Neural Network-based Taxonomic Clustering for Metagenomics

Steven D. Essinger, *Student Member,* Robi Polikar, *Member,* and Gail L. Rosen, *Member, IEEE*

Abstract— **Metagenomic studies inherently involve sampling genetic information from an environment potentially containing thousands of distinctly different microbial organisms. This genetic information is sequenced producing many short fragments (<500 base pair (bp)); each is tentatively a small representative of the DNA coding structure. Any of the fragments may belong to any of the organisms in the sample, but the relationship is unknown a priori. Furthermore, most of these organisms have not been identified and correspondingly are not represented in any of the publicly available search databases. Our goal is to be able to predict the taxonomic classification of an organism based on the fragments obtained from an environmental sample that may include many (some previously unidentified) organisms. To elucidate the diversity and composition of the sample, we first use a supervised naïve Bayes classifier to score the fragments of known genomes, followed by an unsupervised clustering to group fragments from similar organisms together. We are then free to analyze each cluster separately. This is challenging since we are not interested in similar sequences, but sequences that come from similar genomes, which are known to vary widely intra-genomically. Our dataset comprises of an extremely challenging scenario involving clustering fragments at the phyla level, where none of the phyla have been previously seen or identified. We present two variations of our proposed approach, one based on ART and K-means. We show that ART can cluster 500bp fragments from 17 novel phyla at an overall isolation/grouping that is 10% better than K-means and nearly 7 times over chance.**

I. Introduction

Metagenomics is primarily concerned with the composition and community function of microbial organisms in their native environments on a molecular basis. Understanding which microbes are present in an environment permits the characterization of the community in terms of diversity. Additionally, and perhaps more importantly, the function and current state of the environment may be described. It is well understood that communities of microbes have substantial impact on the behavior of their environment [1, 2].

For example, it is estimated that humans are comprised of approximately 100 trillion cells [2]. Surprisingly, it has been suggested that there are 10-to-20 times as many microbes inhabiting the human body collectively accounting for 1-2% of our body mass [2]. These microbes are believed to be incredibly diverse across the body varying between sites and between individuals [3]. The majority of these microbes are known to form a symbiotic relationship with the body, for example, microbes in the intestines assist in the production of vitamins such as vitamin B12 [4].

Studies are currently under way investigating the impact of microbial communities in the human gut causing ailments such as obesity and Crohn's Disease [5]. The findings of such studies may not only elucidate the cause of the disease, but hopefully will lead to novel medical treatments. The impact of microbes inhabiting the human body is thought to be so significant that the Human Microbiome Project has been commissioned by the National Institutes of Health to catalog the diversity and function of all of the microbes inhabiting the human body. A few studies under this project already have targeted microbes inhabiting the human gut environment [6].

II. Background

To study a microbe from an environmental sample, traditional genomics prefers first to isolate a microbe from its environment, and culture a small population in the laboratory. The genomic DNA is extracted from this microbe and is sent to a sequencer to call the bases (nucleotides) of the strand of DNA (i.e. **A**denine, **C**ytosine, **T**hymine, and **G**uanine). Contemporary sequencing technology produces short fragments of sequenced DNA anywhere from 35 to 1000bp in length. Pyrosequencing technology such as Illumina® produces read lengths of about 35bp while 454 Sequencing™ produces 400bp reads. Traditional Sanger sequencing produces reads in length upwards of 1000bp. Regardless of the sequencing technology, the outputted contiguous fragments are aligned and assembled to produce the finished sequence of DNA. Baring any contamination in the experiment, we know that this assembled sequence of DNA is germane to the microbe under study since there is only one type of microbe in the culture and therefore one genome.

