Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2008, Article ID 205969, 12 pages
doi:10.1155/2008/205969

Research Article

Metagenome Fragment Classification Using *N*-Mer Frequency Profiles

Gail Rosen,[1] Elaine Garbarine,[1] Diamantino Caseiro,[2] Robi Polikar,[3] and Bahrad Sokhansanj[4]

[1] *Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA*
[2] *Spoken Language Systems Laboratory, INESC-ID, 1000 Lisbon, Portugal*
[3] *Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA*
[4] *School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA 19130, USA*

Correspondence should be addressed to Gail Rosen, gailr@ece.drexel.edu

Received 5 June 2008; Revised 19 September 2008; Accepted 30 September 2008

Recommended by Rita Casadio

A vast amount of microbial sequencing data is being generated through large-scale projects in ecology, agriculture, and human health. Efficient high-throughput methods are needed to analyze the mass amounts of metagenomic data, all DNA present in an environmental sample. A major obstacle in metagenomics is the inability to obtain accuracy using technology that yields short reads. We construct the unique *N*-mer frequency profiles of 635 microbial genomes publicly available as of February 2008. These profiles are used to train a naive Bayes classifier (NBC) that can be used to identify the genome of any fragment. We show that our method is comparable to BLAST for small 25 bp fragments but does not have the ambiguity of BLAST's tied top scores. We demonstrate that this approach is scalable to identify any fragment from hundreds of genomes. It also performs quite well at the strain, species, and genera levels and achieves strain resolution despite classifying ubiquitous genomic fragments (gene and nongene regions). Cross-validation analysis demonstrates that species-accuracy achieves 90% for highly-represented species containing an average of 8 strains. We demonstrate that such a tool can be used on the Sargasso Sea dataset, and our analysis shows that NBC can be further enhanced.

1. Introduction

While pattern recognition methods have been used in intron/exon identification [1], motif-finding [2], and microRNA prediction [3], these methods have not been applied to whole-genome identification and taxonomical relationships until recently. Now, there are a rapidly growing number and diversity of sequenced genomes across the evolutionary spectrum enabling a systematic study. This makes it possible to use these methods combined with biological insight to identify meaningful features and patterns to reveal relationships among DNA sequences that are not just limited to specific 16S rRNA genes but to any random fragment. A direct parallel between text classification and DNA classification can be made and seen in Figure 1. Until recently, bioinformatics approaches to metagenomics have been limited due to their lack of available data. Because of the lack of knowledge about genome diversity, most phylogenetic studies of metagenomic samples examine 16S ribosomal RNA genes for diversity [4]. This is because 16S rRNA sequences produce the fundamental protein needed for transcription, and therefore they are highly conserved across all species of life. Also, they contain insertion and deletion variation that makes their information content unique to various genera and species [4]. However, it has been shown that organisms that are identical or cluster tightly under 16S criterion cannot be concluded to share all or, in some cases, essential physiological similarities [5]. Thus, definition of species on this basis is not adequate for assessing the functional diversity of prokaryotic communities. In fact, it has been noted that the hot-spring microbes have ecologically important differences that have less than 1% 16S rRNA sequence divergence [5].

This has led scientists to consider new ways to identify the species/strain content of a clinical or environmental sample.

Unfortunately, in a less than ideal metagenomic sample, scientists do not always have the luxury of extracting these 16S genes. If blind methods existed to assess the taxonomical content of the sample from these random fragments, it would yield a high-throughput analysis especially when combined with short-read next-generation sequencing technology. Next-generation sequencing promises extremely high throughput, but at a price, it yields short reads. Currently, many metagenomic tools use BLAST as a first step to identify a sample's content [6–8]. But BLAST's [9] 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 100–200 bp reads for metagenomics data [7]. Huson et al. suggest that a "sweet" spot may exist around the 200 bp threshold for accuracy rates [6]. Wang et al. verify 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 test set [10]. Krause et al. suggest that with 80–120 bp reads, the superkingdom can be classified with 81% accuracy and the order can be classified with 64% accuracy [11]. Of course, most of these techniques use different sets of corpora and therefore the methods are difficult to compare although the main goal in identification is to gain as good of accuracy rates as possible. In general, researchers have deduced that fragments longer than 200 bp are needed in metagenomic applications. Yet, newer and faster sequencing technologies yield 20–35 bp reads in order to parallel the process, and scientists are questioning whether the technology is worth it due to the short reads [7]. Therefore, the holy grail of high throughput metagenomics is short-read DNA classification with reasonable accuracy.

