Comparison of Statistical Methods to Classify Environmental Genomic Fragments

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

***Abstract*—"Binning" (or taxonomic classification) of DNA sequence reads is an initial step to analyzing an environmental biological sample. Currently, a homology-based tool, BLAST, is one of the most commonly used tools to label DNA reads, but it is argued that BLAST will quickly lose its classification ability as the genome databases grow. In this paper, we compare the accuracies of a naïve Bayes classifier (NBC) and statistical language model to BLAST for binning reads and demonstrate that NBC obtains good performance for the low cost of computational complexity. On the other hand, the back-off n-gram language model can improve accuracy when only partial training data is available (such as in-progress sequencing projects). NBC demonstrates comparable performance to BLAST and can also be optimized on partial training datasets by adjusting the word feature size. A fivefold cross validation is conducted to compare each method's accuracy for determining novel genomes at different taxonomic levels, with NBC outperforming BLAST for species-level classification but BLAST outperforming NBC for genus-level and phyla-level classification. In conclusion, the NBC is a competitive taxonomic classifier, and language models can improve performance when only partial training data is available.**

***Index Terms*—Bayesian classification, DNA, language models, metagenomics.**

I. Environmental DNA Classification

OUR ABILITY to sequence and study whole microbial genomes is impaired because most microbes (over 99%) cannot be cultured in isolation [1]. High-throughput approaches, or metagenomic methods, propose to sequence the DNA that is present in a sample en-masse without the need for prior cultivation using next-generation sequencing technology. But this technology processes and fragments all DNA in a sample indiscriminately, and therefore the fragments need to be labeled as particular taxa, organisms or higher level families of organisms, efficiently and accurately. A graphical illustration comparing traditional genomics to the problem of metagenomics is shown in Fig. 1.

Traditionally, sequence classification methods align two sequences (usually homologous genes) to compare their similarity and are based on dynamic programming techniques, with the most popular tool being BLAST [2]. Relying on homology is feasible, but homology-based methods' ability to assign short-reads to strains in the database yields many ambiguous results, and it has been recently reported that BLAST breaks down when going from long 600–900 bp reads to short-reads for metagenomics data [3]. It is also hypothesized that as the database of organisms grows, the complexity of the search will grow which will also disadvantage BLAST. Therefore, we seek a framework that represents the entire DNA in a sample without prior knowledge of the genes, promoters, etc. in the DNA sequences.

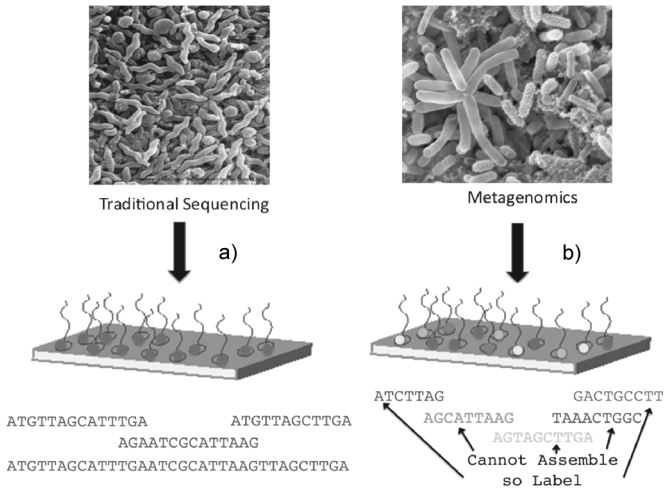

Fig. 1. (a) In classical genomics, one type of bacteria is usually cultured then its DNA is isolated, sequenced, and assembled. (b) In metagenomics, each fragment can be from a different organism, therefore assembly is not viable, and methods are needed to decipher the origins of the fragments.

This field of analyzing complex mixtures from environmental samples is coined metagenomics. The increased complexity of the data from various environments poses challenges in assembling, annotating, and classifying genomic fragments from multiple organisms. Complications also stem from the difficulty of assembling, annotating, and classifying the short sequence fragments typically obtained with next-generation sequencing methods. So, novel computational methods are needed to address these issues and the massive amounts of sequence data that have become available through recent technological advances.

Currently, the most widely used tool for DNA string search is BLAST (Basic Local Alignment Search Tool), an approach based on dynamic programming [2]. Yet, it has been shown that the closest BLAST hit is often not the nearest taxonomic neighbor [11] and without questioning the results, most metagenomic analysis relies on BLAST [6], [16], [34], [35]. Only recently researchers have begun to analyze and compare the performance of BLAST for metagenomic datasets [3], [18]. Simply

Manuscript received March 25, 2010; accepted September 20, 2010. Date of publication September 27, 2010; date of current version February 02, 2011. This work in this paper is supported in part by a National Science Foundation CAREER award #0845827 and a Department of Energy award DE-SC0004335. *Asterisk indicates corresponding author.*

*G. L. Rosen is with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, 19104 USA (e-mail: gailr@ece.drexel.edu; sde22@drexel.edu).