Metagenomics on the other hand involves characterizing the entire environmental sample, which potentially contains thousands of different microbe strains. It is nearly impossible to isolate and culture each of these strains since most cannot exist outside of their environment. Notably, approximately 98% of all microbes cannot be isolated and cultured and therefore must be studied in their natural environment [7]. In order to circumvent this hurdle, genomic DNA is sampled from the entire environmental sample and is passed through the sequencer. Contradictory to traditional genomics, the fragments cannot be immediately assembled since we do not know a priori which fragment belongs to which microbe. Therefore a logical first step before analysis is to classify fragments that come from known genomes using

Manuscript received February 7, 2010. This material is based upon work supported by the National Science Foundation Grant No: 0845827. Steven D. Essinger is with the Department of Electrical and Computer Engineering, Drexel University, 3141 Chestnut Street, Philadelphia PA 19104 (phone: 609-709-6742; e-mail: sessinger@drexel.edu). Robi Polikar is with the Department of Electrical and Computer Engineering, Rowan University, 201 Mullica Hill Road, Glassboro NJ 08028 (email: polikar@rowan.edu). Gail L. Rosen is with the Department of Electrical and Computer Engineering, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 (email: gailr@ece.drexel.edu).

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

prior knowledge from a database (for example, using supervised learning). For notational purposes, we will refer to these methods as supervised classification while unsupervised/semi-supervised clustering will be referred to as binning.

Most supervised classification methods for metagenomics employ either a homology-based alignment or a composition-based frequency model [8, 9, 10, 11]. However, as mentioned above, either of these techniques can only identify the 1-2% of organisms (those that are known), and perhaps classify another 50-70% to a higher taxonomic level (such as order or phylum). But if no examples of this higher taxonomic level exist in the database, accuracies have been shown to decrease sharply [11]. Therefore, it will be desirable to group metagenomic reads into taxonomic subdivisions when the higher order taxa is not known. For example, if we conclude sequences obtained from an environmental sample must be from a new species – we can use clustering to help us discern if all the fragments are from the same species or cluster into different species, which can then help us to infer new organisms. The total number of clusters created reflects the diversity of the environmental sample; namely the number of different taxonomies of microbes that are present, even if they are unknown.

Currently, most methods for genomic data classification are supervised approaches, with unsupervised and semi-supervised binning only recently emerging. The only fully unsupervised method to our knowledge is LikelyBin, an algorithm that learns the different "source" genomes via a Monte Carlo Markov Chain approach [12]. However, the method is only valid for low complexity samples (2-10 species) and was tested on samples that were sufficiently divergent according to derived criteria. CompostBin is a semi-supervised algorithm for grouping fragments that uses an initial set of sparse labeled data and principle component analysis [13]. They have demonstrated error on several datasets bounded by 10%; however, their datasets are also low-complexity with only 2-6 organisms. Also, self-organizing maps (SOMs) have shown promise, with SOM [14] and Growing-SOM [15] being implemented, with the latter achieving 99+% on low-complexity and 90% on medium complexity samples. The caveat with SOMs is that it was shown to work well only on DNA fragments that are longer than 8kbp and lose much accuracy by 1kbp. Nonetheless, the G-SOM method by Chan uses sparsely labeled data and does not require knowledge of known genomes, making it less supervised than many semi-supervised methods. So far, all clustering methods are infeasible for next-generation sequencing (that produces fragments that are much shorter than 1kbp, some as short as 25bp) and on high complexity samples.

III. METHODS

In this paper, we propose a semi-supervised approach that uses the output of a supervised classifier (e.g. Naïve Bayes classifier) as features that can be fed into an unsupervised clustering method. The novelty of this approach is based on using features as outputs of the supervised classification in such a way that allows us to place the fragment in context to all organisms currently known. In turn, such an approach provides a basis to launch from when attempting to cluster an unknown fragment with like organisms.

Conceptually, the proposed method of processing metagenomic fragments is similar to processing a segment of recorded speech. First, the DNA (audio signal) is extracted from the environmental sample (recording). Next, the DNA is run through a sequencer producing fragments (audio segmentation) of sequenced DNA. The features are then extracted from the fragments via the supervised classifier and then clustered using an unsupervised algorithm (e.g. K-means/ART/SOM). Finally, the results are validated using confidence measures such as bootstrapping or cross validation. The intended result is to produce clusters of fragments, each of which originate from a similar taxonomic class of microbes.