In this paper, we construct the unique N-mer frequency profiles of 635 microbial genomes (including 470 unique species and 260 unique genera), publicly available as of February 2008. These profiles are used to train a naive Bayes classifier (NBC) that can be used to identify the genome from which a fragment may have been sequenced as part of a metagenomic data set. In Section 3, the methodology for naive Bayes classifier is presented, an example is given, the word frequency computations are discussed, and the methodology to obtain the confidence of our classifier validation is presented. In Section 4, NBC for the small (25 bp) fragment case is compared to the most widely used identification method, BLAST. We then assess the method's cross-validation performance (unseen-strains) for species-level classification. Finally, we test the NBC on the Sargasso Sea set and compare the results to MEGAN, a BLAST-based taxonomy presentation. The preliminary results show that an N-mer-count global perspective can yield a reasonable classification of metagenomic sequences that does not produce ambiguity. In Section 5, a discussion of the advantages and disadvantages of the method is shown. With further enhancements, this approach can yield a promising way to solve the strain resolution that BLAST has no chance to resolve with sequence identity scoring.

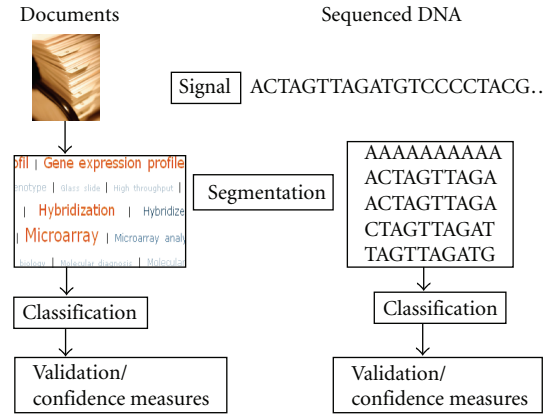

Figure 1: Comparison of text classification to the DNA classification problem.

2. Background

Sequence classification methods have traditionally aligned two sequences (usually homologous genes) to compare their similarity. The progress has been slow due to lack of demand, with the Needleman-Wunsch [12] algorithm introduced in 1970 and the Smith-Waterman algorithm [13] following over a decade later. Multiple-sequence alignment is an extremely important tool for phylogeny but did not have viable tools until the late 80s [14] due to the lack of sequenced genomes. Counting on BLAST to find homologous genes and sequence phylogenies is feasible, but it would be simpler to identify characteristic features unique to a group, such as a genotype signature representing all pathogenic E. coli, independent of encoded genes. In fact, most comparative techniques focus on the comparison of genes because they signify conserved regions and functions related to a phenotype [15]. Also, they conveniently ignore horizontal transfer which can insert locally anomalous characteristics [16]. In bacteria, this is especially true and phylogenetic footprinting uses gene homologs although there has been mounting evidence of use of noncoding RNAs [17]. Also, standard methods that ignore horizontal gene transfer cannot analyze the complete evolution of a microbial community or identify characteristic markers that may exist. Therefore, we seek a framework that represents the entire DNA in a sample without prior knowledge of the genes, promoters, and so forth. In the DNA sequences. We propose such approach that uses the naive Bayes classifier, which is able to identify significant features in a blind and high-throughput manner.

Existing methods to identify metagenomic sample content involves profiling clones with microarrays that identify previously unknown genes in environmental samples [18], subtractive hybridization to eliminate all sequences that hybridize with another environment, or subtractive hybridization to identify differentially expressed genes [19], and genomic signature tags [20]. The latter method is a way to extract particular 21-22 bp tag sequences that can be used to examine intraspecific genomic variation and, if genome information is available, provide immediate species identity.

Further, it pinpoints areas of a genome that might have undergone changes which add or delete restriction sites.

Our approach is to use N-mer frequencies, or words, of sequences as features to classify genomic fragments. Using DNA words as genomic features for discrimination and phylogenetic measures has previously been explored. For example, when faced with a contig that originated from an unknown, and never-to-be-recovered bacterial cell, Glöckner described how multivariate analyses of small-scale DNA architecture (e.g., comparing tetra-nucleotide usage) revealed a reasonable measure of fragment relatedness [21]. For tetra-nucleotides, it has even been demonstrated that their frequencies carry an innate but weak phylogenetic signal [22]. Other researchers have explored observing the patterns in oligonucleotide frequencies and unusual features [23–25]. A recent notable work is to construct a phylogenetic tree via variable-length segments and their frequency occurrence [26].

3. Methods

3.1. Naive Bayes Classifier

The term "classifier" is used in the sense of a statistical tool, trained using the full genomic data, to discriminate between "classes." Each "class" is a strain, species, family, and so forth, which depends on the particular class label definitions. In our work, we examine the cases where the classes are strains, species, and genera. The classification method will provide us with a way to predict class labels from fragment features, and the results are assessed for varying length of features and fragment sizes.

A naive Bayes classifier (NBC) is based on applying Bayes's theorem assuming that each feature in the classification is independent of each other. This strong independence assumption is the naivity of the algorithm, but the NBC has been shown to perform well in complex situations [27].