S. D. Essinger with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, 19104 USA (e-mail: gailr@ece.drexel.edu; sde22@drexel.edu).

Digital Object Identifier 10.1109/TNB.2010.2081375

TABLE I
METHODS FOR TAXONOMIC CLASSIFICATION

Features	Classifier	Published Method
Homology-based	Nearest-Neighbor	BLAST [2]
	Nearest-Neighbor & Last Common Ancestor	MEGAN [6]
Composition-based	Naïve Bayesian	Sandberg et al. [7]
		RDP classifier (16S sequences only) [8]
		Rosen et al. [9]
	Support Vector Machines	PhyloPythia [10]

classifying genomic fragments based on a best BLAST hit will yield reliable results only if close relatives are available for comparison. While recently published MEGAN software relies on BLAST for analysis, it attempts to address this problem by classifying DNA fragments based on a lowest common ancestor algorithm (LCA) [6]. LCA allows fragments to generalize to a higher branch in the tree and not the nearest neighbor. Mavromatis *et al.* [19] show that homology-based approaches have lower specificity and hence are not very accurate. But, it has been shown that utilizing all the random sequence reads (RSRs) in a sample has comparable performance and can be faster and cheaper than extracting 16S rRNA genes alone [18]. Therefore, the naïve Bayes classifier and language models offer an interesting alternative to BLAST for processing mass amounts of randomly fragmented segments.

Signal processing and machine learning disciplines are well-equipped to solve problems where background noise, clutter, and jamming signals are commonplace. Hidden Markov models (HMMs), originally popularized for speech processing, have been used for over a decade for gene recognition [4], and it has been found that many techniques used in speech and text mining can now be applied to biology. Metagenomics allows the classification of millions of organisms and their genes, including identifying particular community differences and markers. Supervised and unsupervised machine learning methods, language models, linear classifiers, advanced Bayesian techniques, etc. are all promising to advance rapid annotation and comparison of samples. In this paper, we compare a Bayes classifier and statistical language model, which are able to identify significant features and classify sequences in a blind and high-throughput manner.

II. SUPERVISED TAXONOMIC CLASSIFICATION

Supervised classification methods have traditionally been more popular, since unsupervised methods rely on intrinsic, possibly false, assumptions of the data. The disadvantage of supervised methods is the lack of sufficient data for training. Only a fraction of the species diversity exists in the current databases, and estimating diversity has been seen as unknowable as it is in constant change [5], making supervised approaches difficult to apply. However, as our knowledge of genomes expands, supervised methods hold promise to learn the data that will become available.

In this section, we review several methods in Table I.

1) Homology-Based Approaches: Many current approaches align sequenced fragments to known genomes using homology [3], [6], [11]–[16]. DNA is fragmented during sequencing so that the sequencer can "read" (or call the bases of) a relatively short length of DNA. Usually, the shorter the fragment, the shorter the time it takes to sequence, thereby driving next-generation technology. Short-reads are generally not unique, thus yielding ambiguous classifications, and this has cast doubt about their applicability to metagenomics [3], [11], [15]. Therefore, when classifying sequences, an important aspect is to assess methods for these short-reads.

When the Venter Institute first shotgun-sequenced fragments from the Sargasso Sea, the natural first step was to BLAST these sequences against the comprehensive Genbank database [12], [17]. However, the closest BLAST hit is often not the nearest neighbor [11]. Yet, without questioning the results, most metagenomic analysis relies on BLAST [6], [13], [16]. Only recently researchers have begun to analyze and compare the performance of BLAST for metagenomic datasets [3], [18]. Simply classifying genomic fragments based on a best BLAST hit will yield reliable results only if close relatives are available for comparison. While recently published MEGAN software relies on BLAST for analysis, it attempts to address this problem by classifying DNA fragments based on a lowest common ancestor algorithm (LCA) [6]. LCA allows fragments to generalize to a higher branch in the tree and not the nearest neighbor. Mavromatis *et al.* [19] show that homology-based approaches have lower specificity and hence are not very accurate. But, it has been shown that BLASTing all random sequence reads (RSRs) in a sample has comparable performance and can be faster and cheaper than extracting 16S sequences alone [18].