Our unsupervised clustering pipeline begins with scoring each fragment against a training database of known microbe genomes using a Naïve Bayes classifier. This classifier was first implemented for organism classification by Sandberg in 2001 on a small set of just 28 genomes, and has since been further extended to a larger database of 635 genomes by Rosen [9, 16]. The outputted scores for each fragment are then submitted as features to an unsupervised clustering algorithm. In this paper two clustering algorithms have been implemented; K-Means and Adaptive Resonance Theory (ART). [17, 18, 19, 20]. The contents of the clusters are then evaluated and a final score is assigned. The flowchart in Fig. 1 depicts the algorithmic pipeline. A detailed pseudo-code of the algorithm is provided in Fig. 2.

The Naïve Bayes classifier (NBC) begins with calculating the occurrence profile (or frequency profile) of each motif present in a training genome. A motif is an N-mer sequence of DNA, that is, a sequence that consists of N nucleotides. Since there are 4 possible nucleotides, there are 4^N possible different motifs in a genome, and many may occur more than once.

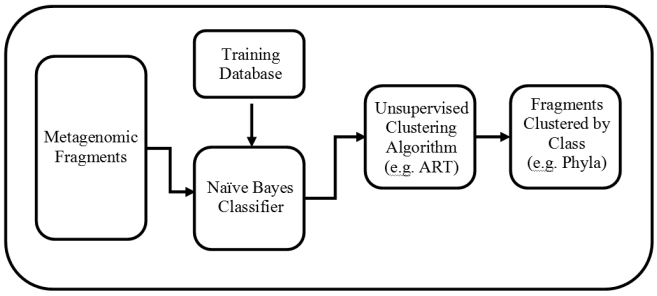

Fig. 1: Proposed algorithmic pipeline. Metagenomic reads (fragments) obtained from sequencer are input to the Naïve Bayes Classifier that has been trained on annotated microbial genomes. Each fragment is scored against each genome in the training database and a corresponding feature vector containing the scores is outputted for each fragment. The fragments are then clustered based on their feature vectors using an unsupervised clustering algorithm. At completion all fragments should be grouped together based on similarity of class (e.g. Phyla).

To run the classifier, a fragment of length S is input to the NBC and all J overlapping motifs are identified. The probability of the fragment belonging to each genome, constitut-

ing the motif frequency profile, is calculated according to Equation (1).

$$P\left(M_j \middle| genome_i\right) = \frac{Frequency\ of\ M_j\ in\ genome_i}{Total\ M\ in\ genome_i} \quad (1)$$
$$i = 1, \cdots, G;\ j = 1, \cdots, 4^N$$

where M_j denote the jth motif, M is the total number of motifs in a given genome, and G is the total number of genomes in the database. Generally, given a fragment f that consists of J motifs, the genome (from the training dataset) with the highest score is selected as the classification decision, where the scores are computed as in Equation (2).

$$S_{f,i} = P\left(f \middle| genome_i\right) = \prod_{j=1}^{J} P\left(M_j \middle| genome_i\right) \quad (2)$$

In other words, the NBC chooses the genome whose frequency profile most closely matched that of the given test fragment. However, the proposed algorithm does not necessarily care about the winning genome, but rather the entire vector of scores of each fragment across all genomes for subsequent unsupervised classification.

For each fragment we use its score across all training genomes as a G-dimensional feature. We handle the unsupervised clustering with an implementation of Fuzzy Adaptive Resonance Theory [19, 20, 21] and alternatively K-means. Our intent was to utilize this tool to classify unknown fragments based on similarity of their feature vectors corresponding to the fragment's strain of origin.