In this case, our features are composed of DNA words (N-mers). N-mers are N-length words of DNA that may or may not be overlapping. The foundation of our analysis is correlating the frequencies of these N-mers in a sequence to its overall identity. It is analogous to predicting the genre of a book from its word content. For example, a book about law is more likely to contain high frequencies of "law," "court," and "ruling" than this article which contains high frequencies of "genome," N-mer," and "fragment."

Let $\mathbf{w} = [w_1, w_2, w_3, \ldots w_K]^T$ be the feature vector, composed of a set of words (or N-mers) in an L-length fragment, $\mathbf{f}$. To label $\mathbf{w}$ in one of the M genome classes, $C_1, C_2, \ldots C_M$, the posterior probability of a particular class, C_i, given the feature vector, $\mathbf{w}$, is $P(C_i \mid \mathbf{w})$. The Bayes classifier chooses the predicted class, $\hat{C}$, with the largest posterior probability given that $\mathbf{w}$ is observed

$$\hat{C} = \underset{i}{\operatorname{argmax}}\, P(C_i \mid \mathbf{w}). \tag{1}$$

This expression guarantees minimum error across the whole space spanned by the K features in $\mathbf{w}$.

The posterior probability, $P(C_i \mid \mathbf{w})$, can be calculated by using the Bayes rule:

$$P(C_i \mid \mathbf{w}) = \frac{P(\mathbf{w} \mid C_i) \cdot P(C_i)}{P(\mathbf{w})}. \tag{2}$$

In other words, the probability, $P(C_i \mid \mathbf{w})$, of the genome class given the word features is equal to the probability, $P(\mathbf{w} \mid C_i)$, of the words given the class times the prior probability of observing that genome class, $P(C_i)$, divided by the unconditional probability of observing the words, $P(\mathbf{w})$, that compose a fragment, $\mathbf{f}$. The $P(\mathbf{w})$ is constant given a particular fragment.

The naive Bayes classifier assumes conditional independence between the N-mer features and calculates the class-conditional probability as a product of K individual probabilities:

$$P(\mathbf{w} \mid C_i) = \prod_{j=1}^{K} P(w_j \mid C_i), \tag{3}$$

where $K = L - (N - 1)$ is the number of overlapping N-mers in the fragment, $\mathbf{f}$.

The individual conditional probabilities, $P(w_j \mid C_i)$, are obtained by dividing the number of each fragment N-mer in the genome, $f_N(w_j \mid C_i)$, by the total number of N-mers in that genome $P(w_j \mid C_i) = f_N(w_j \mid C_i)/(|C_i|)$, where $|C_i|$ is the length of C_i.

In (2) , the prior probability of the genome, $P(C_i)$, is assumed to be in our hypothetical environmental sample. We make the assumption that our sample is uniform, or each genome is equally likely. In this case, our sample content is unknown, and in the absence of such prior knowledge, equal priors are typically used. With prior knowledge about the environment, a better estimate can be obtained. We also do not know the probability of obtaining a fragment with a set of words, $P(\mathbf{w})$, but this unconditional probability will be constant across the scoring function in (1), so it can be omitted. Therefore, we omit both $P(\mathbf{w})$ and $P(C_i)$ components in (1) and use the following scoring function for our work:

$$\hat{C} = \underset{i}{\operatorname{argmax}} \prod_{j=1}^{K} P(w_j \mid C_i). \tag{4}$$

As K increases, the score, $\prod_{j=1}^{K} P(w_j \mid C_i)$, can become very small and introduce precision errors into the computation. Due to numerical precision errors, we take the log probability to obtain our final scoring function

$$\hat{C} = \underset{i}{\operatorname{argmax}} \sum_{j=1}^{K} \log(P(w_j \mid C_i). \tag{5}$$

3.2. Calculation of N-mer—Frequencies, $f_N(\mathbf{Nmer} \mid C_i)$

Since we need to know the oligo words (or N-mers) as genomic features in the naive Bayes algorithm, an efficient

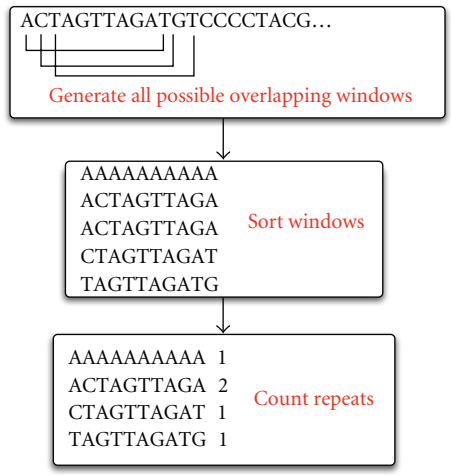

Figure 2: Example of the general algorithm for computing N-mer frequencies.