A notably relevant analysis demonstrates the drawbacks of using BLAST to identify short-reads from next-generation technology. For most metagenomics datasets to date, the significant BLAST hits only account for 35% of the sample [3]. Wommack *et al.* [3] take long-read metagenomic samples and randomly chooses a shorter read within the larger one. The performance of BLAST nucleotide annotation is compared to BLAST for protein function classification using Clusters of Orthologous Genes (COGs). Short-reads retrieve up to 11% of the sample with correct BLAST hits and significance. They find that short-reads tend to miss distantly related sequences and miss a significant amount of homologs found with long-reads. Therefore, improving short-read (less than 400 bp) taxonomic and functional classification are open problems.

2) Composition-Based Approaches: Besides homology, there are many sequence-composition based approaches [7]–[10], [20]–[28]. Compositional approaches use features of length-N motifs, or N-mers, and usually build models based on the motif frequencies of occurrence. Intrinsic compositional structure has been instrumental in gene recognition through Markov models [4] and in tandem repeat detection [29], [30]. In [20]–[22], [24]–[28], evolutionary and classification methods are based on di-, tri-, and tetra-nucleotide compositions, which

soon lead researchers to look at longer oligos for genomic signatures [23]. Wang *et al.* [8] use a naïve Bayes classifier with 8 mers (N-mers of length 8) for 16S recognition. Researchers have since investigated ranges of different oligo-sized frequencies, with the initial pioneering work and the first naïve Bayes implementation by Sandberg *et al.* [7]. McHardy *et al.* [10] found that 5 mer and 6 mer signatures worked the best for support vector machine (SVM) classification, but they concluded that accurate classification only occurs for read-lengths that are $\geq$1000 bp. Sandberg *et al.* were able to obtain over 85% genome-accuracy performance for 400 bp fragments using 9 mers on a dataset of 28 species. Rosen *et al.* [9] took this further to show that the method can achieve 88% for 500 bp fragments, but more impressively, it can achieve 76% for strain-accuracy for 25 bp fragments.

Wang *et al.* [8] shows reasonable classification of 16S rRNA sequences while Rosen *et al.*'s [9] technique can use any fragment including reasonable performance on short-sequence reads. Because Manichanh *et al.* [18] shows RSR-based classification is advantageous to 16S, Rosen *et al.*'s approach has its advantages, especially since the approach achieves 76% accuracy for ALL 25 bp reads at the strain-level. Wang *et al.* verifies that with 16S rRNA sequences, one can get 83.2% accuracy (200 bp fragments) and 51.5% (50 bp) on the genus-level via a leave-one-out cross validation (CV) test set. For comparison, Rosen *et al.*'s naïve Bayes classifier (NBC) achieves 95% accuracy for 100 bp and 90% accuracy for 25 bp fragments on the species level.

The 1000+ completely sequenced microbial genomes, as of March 2010, are still an incomplete representation of extant diversity, as the microbial sequencing projects grow exponentially. Metagenomic data will produce a significant set of sequences that cannot be assigned to any known taxon, and the question arises how to estimate the number of unknown species. For example, Huson *et al.* show that anywhere between 10% and 90% of all reads may fail to produce any hits [6].

III. Classical Bayesian Classification and Language Modeling

A naïve Bayes classifier (NBC) applies Bayes theorem for classification and is based on the assumption that each feature, in this case words, in the classification is independent of each other. This assumption has the advantage of greatly simplifying maximum likelihood estimation of unknown genome-conditional word occurrence probabilities. However, in statistical language models, these estimates are usually modified by application of a heuristic parameter-smoothing technique, that uses lower order word-lengths to avoid (overfitted) null estimates of words occurrences. It is hypothesized that such estimates will improve performance, and we show that it does in the case of higher level taxa recognition using partial training data. But we also show that a simple Bayesian classifier based on naïve assumptions is a competitive classifier for recognizing **novel** genomes of known taxa. We compare NBC and back-off n-gram language modeling with the current method used in bioinformatics, BLAST (Basic Local Alignment Search Tool).

A. Naïve Bayes Classification

The NBC algorithm has been shown to perform well in complex situations, despite its strong independence assumption [31]. In this case, our features are composed of DNA words (N-mers). N-mers are DNA base sequences of length N that may or may not be overlapping, but are overlapping for our classifier. The classifier that is used maximizes the likelihood that a particular fragment comes from a specific genome and is defined as follows:

$$\begin{aligned} l_{\max} &= \arg\max_j log\left(P(\mathbf{w}|C_j)\right) \\ &= \max_j \sum_{i=1}^{K} log\left(P\left(b_i^{i+N}|C_j\right)\right) \end{aligned} \tag{1}$$

where $\mathbf{w}$ is the fragment made up of : K N mers, $\mathbf{w} = [b_1^{1+N} b_2^{2+N} \ldots b_{K*N}^{K*N+N}]$, with each ith N-mer consisting of nucleotide base history $b_i^{i+N} = [b_i, b_{i+1}, ..b_{i+N}]$. C_j is the jth genome, and K is the number of words in the $K * N$-length fragment.