Proposed Algorithm

Input:

- Metagenomic reads (fragments) from next-gen sequencing technology
- Training database (TDB) – consists of G labeled genomes, previously acquired
- Unsupervised clustering algorithm (e.g. ART, K-means)
- Set free parameters (e.g. K in K-means and ν in ART)

Algorithm:

A. Train Naïve Bayes Classifier (NBC) motifs, M of G genome probability profiles
   Do: $i = 1, \ldots, G$
   Do: $j = 1, \ldots 4^N$ (# of diff. motif perm.)
   $$P(M_j|genome_i) = \frac{Freq.\ of\ M_j\ in\ genome_i}{Total\ M\ in\ genome_i} \quad (1)$$
   End
   End

B. Score fragments, evaluate fragment, f using NBC
   Do: $f = 1, \ldots, F$ (# of fragments)
   1. Identify $J\ (N-1)$ overlapping motifs each of length N in fragment, f: $[M_1, M_2, M_3, \ldots, M_J]^T$
   2. Calculate probability of fragment belonging to $genome_i$ in TDB:
   $$Score, S_{f,i} = P(f|genome_i) = \prod_{j=1}^{J} P(M_j|genome_i) \quad (2)$$
   End

C. Build feature matrix for unsupervised classifier

NBC Scores		Features			
		$genome_1$	$genome_2$	…	$genome_G$
Objects	Frag1	S1,1	S1,2	.	S1,G
	Frag2	S2,1	S2,2	.	.
	. . .	.	.	.	
	FragF	SF,1	.	.	SF,G

D. Call unsupervised clustering algorithm
   - Cluster each fragment using corresponding feature vector of dimension G

Output:

- Fragments clustered by taxonomic class (e.g. Phyla, Genus, Strain, etc.)

Test: Figures of Merit

- Accuracy to group similar classes together
$$A_{unity} = \frac{1}{F}\sum_{p=1}^{P}\left[\underset{f_{cp}}{argmax}(f_{cp}|p)\right] \quad (3)$$
- Accuracy of algorithm to isolate dissimilar classes
$$A_{isolate} = \sum_{c=1}^{C}\frac{f_c}{F}\left[\frac{\underset{f_t'}{argmax}(f_t'|c)}{f_c}\right] = \frac{1}{F}\sum_{c=1}^{C}\left[\underset{f_t'}{argmax}(f_t'|c)\right] \quad (4)$$

C: # of clusters
P: # of taxonomic classes (e.g. phyla)
f_c: # of frag. in cluster, c
f_{cp}:# of frag. in cluster, c belonging to taxonomic class, p
f_t': # of fragments from taxonomic class, p
F: total number of fragments in all phyla

Fig. 2: Description of the proposed algorithm in pseudocode. This algorithm is also represented in block diagram form in Fig. 1.

ART analyzes a fragment and assigns it to a category or cluster. If a fragment is found to be substantially different from the existing clusters than a new cluster is created. The stringency of the ART algorithm to discern the similarity between fragments is controlled by a vigilance parameter. This parameter varies between 0 and 1 with 1 imposing the most stringent requirement for similarity. Therefore, one would expect the number of clusters to increase as the vigilance parameter is increased.

The network may grow according to the diversity of the fragment feature vectors without erasing previously assigned clusters. For example, if three clusters have already been formed based on fragments belonging to three different phyla and a fourth fragment has been determined to be significantly different that the other three clusters, a new cluster is formed and thus we have learned (or discovered) a fourth phyla.

We also chose to implement the K-means clustering algorithm for comparison purposes since it is a ubiquitous, easily accessible method [17]. The K-means algorithm clusters the fragments together based on the Euclidean distance between their *G*-dimensional feature vectors. Therefore we would expect excellent performance from this algorithm if our data were distributed spherically.

To assess the performance of the algorithm on our test datasets we invoked two different figures of merit. Our first figure assesses the accuracy of the algorithm to cluster similar classes together as shown in Equation (3).

$$A_{unity} = \frac{1}{F}\sum_{p=1}^{P} \arg\max_{f_{c_p}} \left(f_{c_p} \middle| p \right) \tag{3}$$

Where $f_{c,p}$ is the number of fragments in cluster c that belongs to the taxonomic class (phyla) *p*, *P* is the number of such taxonomic classes, and *F* is the total number of fragments in all of the phyla. In our database, we had *P*=17 previously unseen phyla. For a test dataset including fragments from 17 different phyla, we would ideally expect to see 17 different clusters each containing a different phylum. In such cases, this figure of merit obtains its highest value of 1.