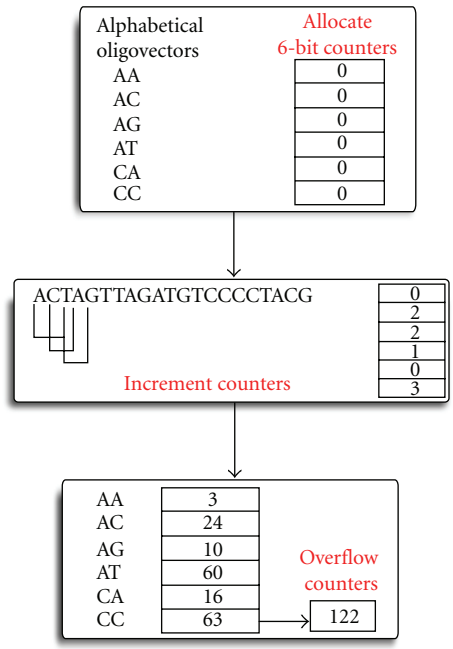

Figure 3: Example of the frequency counting algorithm optimized for shorter N-mers. 2-mers are chosen for simplicity and an arbitrary 6-bit counter is shown.

implementation is devised to compute the frequencies of all N-mers. We denote their frequencies as $f_N(\mathbf{w} \mid C_i)$ for each genome. We have implemented two such methods, one optimized for short words and the other optimized for long ones.

The first one is a general method that works for any sequence length, which generates all possible N-mers that overlap by $N - 1$ nucleotides. Once all possible N-mers are generated, they are sorted and then the cardinalities of recurring N-mers are computed. An illustration of this method is seen in Figure 2.

An optimized count mechanism is used when N is "small" (defined as $N = 20$ or less). By storing a finite bit counter for each N-mer in memory for $N \leq 20$, time and memory can be saved because the algorithm does not have to store each word in memory like the first method. We generate a list of the size of all possible 4^N combinations of N-mers. Each entry in the table stores an M-bit counter for each alphabetically-sorted N-mers. M is heuristically determined by examining the sequence length and the 4^N possible N-mers—if the sequence length is much less than 4^N, then M is low, otherwise, M is increased accordingly. Then incrementing down the sequence with an N-length window, the counter that corresponds to each N-mer in the table is incremented. If a counter overflows, another M-bit counter is mapped from the first counter to account for the extra counts. While this slows the algorithm down, it is unlikely to occur. This phenomenon is related to Zipf's law [28] which is a power law that states that the frequency of any word is inversely proportional to its rank, $f(k) = 1/k$, where k is the rank of the word, in the frequency table. Therefore, only a few N-mers will have high frequencies that need additional counters. The algorithm is summarized in Figure 3.

To further illustrate Zipf's law, we illustrate the 12-mer frequencies of 3 different strains of E. coli in Figure 4. A trend close to Zipf's law (the inverse rank-frequency relationship). Zipf's law curve can be modeled with an exponent as

$$\begin{aligned} &\text{Zipf\_freq}(N\text{mer\_rank}, s = 1/4, N = 12) \\ &= 1000 * \frac{((1/N\text{mer\_rank})^{1/4})}{\sum_{n=1}^{N=12} n^{1/4}}, \end{aligned} \tag{6}$$

where s is the exponent order, N is N-mer length, and the N-mer_rank is the order of the frequency rank on the x-axis. We can see that the log-log E. coli curves tend to follow this law.

A comparison of the algorithm run times for $N = 10$ and $N = 100$ for various genomes can be seen in Table 1. The optimized run times are similar to those seen in other computational methods for frequencies [29], but other methods rarely calculate N-mers larger than 20-mers [30]. We can compute any size, and one of the parameters we will be looking for is the optimum N-mer size for separability among the data sets.

3.3. Confidence Intervals for Accuracy Calculations

To validate our model, we choose 100 random fragments from each training-set genome, totaling 63 500 fragments tested. Once we receive the result of the scoring algorithm, the genome that scores the highest is marked as correct or incorrect using prior knowledge of the true genome. This enables us to average the binomial distribution of correct(1)/incorrect(0) labels to produce an average accuracy per genome (as seen in Figure 5). The confidence of our average accuracy over 100 random fragments can be computed by using the formula for computing the confidence interval

Table 1: Comparison of run times for $N = 10$ (general and optimized methods) and $N = 100$ (general method only). Simulations are run on one 2×2.0 GHz Intel dual-core Xeon, 2 GB RAM, and 2×80 GB HD. Many temporary files are needed for the sort process and are saved to a 3 TB RAID drive which is connected through a cluster-head terminal machine.