For small N, all possible N-mers are expected to exist. But as N gets large, the average word occurrence frequency decreases, and some words are unseen, or have null estimates, termed nullomers. The nullomers in the NBC case are given a low value, arbitrarily chosen as 3.8×10^{-14} instead of 0; this was calculated by taking 1 over the longest genome size. This ensures that when the log computations are performed, that a log probability of $-\infty$ does not skew the summation, and that a final unique score can be given. The NBC assumptions are rudimentary; therefore, it is hypothesized that a more intelligent estimation of null estimates will improve performance. We aim to estimate word occurrences with a back-off n-gram language model in the next section.

B. Back-Off Modeling

Back-off n-gram language modeling (shortened to "back-off modeling" for the rest of the paper) does not assume independence and is based on a conditional probability model. In the back-off model [32], the conditional probability of a series of nucleotide bases, $\mathbf{b}$, given history h, $p(\mathbf{b}|h)$, is estimated according to the $n-1$ precedent bases in h

$$P(b_i|h) = P\left(b_i|b_{i-(n-1)}^{i-1}\right) \tag{2}$$

where $b_{i-(n-1)}^{i-1} = [b_{i-(n-1)}, b_{i-(n-2)}, ..b_{i-1}]$.

N-mers can be estimated according to the recursive context

$$P\left(b_i|b_{i-(n-1)}^{i-1}\right) = \begin{cases} \hat{P}\left(b_i|b_{i-(n-1)}^{i-1}\right) & if\ N\left(b_{i-(n-1)}^{i}\right) > 0 \\ \alpha\left(b_{i-(n-1)}^{i-1}\right) P\left(b_i|b_{i-(n-2)}^{i-1}\right) & \text{otherwise} \end{cases}$$

where α is a normalizing constant to constrain the area of the probability distribution function to 1, $N(b_{i-(n-1)}^{i})$ is the frequency occurrence of the $n-1$th N-mer, and $\hat{P}$ is the smoothed probability model. This model assumes that the

frequency occurrences will be used if they exist, otherwise lower order N-mers will approximate the probability of an unseen N-mer.

NBC does not take into account the disparities in probability mass of different N-mer occurrences, especially those N-mers that do occur to those that do not occur. Therefore, Good–Turing [33] frequency estimator compensates for probability distribution inconsistencies and in effect smooths them to get a better estimate. The derivation will not be explained here, but essentially, a coefficient, based on the previous bases, is multiplied to the conditional probability estimator to discount the probability mass smoothly

$$\hat{P}\left(b_i|b_{i-(n-1)}^{i-1}\right) = d_{N\left(b_{i-(n-1)}^{i}\right)} \frac{N\left(b_{i-(n-1)}^{i}\right)}{N\left(b_{i-(n-1)}^{i-1}\right)} \quad (3)$$

The coefficient, d_N, is determined based on the history of the N-mer.

IV. Dataset

A total of 635 distinct microbial strains, downloaded in 2008, were used. The standard hierarchy of the taxonomy used in this paper is Phyla → Order → Family → Genus → Species → Strains. The 635 microbes belong to 470 distinct species and 260 distinct genera in this dataset. While 66 species contain more than one strain, 89 genera contain more than one strain. This shows that some knowledge will be lacking when it comes to species- and genus- class diversity. The microbial strains genome lengths range from 160 Kbp (base pairs) for Candidatus Carsonella to 13 Mbp for Sorangium Cellulosum.

V. Results

Although all the microbial strains that were acquired for training are listed as "completed" in Genbank [36], it is of interest to test the performance of classifiers using partial knowledge. There are many microbial projects only partially completed and listed as microbial "genomes-in-progress" in Genbank. In fact, as of this writing, there are approximately 2 uncompleted microbial genomes for every completed one. Another interesting aspect is that only around 1000 microbes are completely sequenced out of millions in the environment. So we seek to answer the question: Using a subset of microbial genomes for training, is it possible to predict taxonomies of novel genomes?

In this section, we show the comparison of NBC, using full and partial training datasets. The performance of NBC versus the back-off models are shown for partial training data for 635 genomes. Finally, we show how the classifiers perform for identifying various taxonomic levels of novel strains, via cross validation analysis.