The second figure of merit assesses the accuracy of the algorithm to isolate dissimilar classes as described by Equation (4).

$$A_{isolate} = \sum_{c=1}^{C} \frac{f_c}{f} \left[\frac{\arg\max_{f_t'} \left(f_t' \middle| c \right)}{f_c} \right] = \frac{1}{F}\sum_{c=1}^{C} \arg\max_{f_t'} \left(f_t' \middle| c \right) \tag{4}$$

where f_c is the total number of fragments in cluster c, f_t' is the total number of fragments from the taxonomic class p, and C is the total number of clusters obtained by the unsupervised algorithm. Ideally, we would expect each cluster to contain one class (e.g. one phylum); in which case this figure of merit obtains its highest value of 1, but it's possible that a cluster may contain many classes so it's important to understand the class forming the majority of the distribution within each cluster.

The integrated approach of using unsupervised and supervised techniques together stems from the fact that we have very limited labeled data, and potentially unlimited unlabeled data with unknown future classes. We note that because the number and nature of the classes that may appear in future data are also unknown, most semi-supervised approaches are not immediately applicable to this vast and challenging genomic data. In summary, we use the limited data with available labels to determine the frequency profiles of the features with respect to the known genomes using the naïve Bayes classifier. The posterior probability scores of these features, a measure of how close a given feature vector is to existing genomes - are then used in an unsupervised fashion to establish the inherent clusters in the data obtained from previously unknown genomes. While the number and nature of future classes are unknown, the feature vectors themselves are fixed, since we use all combinations of nucleotides for a given *N*-mer.

IV. Dataset

To evaluate the performance of our approach, we have developed three different experiments, each consisting of a different subset of test data. All experiments were geared towards grouping the test fragments at the phyla taxonomic level. We selected this level since it is comprised of microbes that are much more diverse than those belonging to the levels of genus or species. Therefore, our experiments focus on analyzing extremely challenging scenarios for the classification of short metagenomic fragment reads for taxonomic purposes.

The 635 strains that comprised our dataset were obtained from the National Center for Biotechnology Information (NCBI) repository. These are strains of microbes that have fully sequenced genomes. We have found that the strains in this database span 19 different phyla. Our goal was to use our algorithmic pipeline to cluster fragments that were randomly sampled from these strains into groups based on their phyla class. Each strain was randomly sampled 100 times with each sample consisting of a fragment 500bp (base-pair) in length.

In experiment 1, we set aside the strains belonging to the two largest classes of phyla for the NBC training dataset. This training dataset consisted of 431 strains. The remaining 204 strains spanned 17 different phyla and were used as the test dataset. Therefore, we attempted to cluster fragments 500bp in length into 17 different groups using a training database that did not include any examples of the test dataset, or even any examples of any of the phyla in the test dataset (in other words, the instances in the training dataset did not include any samples from any of the classes in the test dataset). This is an extremely challenging scenario, different from the standard procedure of just keeping training and test datasets mutually exclusive, but consisting of examples of the same classes. This is because our entire set of test fragments is novel to the classifier.

In experiment 2 we reversed the set up of experiment 1. We trained the NBC using the 17 phyla used as the test set in

the prior experiment. We then attempted to cluster the 2 largest phyla into 2 groups. Again, the test fragments were completely novel to the NBC classifier, resulting in a similarly challenging scenario for classification.

In experiment 3, we followed a different approach from the prior two experiments by partitioning the number of strains of each of the 19 classes into two groups. One group of 320 strains was used for training the NBC while the remaining 315 strains were used in the test set. As in all of the experiments, the test set consisted of 100 fragments, 500bp in length, for each strain while the training set consisted of whole-genomes for each training strain.

The results of all three experiments are provided in the next section. Each experiment was bootstrapped 25 times to provide a level of confidence in the results [22, 23].