Name	Genome size (Mbp)	Optimized alg. $N = 10$ (min)	General alg. $N = 10$ (min)	General alg. $N = 100$ (min)
Human	3142.05	2.4	212.7	1877.2
Zebrafish	1578.26	1.1	106.25	506.5
Fruitfly	135.25	0.1	6.5	26.4
E. coli K12	4.7	0.01	0.18	0.63

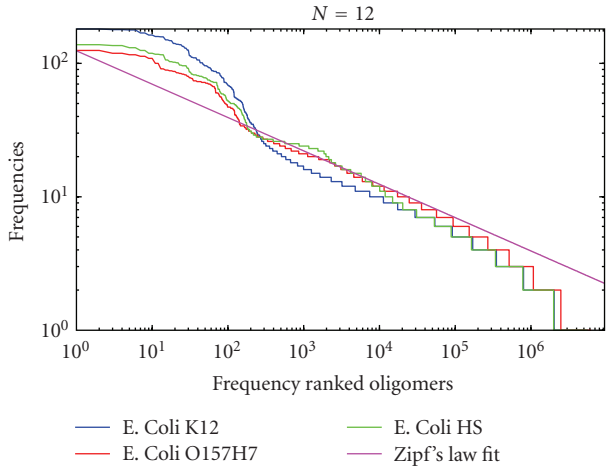

Figure 4: A log-log plot of the N-mer frequences versus N-mers in ranked order for various E. coli strains (K12 is the commensal strain, O157H7 is highly pathogenic, and HS is the commensal isolate from the human gastrointestinal tract). E. coli has a characteristic curve for all strains in this domain. This curvature is then compared to Zipf's law which states that N-mer frequency is directly related to inverse rank order. While E. coli generally obeys this law, the curvature deviation from the straight line shows that higher ranking of words has higher normal frequency.

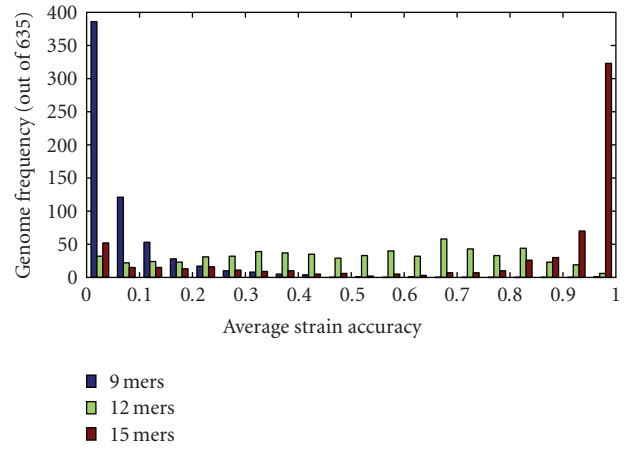

Figure 5: A histogram of the average strain accuracy for 25 bp fragments for $N = 9, 12, 15$. The overall strain accuracy of the $N = 9$ mers is 3.5% and is apparent from the *blue* distribution of average accuracy (averaged over 100 fragments) per strain. For $N = 12$ mers, the overall average accuracy is 49.3% and this is reflected in *green* average strain accuracies; this shows that some strains begin to classify well with $N = 12$ mers while others do not. It is interesting to note that $N = 12$ mers do not yield many strains with over 95% accuracy. For $N = 15$ mers, while the overall accuracy is 75.7%, we can see that over 50% of the strains have over 95% accuracy. We can conclude that most strains perform well with the NBC using $N = 15$ mers, but some strains have poor accuracy and cannot be resolved.

for a binomial distribution. The binomial distribution is approximated by a normal distribution. It has been shown that for over 30 trials, a binomial distribution obeys the normal distribution due to the central limit theorem. The true accuracy with its confidence interval is defined as

$$\text{True Accuracy} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{7}$$

where $\hat{p}$ is the estimated average accuracy, $z_{\alpha/2}$ is the critical value corresponding to the $\alpha/2$ percentile of the standard normal distribution, n is the sample size, and $\sqrt{\hat{p}(1-\hat{p})/n}$ is the standard deviation of the binomial distribution.

4. Results

The naive Bayes classification is performed on all completed microbial sequences in the NCBI Genbank as of February 2008, which totaled 635 distinct microbial strains. The 635 microbes belong to 470 distinct species and 260 distinct genera in this data set. 404 strains are the sole member of their species class while 171 strains are the sole member of their genus in the data set. This shows that some knowledge will be lacking when it comes to species- and genus-class diversity. While 66 species contain more than one strain, 89 genera contain more than one strain. The microbial strains genome lengths range from 160 K(bp) for Candidatus Carsonella to 13 Mil(bp) for Sorangium Cellulosum.