A. Comparison of Partial Training Sizes on NBC Performance

To show the effect of training data size on the NBC performance, we test the classifier on three different training data scenarios (shown in Fig. 2): 1) full training-data, 2) 5 Mbp per genome (100 × 50 Kbp fragments), and 3) 1 Mbp per genome (100 × 10 Kbp fragments). The test fragments were chosen to be 500 bp long (which is approximately the modal read length from current pyrosequencer technology) [37]. For each genome, 100 test fragments were scored and averaged, totaling 63 500. Because some genomes are shorter than 5 Mbp or 1 Mbp, these genomes are oversampled for these scenarios. This is a comprehensive study of what performance can be expected when all testing genomes come from the training set but only partial genomes are available for training.

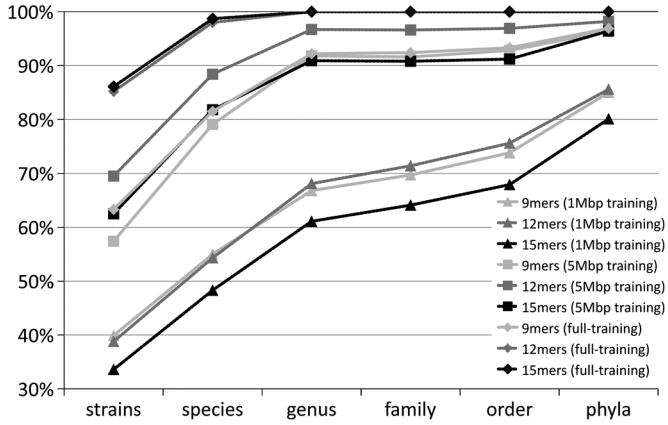

Fig. 2. A comparison of varying partial training data sizes on NBC accuracy (%) versus taxonomic level. For each of the 635 microbes, the full genomes, 100 × 50 Kbp random fragments (5 Mbp of each genome), and 100 × 10 Kbp random fragments (1 Mbp of each genome) were trained on. The same 100 × 500 bp fragments per genome were used for each test. This demonstrates that the less training data available for organisms, the lower the accuracy for lower taxonomic levels, but upper levels will still be able to resolve with fairly high accuracy.

The results in Fig. 2 show that when the full training-data is used, the 15 mers perform the best, with little improvement over 12 mers. Significant performance is lost for 9 mers in the full-training data case. 3%–15% (phyla/strains) accuracy is lost when only using 5 Mbp per genome as opposed to the full genome in the training. In this case, it is interesting to note that 12 mers significantly outperform 15 mers. Finally, for the 1 Mbp partial training data, over 40% accuracy is lost as compared to the full training data case for strains (which shows that strain genotype signatures are less likely to be captured). But for phyla-recognition, only 15% accuracy is lost, showing that higher level taxa signatures can still be characterized with less training data and 9 mers/12 mers perform significantly better with less training data than 15 mers.

B. NBC Versus Back-Off Using Partial Training Data for the 635 Microbial Genomes

In order to compare NBC to back-off models, the partial-training set of 1 Mbp per genome (100 × 10 Kbp fragments) and test set of 100 × 500 bp fragments per genome were used. As noted before, NBC does not perform optimally with 15 mers with this training database. In Table II, the accuracy of 9 mers versus 15 mers for both NBC and back-off models are compared. As shown in Fig. 2, NBC's performance degrades with 15 mers. The interesting aspect here is that the back-off model improves performance for genus level and higher. In fact, the back-off models for 9 mers perform better than the 12 mer-NBC for the levels of Order (NBC 12 mers: 75.6% versus back-off

TABLE II
THE TAXONOMIC CLASSIFICATION OF 100 500 bp FRAGMENTS COMING FROM 635 GENOMES. ALL 635 GENOMES THAT WERE TESTED WERE ALSO TRAINED ON BUT WITH ONLY 100 × 10 Kbp (1 Mbp) PARTIAL DATA. IF SOME AMBIGUOUS HITS FOR BLAST ARE INCORRECT, THE MATCH IS MARKED AS INCORRECT. THIS TABLE DEMONSTRATES THAT BACK-OFF MODELING CAN IMPROVE PERFORMANCE FOR RECOGNITION OF GENOMES THAT WERE ONLY PARTIALLY TRAINED ON (e.g. INCOMPLETE GENOMES), AND THAT NBC OUTPERFORMANCES BLAST BY 5–10%