V. Results

A. Experiment 1: Training on 2 large phyla to cluster 17 smaller phyla

A training dataset of 431 strains was constructed spanning 2 phyla. The remaining 204 strains spanning 17 different phyla were used as test strains as described above. The training dataset consisted of whole genomes while the test fragments were obtained from the test strains by randomly sampling each of them 100 times extracting 500bp nucleotide reads each. Both K-means and ART were implemented to cluster the fragments using the NBC scores as feature vectors. The results are summarized in Table I.

Experiment 1 Performance	Phyla			
	K-M		ART	
	Avg	Std	Avg	Std
Class Unity	0.25	0.05	0.43	0.05
Class Isolation	0.31	0.04	0.34	0.03
# of Clusters	17		17	

Table I. The results of clustering 20400 fragments spanning 17 different phyla when trained on another 2 different phyla using the two figures of merit described in the methods section. The free parameter for K-means was set to 17 and the vigilance parameter, ν, for ART was set to 0.1. Grouping these fragments by chance into clusters of similar phyla we would expect accuracy of 1/17 or 5.9%.

Both algorithms grouped all of the fragments into 17 different clusters as intended. The ART algorithm performed better using both figures of merit than the K-means algorithm. Interestingly, K-means was able to isolate different phyla better than it was able to group similar phyla together while for ART the converse is true. The results imply that ART is grouping similar phyla together, but the clusters are containing more than 1 phyla thereby driving down the score for isolating different phyla. The opposite is true for K-means, suggesting that similar phyla are distributed among several clusters rather than one, but not across all clusters, otherwise we would expect a similar score for isolating different phyla.

B. Experiment 2: Training on 17 smaller phyla to cluster 2 large phyla

A training dataset of 204 strains was constructed spanning 17 phyla; the opposite of experiment 1. The remaining 431 strains spanning 2 different phyla were used as test strains. The training dataset consisted of whole genomes while the test fragments were obtained from the test strains by randomly sampling each of them 100 times extracting 500bp nucleotide reads each. Both K-means and ART were implemented to cluster the fragments using the NBC scores as feature vectors. The results are summarized in Table II.

Experiment 2 Performance	Phyla			
	K-M		ART	
	Avg	Std	Avg	Std
Class Unity	0.73	0.03	0.73	0.03
Class Isolation	0.74	0.04	0.86	0.04
# of Clusters	2		4	

Table II. Results of clustering 43100 fragments spanning 2 different phyla when trained on another 17 different phyla using the two figures of merit described in the methods section. The free parameter for K-means was set to 2. Grouping these fragments by chance into clusters of similar phyla we would expect accuracy of 1/2or 50%. ART grouped these fragments into 4 clusters with the vigilance parameter, ν, set at 0.025.

While K-means was programmed to group all of the fragments into 2 clusters, the ART algorithm exhibited the best performance when grouping all fragments into 4 clusters. The ART algorithm performed slightly better using both figures of merit than the K-means algorithm. K-means was able to isolate different phyla marginally better than it was able to group similar phyla together and the same is true for ART albeit at a larger margin. The performance of the algorithms appear to be much better for experiment 2 than the previous, but it is important to note that chance was 5.8% in experiment 1, while chance is 50% in this experiment.

C. Experiment 3: Training on examples of each phyla to cluster the rest

A training dataset of 320 strains was constructed spanning 19 phyla. The remaining 315 strains also spanned the same 19 different phyla and were used as test strains. The training dataset consisted of whole genomes while the test fragments were obtained from the test strains by randomly sampling each of them 100 times extracting 500bp nucleotide reads each. Both K-means and ART were implemented to cluster the fragments using the NBC scores as feature vectors. The results are summarized in Table III.

Experiment 3 Performance	Phyla			
	K-M		ART	
	Avg	Std	Avg	Std
Class Unity	0.52	0.04	0.51	0.05
Class Isolation	0.22	0.06	0.53	0.05
# of Clusters	19		18	

Table III. The results of clustering 31500 fragments spanning 19 different phyla when trained genomes belonging to the same 19 phyla using the two figures of merit described in the methods section. The free parameter for K-means was set to 19. Grouping these fragments by chance into clusters of similar phyla we would expect accuracy of 1/19 or 5.2%. ART grouped these fragments into 18 clusters with the vigilance parameter, ν, set at 0.105.