4.1. The Naive Bayes Classification of the 635-Strain Genome Data Set

4.1.1. Matching to the Nearest Strain

To evaluate the performance of our classifier's ability to classify a given fragment in our database, we test over varying N and fragment lengths used in the scoring function (5). These two parameters are varied and the scoring function

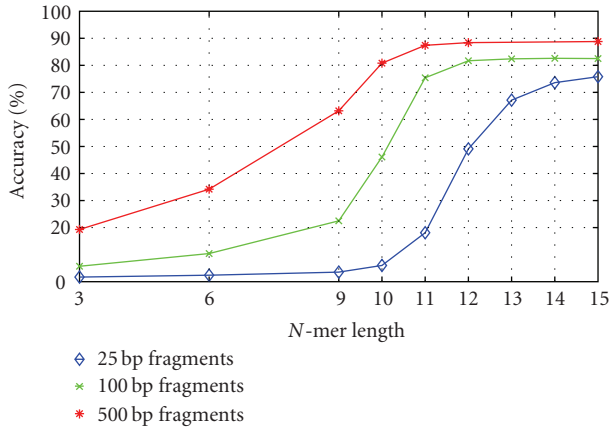

Figure 6: The accuracy of the naive Bayes classifier versus N-mer length versus fragment length for strain classifications for the 635 completed microbial genomes. This graph clearly shows that accuracy improves when the longer N-mers are used in the scoring function. As expected, 500 bp fragments performed the best, reaching 88.8% accuracy in strains and 82.5% for 100 bp fragments. The 25 bp fragments surprisingly increased performance when using 15 mers, yielding 75.8%. There is a jump in accuracy at around the $9 < N < 12$ range which provides insight into the order needed for classification.

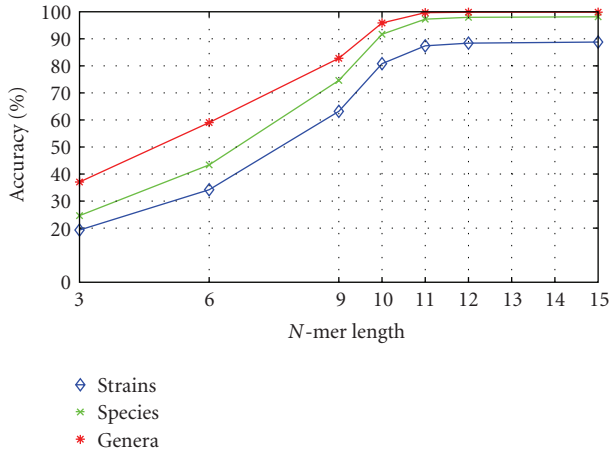

Figure 7: A comparison of the strain, species, and genus classification for 500 bp fragments on the training data. Above $N = 11$, the classification accuracy cannot be improved with strain accuracy being around 88%, species accuracy around 97%, and genus accuracy around 99.7%. It is interesting to note that the increase of accuracy from $N = 6$ to $N = 9$ is dramatic.

is calculated for all 635 microbial genomes. The fragment length is chosen as 500 bp, 100 bp, and 25 bp to simulate long and short reads. The N-mer lengths is varied for 3-, 6-, 9-, 10-, 11-, 12-, 13-, 14-, and, 15-mers to test performance improvement over these lengths.

To validate our model, we choose 100 random fragments from each training-set genome, totaling 63 500 fragments tested. The 100 fragments are averaged to obtain a strain accuracy per genome. Figure 5 demonstrates how increasing N changes the individual strain accuracy rates. For $N = 9$, most strains have a very poor 0–5% classification rate, and interestingly various strains have performance across the board with 12 mers.

For a 95% confidence interval, the critical value is 1.96. Therefore, for the strains that have 50% average accuracy, we are 95% confident that they are between 40% and 60% using (7), with $1 - \alpha/2 = 0.95$, $z_{\alpha/2} = 1.96$, $n = 100$, and $\hat{p} = 0.5\,(50\%)$. The ± 9.8 interval is an upper bound. The interval has a quadratic drop-off as the binomial estimates tend towards 0% or 100%.

The accuracy per genome is then averaged and produces a composite "overall" accuracy for the genome strains in our data set. This overall accuracy is computed for each fragment and N-mer length. The accuracy of each strain classification can be seen in Figure 6. The strain average accuracies are then averaged together to form an overall average of the 63,500 fragments. The upper bound on the confidence interval for the overall accuracy is $\pm 0.4\%$. To calculate this bound, the same parameters for (7) are used except $n = 100 * 635$. The accuracy seems to level off for 12 mers for 500 bp and 100 bp fragments while 25 bp fragments do the best with 15-mer calculations (and probably beyond).

Because Sandberg et al. [31] never ventured past $N = 9$ for the N-mer size, the result of a jump in performance is never discovered. Again, we believe this is due to the fact that N-mer sparsity begins at around $N = 9$ because that is when the number of possible combinations surpasses the lengths of the microbial genomes.