	9mers (NBC)	9mers (Back-off)	15mers (NBC)	15mers (Back-off)	BLAST
Strains	**39.9%**	37.0%	33.6%	25.3%	27.4%
Species	**55.0%**	52.9%	48.3%	38.3%	44.0%
Genus	66.8%	**67.8%**	61.1%	44.9%	61.1%
Family	69.7%	**71.1%**	64.1%	46.4%	65.7%
Order	73.8%	**76.0%**	67.9%	49.4%	70.0%
Phyla	85.0%	**87.1%**	80.1%	67.8%	82.0%

TABLE III
SPECIES AND GENUS FIVEFOLD CROSS VALIDATION USING THE FULL GENOME TRAINING DATA. NBC PERFORMS BETTER THAN BLAST BY 1% FOR THE FINE RESOLUTION OF SPECIES WHILE BLAST OUTPERFORMS NBC BY 5% FOR GENERA

	Species 5-CV (77 strains)		Genus 5-CV (216 strains)	
	9mers	15mers	9mers	15mers
NBC	$85.8\% \pm 1.1\%$	$\mathbf{97.3}\% \pm 1.1\%$	$75.6\% \pm 2.7\%$	$81.2\% \pm 4.6\%$
Back-off	$53.5\% \pm 3.6\%$	$84.5\% \pm 2.8\%$	$58.2\% \pm 4.3\%$	$45.3\% \pm 5.6\%$
BLAST	$96.1\% \pm 0.6\%$		$\mathbf{86.4}\% \pm 1.9\%$	

9 mers: 76%) and Phyla (NBC 12 mers: 85.6% versus back-off 9 mers: 87.1%). It is hypothesized that the back-off models help smooth out some of the missing information.

C. Cross Validation Analysis for Taxonomic Classification

In this section, we wish to answer the question that if we have a partially populated database representing a finite set of taxa, with what accuracy can novel strains be classified into various taxonomic-"resolutions"? The full genomes were used in the training data, and again 500 bp fragments were used for testing. Yet for this time, datasets that have sufficient representations of various taxonomic levels are carefully selected, and "test strains" are left out each trial and evaluated to see how well they are classified into the sufficiently represented taxonomies. The results are compared to BLAST that also only contains the "training genomes" in its database and have test-strain fragments as the query sequences.

1) Species: Nine species-classes have five or more example strains, and therefore we determine fivefold cross validation to be sufficient for this small dataset. The 9 species-classes, containing 77 strains (approximately 8–9 strains per species), are selected. For each fivefold cross validation set, about 62 strains are trained on while about 15 strains are left out (approximately 1/5 of each class, or 1–2 strains per species are left out for each trial).

The results in Table III show that compared to the language models, NBC using 15 mers has the highest accuracy, with 97.3% and is slightly better than BLAST's 96.1%. The back-off model in this case receive considerably reduced performance. For both 9 mers and 15 mers, performance decreases by using a back-off model.

TABLE IV
PHYLA FIVEFOLD CROSS VALIDATION USING THE FULL GENOME TRAINING DATA. THE BEST PERFORMANCE IS BLAST BUT ONLY 2% BETTER THAN NBC

Phyla 5-CV (100 strains)		
	9mers	15mers
NBC	$82.2\% \pm 5.4\%$	$78.6\% \pm 6.5\%$
Back-off	$81.6 \pm 5.1\%$	$61.6\% \pm 8.4\%$
BLAST	$\mathbf{84}\% \pm 4.1\%$	

Previously, species fivefold cross validation was also compared for 25 bp sequences, and performed with 90.2% ± 1.2% accuracy while BLAST performed at 89.2% ± 1.9% [9]. This demonstrates that as long as species-classes are well-represented, that new species can be predicted with high accuracy. Also, NBC classifies better when the resolution of the taxonomy is fine (e.g. species). In fact, for strain resolution, BLAST is only able to resolve 67% of the strains uniquely while NBC achieves 75% [9]. It is hypothesized that NBC is able to learn the tight clusterings of lower level classes better than the large umbrella (large variance) of higher level taxa.

2) Genera: In order to get a reasonable sample of strains for genus-level fivefold cross validation analysis, we selected strains that have at least 10 strains per genera. There are 15 such genera that meet this criterion and on average, 2–3 strains can be left out of the training per run, to give a fair fivefold cross validation. The 15 genera contain 216 strains (almost 1/3 of the original dataset). For each cross validation set, around 173 strains (11–12 strains per genus) are trained on and around 43 strains (2–3 strains) are left out for each trial.

The results in Table III show that out of the composition methods, NBC using 15 mers perform the best, and the back-off models only degrade performance. BLAST outperforms NBC by about 5%.