K-means was programmed to group all of the fragments into 19 clusters, but the ART algorithm exhibited the best performance when grouping all of the fragments into 18 clusters. K-means was slightly better at grouping similar phyla together, but substantially worse at isolating different phyla. ART scored nearly the same for both figures of merit implying that the distribution of fragments is similar across all of the clusters.

VI. Discussion

Grouping unknown metagenomic fragment reads is inherently a challenging problem. We simply do not know the distribution of classes of taxa in a metagenomic sample prior to sequencing. Therefore we need to rely on clustering these fragments to provide a sense of diversity of the sample. Furthermore, we need the fragments grouped together if we wish to further investigate the sample for functionality since a single fragment by itself may not contain enough information to describe a gene.

The challenge we face is that we cannot simply cluster fragments together that are similar in composition as many clustering methods tend to do. Our problem is to group fragments that belong to similar groups of taxa (e.g. phyla) together. The difference is that fragments belonging to a strain vary greatly in composition because each one represents a different part of the genome. While two strains may be similar inter-genomically, each generally will vary greatly intra-genomically. Since the fragments we are clustering represent short samples of each strain's genome, we expect that the fragments in each cluster will vary greatly. We therefore need clustering methods of greater complexity over simply using basic unsupervised clustering algorithms. This has motivated our use of the NBC for feature extraction in our algorithmic pipeline.

Our experiments have been constructed to simulate challenging scenarios for clustering unknown fragments. The experiments purposely ensured that no class in the test set was represented in the training database. This reflects the current state of metagenomic analysis since most fragments obtained from a metagenomic sample are novel and need to be clustered based on their relation to known classes.

. The first two experiments purposely ensured that no class in the test set was represented in the training database. The third experiment on the other hand was constructed so that the training database contained an example from each class in the test set. This enabled us to observe the performance of the algorithm on the other extreme; complete representation in the training database.

From our experiments we have found that the proposed approach is able to cluster the test fragments substantially better than chance. For example, we expect 5.8% of the test fragments in experiment 1 to group together by chance since the test set spans 17 phyla. ART was able to cluster the fragments with accuracy of 42.9% and 33.5% using the two assessment criteria respectively. Accuracy increased in Experiment 2 and Experiment 3. As anticipated, the accuracy further increased in Experiment 3 over the prior experiments since there was full representation of the test classes in the database.

Throughout all experiments using both assessment criteria, we find that ART generally performs better than K-means. K-means is a simple clustering method that creates clusters based on the Euclidean distance between points. In our experiments the K-means algorithm is calculating the distance between the feature vectors of fragment scores against strains in the database. If the feature vector scores were spherical in nature, then we would expect K-means to perform rather well on the test set. However, the further that the feature vectors deviate from spheres the poorer the performance of the algorithm will be. It is clear that our feature vectors are not spherical in nature. The ART algorithm is a more sophisticated neural network based model that "learns" the pattern of feature vector scores associated with a set of fragments and groups them together based on their similarity. Therefore, we expect that ART would have an advantage over the K-means algorithm since we are grouping based on pattern and not distance.

VII. Conclusion

Compared to other unsupervised and semi-supervised approaches, we cluster shorter reads (500bp) and more strains (200 to 400) than any other method, to show the clustering method's feasibilities on real metagenomics datasets. Most current un/semi- supervised methods have only discriminated up to 10 strains using short reads or long fragments of over 5Kbp for medium complexity samples. Each of these constraints makes these methods disadvantageous for next-generation sequencing of environmental samples. Also, we demonstrate that adaptive resonance theory is able to cluster novel phyla better than K-means when there are a large number of fragments to cluster. We believe this is due to the incremental learning capability of ART and its ability to learn non-spherical clusters. In conclusion, on an extremely challenging dataset of grouping 500bp reads from 204 strains spanning 17 phyla, ART is able to accomplish this with 43% accuracy, demonstrating that this problem is a challenging one.