4.1.2. Classification to Higher-Level Classes: Examples of Species and Genera

One of the reasons for misclassification of fragments is the sequence overlap between different strains within the same species, and possibly within species belonging to the same genus. In particular, for the case of strains, different strains may be characterized by the loss of genes, addition of genes, or possibly the addition of extrachromosomal genes through the addition of a plasmid. In those cases, there may be only random single base changes in the remainder of their genome, if any. Thus, fragments taken from them using our procedure described above may identically appear in multiple organism sequences. Moreover, if only one base differs, the N-mer frequency profiles may be sufficiently similar for the NBC to misclassify the fragment. To study this issue, we consider the performance of fragment identification by pooling the results based on species and genus identity. In doing so, we define genus, species, and strain identifies based on the conventional NCBI taxonomy. For example, Yersinia pestis CO-92 and Yersinia pestis KIM-9 are two strains of the same species; Yersinia pestis and Yersinia pseudotuberculosis are two species of the same genus.

Therefore, the classification with pooling is performed for "species" and "genus" classes instead of individual strain classes. In other words, as long as the genome is classified to a genome within the same species or genera, it is considered correct for that classification. A comparison of the strains, species, and genera classifications can be seen for 500 bp

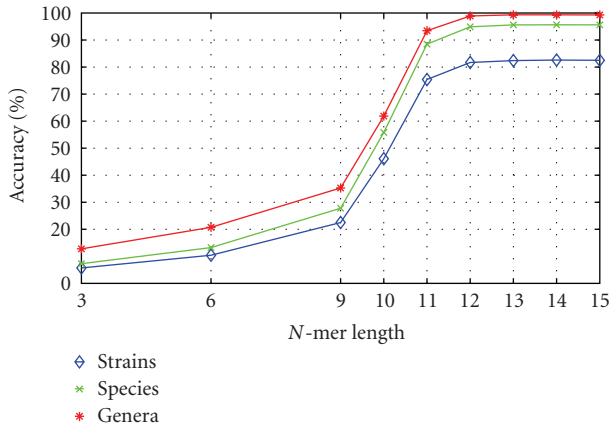

Figure 8: A comparison of the strain, species, and genus classification for 100 bp fragments on the training data. Above $N = 12$, the classification accuracy cannot be improved with strain accuracy being around 82%, species accuracy around 95%, and genus accuracy around 99%. It is interesting to note that the increase of accuracy from $N = 9$ to $N = 12$ is dramatic.

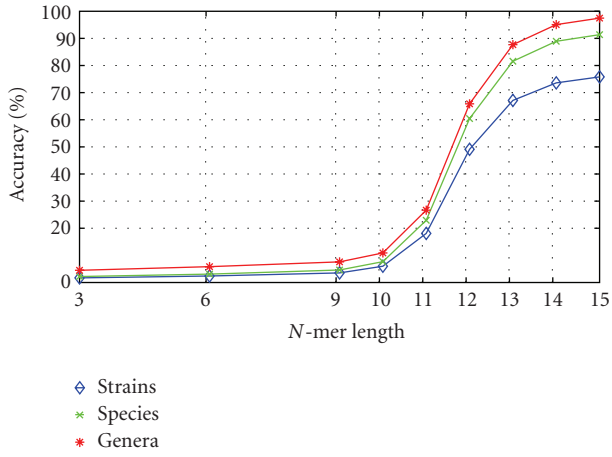

Figure 9: A comparison of the strain, species, and genus classification for 25 bp fragments on the training data. In this case, the level-off point has not yet been reached, although there is a dramatic increase from $N = 10$ to $N = 13$. By $N = 15$, strain accuracy reaches 76%, species accuracy around 91%, and genus accuracy around 97%. It is interesting to note that the increase of accuracy from $N = 9$ to $N = 12$ is dramatic.

fragments in Figure 7, 100 bp fragments in Figure 8, and 25 bp fragments in Figure 9, respectively. The accuracy for genera is better as expected but follows the general trends for increasing N. For genera, the accuracy levels off at 99.8% for 500 bp, 99.3% for 100 bp, and 97.5% 25 bp fragments, respectively, and shows the potential power of the method.

4.2. Comparison Against BLAST Results for 25 bp Fragments for the 635 Genome Data Set

BLAST [9] is expected to do very well for long fragment lengths. In this section, a direct comparison of how the naive Bayes classifier compares to BLAST is shown. It has been reported that BLAST does not yield sufficient results for 25 bp because of ambiguity [6]. It looks for local and global alignments of sequences to score a particular fragment's identity. But there are a slew of parameters controlling the significance of this score, and when a scientist is looking for the closest matching genome to a particular sequence, we will see that in some rare cases, it is incorrect or does not provide an answer. In many cases, it provides too many of the same top scores, yielding ambiguous results. To conduct the comparison, we took all 63 500 fragments (100 fragments per database genome), and BLASTed them against our 635 genome database. The results were compared to the $N = 15$ NBC case.