3) Phyla: For the phyla dataset, 100 strains were chosen that represent 4 phyla (2 well-represented and 2 underrepresented): proteobacteria, firmicutes, cyanobacteria, and tenericutes. The proteobacteria phyla contains 42 strains, 39 species, 20 genera, and 17 orders. The firmicutes phyla contains 36 strains, 29 species, 20 genera, and 16 orders. The cyanobacteria phyla contains 14 strains, 9 species, 4 genera, and 4 orders. The smallest phyla contians 8 strains, 4 species, 1 genus, and 1 order. Approximately 1/5 of each phyla was left out for each trial.

The results in Table IV show that 9 mers actually perform better for phyla recognition than 15 mers. In fact, NBC obtains similar performance to BLAST for this scenario. Interestingly, the back-off model for the 15 mers does almost as well as NBC in this case. It is surprising to note that although the phyla are sufficiently represented, the maximum performance is 84% recognition, and this demonstrates that not even BLAST results should be held as a ground-truth. Also, as the fragments are classified higher in taxonomy, the accuracy drops, and this is contrary to intuition that it would be easier to classify fragments at least in to the highest taxonomic level. Finally, it is of note that as in the case of partial training data, lower order N-mers (in this case 9 mers rather than 15 mers), perform better and thus show that particular N-mer order may need to be optimized for taxonomic level.

VI. Conclusions

In this paper, we demonstrate that composition-based classifiers, such as the NBC and back-off n-gram language models, are promising for the taxonomic binning problem of metagenomics. NBC obtains better results for lower taxonomic levels, and back-off modeling can improve performance for partial training data which arises in the case of incomplete or partially sequenced genomes. To simulate the situation where new genomes may need to be classified, cross validation tests demonstrate that reasonable accuracy for new genomic fragments can be obtained, as long as the respective taxonomic classes are well-represented. Interestingly, different Nmer sizes may do better for different levels (as seen with Phyla), and this parameter may need to be optimized. While we only focus on 500 bp in this paper for consistent comparison, we have previously evaluated classifier performance across several fragment sizes, and find that performance does not significantly decrease for 100 bp and 25 bp fragments. Therefore, statistical models can obtain reasonable results for shorter reads for full genomes and have an advantage over BLAST for partial knowledge in the case of incomplete genomes.

Acknowledgment

The authors would like to acknowledge the SRILM toolkit at SRI International for easing the implementation of the back-off models.

References

[1] J. Handelsman, *Committee on Metagenomics: Challenges and Functional Applications*. Washington, DC: National Academy Press, 2007.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[3] K. E. Wommack, J. Bhavsar, and J. Ravel, "Metagenomics: Read length matters," *Appl. Environ. Microbiol.*, vol. 74, no. 5, pp. 1453–1463, 2008.

[4] A. V. Lukashin and M. Borodovsky, "Genemark.hmm: New solutions for gene finding," *Nucleic Acids Res.*, vol. 26, no. 4, pp. 1107–1115, 1997.

[5] T. P. Curtis, W. T. Sloan, and J. W. Scannell, "Estimating prokaryotic diversity and its limits," *Proc. Nat. Acad. Sci. USA*, 2002.

[6] D. E. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "Megan analysis of metagenomic data," *Genome Res.*, 2007.

[7] 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 Res.*, vol. 11, no. 8, pp. 1404–1409, 2001.

[8] Q. Wang, G. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayes classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Appl. Environ. Microbiol.*, pp. 5261–5267, 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. Bioinf.*, Nov. 2008, Art. ID 205969.

[10] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, "Accurate phylogenetic classification of variable-length dna fragments," *Nature Methods*, vol. 4, pp. 63–72, 2007.

[11] L. B. Koski and G. B. Golding, "The closest blast hit is often not the nearest neighbor," *J. Mol. Evol.*, vol. 52, no. 6, pp. 540–542, 2001.

[12] Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith, "Environmental genome shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[13] S. L. Havre, B. J. Webb-Robertson, A. Shah, C. Posse, B. Gopalan, and F. J. Brockma, "Bioinformatic insights from metagenomics through visualization," in *Proc. Comput. Syst. Bioinf. Conf.*, 2005, pp. 341–350.

[14] S. Neph and M. Tompa, "Microfootprinter: A tool for phylogenetic footprinting in prokaryotic genomes," *Nucleic Acids Res.*, vol. 34, no. 366–368, 2006.

[15] M. Pignatelli, G. Aparicio, I. Blanquer, V. Hernández, A. Moya, and J. Tamames, "Metagenomics reveals our incomplete knowledge of global diversity," *Bioinformatics*, vol. 24, no. 18, pp. 2124–2125, 2008.