References

[1] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis and P. Hugenholtz,, "A Bioinformatician's Guide to Metagenomics", *Micro Mol Biol Rev.* ,vol. 72, no. 4, 2008.

[2] G. Rosen, B. Sokhansanj, R. Polikar, M. Bruns, J. Russell, E. Garbarine, S. Essinger, and N. Yok, "Signal Processing for Metagenomics: Extracting Information from the Soup", *Current Genomics*, vol. 10, no. 7, pp. 493-510, 2009.

[3] E. Costello, C. Lauber, M. Hamady, N. Fierer, J. Gordon and R. Knight, "Bacterial Community Variation in Human Body Habitats Across Space and Time", *Science,* vol. 320, no. 5960, pp. 1694-1697, 2009.

[4] M. Wilson, *Bacteriology of Humans: An Ecological Perspective.* Malden, MA: Blackwell, 2008.

[5] D. Frank and N. Pace, "Gastrointestinal microbiology enters the metagenomics era", *Curr. Opin. in Gastro.,* vol. 24, no. 1, pp. 4-10, 2008.
[6] J. Peterson et al., "The NIH Human Microbiome Project", G*enome Research*, vol. 19, no 12, pp. 2317-2323, 2009.
[7] J. Handelsman, *Committee on Metagenomics: Challenges and Functional Applications*, N. R. Council, Ed. The National Academies Press, 2007.
[8] D. Huson, A. Auch, J. Qi, and S. C. Schuster, "MEGAN Analysis of Metagenomic Data", *Genome Research,* vol. 17, no. 3, 2007.
[9] G. L. Rosen, E. M. Garbarine, D. A. Caseiro, R. Polikar, and B. A. Sokhansanj, "Metagenome Fragment Classification Using N-Mer Frequency Profiles", Hindawi Adv Bioinfo., vol. 2008, (2008).
[10] T. Madden, *The NCBI Handbook*. Ch. 16, pp. 1-17, 2003.
[11] A. Brady and S. L. Salzberg, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models", *Nature Methods*, vol. 6, no. 9 pp.673-U68, 2009.
[12] A. Kislyuk, S. Bhatnagar, J. Dushoff and J. S. Weitz, "Unsupervised statistical clustering of environmental shotgun sequences", *BMC Bioinformatics*, vol. 10, no. 316, 2009.
[13] Sourav Chatterji, Ichitaro Yamazaki, Zhaojun Bai and Jonathan Eisen, "CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads", RECOMB, 2008.
[14] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, T. B. Ikemura, "Informatics for Unveiling Hidden Genome Signatures", *Genome Research,* vol. 13, no. 4, pp. 693-702, 2003.
[15] C. Chan, A. Hsu, S. Tang and S. Halgamuge, "Using Growing Self-Organizing Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing", *Hindawi Journal of Biomedicine and Biotechnology*, vol 2008, 2008.
[16] R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier," *Genome Research,* vol. 11, no. 8, pp. 1404–1409, 2001.
[17] R. Duda, P. Hart and D. Stork, *Pattern classification.* Wiley, New York (2001).
[18] J. MacQueen, *Some methods for classification and analysis of multivariate observations. In: L.M.L.* Cam and J. Neyman, Editors, Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, Statistics vol. 1, University of California Press, 1967.
[19] G. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine", *Computer Vision, Graphics and Image Processing,* vol 27, pp. 54-115 1987.
[20] G. Carpenter, "Distributed Learning, Recognition, and Prediction by ART and ARTMAP Neural Networks", *Neural Networks,* vol. 11, no. 8, pp. 1473-1494, 1997.
[21] G. Carpenter, S. Grossberg and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system", *Neural Networks,* vol. 4, pp. 759-771, 1991b.
[22] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Proc. 14th Intl. Joint Conf. on Artificial Intelligence* (IJCAI), 1995, pp. 1137–1143.
[23] P. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*, Prentice-Hall, London, 1982.