference on Fuzzy Systems*, pages 199–202, 2001.

[23] RA Welch, V Burland, G 3rd Plunkett, P Redford, P Roesch, D Rasko, EL Buckles, SR Liou, A Boutin, J Hackett, D Stroud, GF Mayhew, DJ Rose, S Zhou, DC Schwartz, NT Perna, HL Mobley, MS Donnenberg, and FR. Blattner. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. *Proc Natl Acad Sci U S A*, **99(26)**:17020–4, 2002.

[24] AC McHardy, HG Martn, A Tsirigos, P Hugenholtz, and I Rigoutsos. Accurate phylogenetic classification of variable-length dna fragments. *Nat Methods*, **4(1)**:63–72, 2007.

[25] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981.

[26] Edmund Pillsbury. A history of genome sequencing. Technical report, Yale University Bioinformatics, 2001.

[27] Mihai Pop, Steven L. Salzberg, and Martin Shumway. Genome sequence assembly: Algorithms and issues. Technical report, 2002.