[16] A. Andersson, M. Lindberg, H. Jakobsson, F. Bäckhed, P. Nyrén, and L.Engstrand, "Comparative analysis of human gut microbiota by barcoded pyrosequencing," *PLoS ONE*, vol. 3, no. 7, 2008.

[17] M. L. Tress, D. Cozzetto, A. Tramontano, and A. Valencia, "An analysis of the Sargasso Sea resource and the consequences for database composition," *BMC Bioinf.*, vol. 7, no. 213, 2006.

[18] C. Manichanh, C. E. Chapple, L. Frangeul, K. Gloux, R. Guigo, and J. Dore, "A comparison of random sequence reads versus 16s rdna sequences for estimating the biodiversity of a metagenomic library," *Nucleic Acids Res.*, vol. 36, no. 16, pp. 5180–5188, 2008.

[19] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides, "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nature Methods*, vol. 4, pp. 495–500, 2007.

[20] S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: A genomic signature," *Trends Genet.*, vol. 11, pp. 283–290, 1995.

[21] S. Karlin, J. Mrazek, and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *J. Bacteriol.*, vol. 179, pp. 3899–3913, 1997.

[22] H. Nakashima, M. Ota, K. Nishikawa, and T. Ooi, "Genes from nine genomes are separated into their organisms in the dinucleotide composition space," *DNA Res.*, vol. 5, pp. 251–259, 1998.

[23] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences," *Mol. Biol. Evol.*, vol. 16, pp. 1391–1399, 1999.

[24] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T.Ikemura, "Informatics for unveiling hidden genome signatures," *Genome Res.*, vol. 13, pp. 693–702, 2003.

[25] D. T. Pride, , R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser, "Evolutionary implications of microbial genome tetranucleotide frequency biases," *Genome Res.*, vol. 13, pp. 145–158, 2003.

[26] 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 Bio.*, vol. 5, no. 163, 2004.

[27] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples," *DNA Res.*, vol. 12, pp. 281–290, 2005.

[28] B. Fertil, M. Massin, S. Lespinats, C. Devic, P. Dumee, and A. Giron, "Genstyle: Exploration and analysis of dna sequences with genomic signature," *Nucleic Acids Res.*, vol. 33, 2005.

[29] G. L. Rosen, "Examining coding structure and redundancy in DNA," *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 1, pp. 62–68, Jan./Feb. 2006, Special Issue on Communication Theory, Coding Theory, and Molecular Biology.

[30] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal processing in sequence analysis: Advances in eukaryotic gene prediction," *IEEE Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 310–321, 2008.

[31] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI-01Workshop Empirical Methods Artif. Intell.*, 2001, pp. 41–46.

[32] I. Zitouni and H.-K. J. Kuo, "Effectiveness of the backoff hierarchical class n-gram language models to model unseen events in speech recognition," in *Proc. IEEE Workshop Autom. Speech Recog. Understand.*, 2003, pp. 560–565.

[33] S. F. Chen and J. T. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. 34th Annu. Meet. Assoc. Comput. Linguistics*, 1996, pp. 310–318.

[34] W. Gerlach, S. Jünemann, F. Tille, A. Goesmann, and J. Stoye, "Webcarma: A web application for the functional and taxonomic classification of unassembled metagenomic reads," *BMC Bioinf.*, vol. 10, no. 430, 2009.

[35] H. M. Monzoorul, T. S. Ghosh, D. Komanduri, and S. S. Mande, "Sortitems: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences," *Bioinformatics*, vol. 25, no. 14, pp. 1722–1730, 2009.

[36] GenBank: National Center for Biotechnology Database [Online]. Available: http://www.ncbi.nlm.nih.gov 2009

[37] 454 Sequencing: Products and solutions [Online]. Available: http://www.454.com/products-solutions/system-features.asp#titanium

Gail L. Rosen (M'06) received the B.S., M.S., and Ph.D. degrees from the Georgia Institute of Technology, Atlanta.

She is an Assistant Professor of Electrical and Computer Engineering at Drexel University, Philadelphia, PA.

Prof. Rosen received a best student paper award in 2006 at the IEEE International Conference on Acoustics, Speech, and Signal Processing. In 2009, she was a recipient of the prestigious NSF CAREER award. She is currently the signal processing in education technical program chair and organizing the 2011 IEEE DSP/SPE workshop.

Steve Essinger (S'08) is working toward the Ph.D. degree in the Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA.

Mr. Essinger is a recipient of travel awards to share his research at the 2009 IEEE Biocomplexity Summer School in Istanbul, Turkey, the 2010 Pacific Symposium on Biocomputing in Hawaii, and the 2010 IEEE World Congress on Computational Intelligence in Barcelona, Spain.