BLAST finds the significance of alignment via an E-value, which is the number of highest scoring pairs (HSPs) expected by chance. Therefore, the higher the E-value, the lower the significance. In our tests, we desired BLAST to give all tied HSPs despite the score; therefore, we desire an infinite E-value. But too many hits were produced by the local BLAST program for an E-value above 3000 causing memory errors. This limited us to use an E-value of 3000, but because this means that 3000 HSPs may occur by chance, it is a reasonable E-value to use in BLAST since it is likely to cause BLAST to produce insignificant scores and hits. More on the E-value and BLAST is discussed in [9].

Despite the high E-value, 287 or 0.5% of the fragments scored "No Hit" which can be interpreted that all matches in the database were insignificant. One must remember that all fragments BLASTed are from the database, so this is an unexpected result from BLAST. Many of these fragments are only found one time in one genome across the database. Because of this uniqueness, NBC is able to classify the correct genome that produced it, 71% of the time. There is also the issue of multiple top-scoring hits because BLAST only gave 66% of the fragments a unique top-scoring hit and is correct for all of them. Comparably, the naive Bayes classifier classifies 99% of those as well. Out of the multiple top-scoring hits, BLAST completely missed 13 of them, meaning that there are multiple top-hits but the correct one is not in that list. The remaining ones have the correct classification embedded in a list that could range from 2 to 200 top-scoring hits. If one "flips a coin" whenever multiple ambiguous choices occur for a top hit, the correct genome can be guessed 29% of the time overall. The NBC chooses the correct genome 31% of the time out of this set. A comparison between the (a) unique BLAST hits, (b) multiple top-hits, and (c) no hits cases can be seen in Table 2.

To summarize, BLAST is able to find the correct genome (even if ambiguous) in 63200 of the reads but can only resolve 41641 uniquely. With the top hits and flipping a coin for the ambiguous multihits, BLAST would get 47889 (75.4%) correct. The NBC scored 48118 (75.8%) correct which is shown in Figure 6. If N is increased, the NBC can potentially get better strain resolution.

The primary issue with BLAST concerning small fragments is that the probability of a unique score becomes lower. Due to NBC's spatial independence, the algorithm can classify correctly 31% of the fragments that are ambiguous

Table 2: 63 500 25 bp fragments, 100 from each genome, are BLASTed and compared to the $N = 15$ NBC. BLAST gives 66% of them unique top-scoring hits, where all of them were correct. Almost 34% of the reads have ambiguous top-scoring hits, meaning that there are multiple organisms that have top scores and E-values. Also, even though the exact string or complement exist in the database, 287 fragments receive no hit from BLAST with an E-value of 3000. NBC is able to correctly identify 71% of those. Being that the multiple top-scoring genomes can be randomly chosen as a top hit, we can compare directly, how often BLAST would get the genome correct compared to the NBC. Taking this and the single top hits into consideration, NBC scored 48118 (75.8%) fragments correct while BLAST matched 47889 (75.4%) fragments correct.

63 500 fragments		
BLAST category	Interpretation of BLAST results	NBC's results for the BLAST category
No. of reads that had *Unique* Top-scoring hits in BLAST	No. that BLAST got correct	No. that NBC got correct
41641	41641	41211
No. of reads that had *Multiple* Top-scoring hits in BLAST	BLAST hits for reads where the *multiple* top-scoring list contained the correct one/no. of *unique* top-hits BLAST would get by chance from ambiguous hits	No. that NBC got a correct, *Unique* Top-hit
21572	21559/6248	6702
Reads that had *No* hits in BLAST (E-value of 3000)	Could not be assigned in BLAST	No. that NBC got correct
287	0	205

in BLAST. The NBC algorithm can be extended to exploit its top multiple scores to obtain better accuracy. With an intelligent examination of the scores, it may be able to get better performance than just predicting the genome with the maximum score. While BLAST gives the same score to multiple organisms, NBC ranks the organisms by score. Surprisingly, NBC never has a tied score for any of the 63 500 fragments. This means that each fragment combination yields a unique probability for the top-ranking genome. This outcome opens up further work in how to exploit the histogram of the genomes' NB prior and posterior probability scores to gain better accuracy. In any case, for 25 bp fragments, it is shown that NBC performs at least as well as BLAST with no augmentations.

4.3. Cross-Validation Performance of NBC Versus BLAST (Using a 9-Species Subset)

In order to fully assess the performance of both methods, we propose to leave some of the data set out for testing. When carefully partitioning the data so that each test set contains a unique subset, this is known as cross-validation and particularly K-fold cross-validation for K partitions. A major obstacle in conducting cross-validation for our data set is choosing the K. We treat each genome as a single strain, training only on full genomes, and do not train on parts of genomes. Thus, for cross-validation, we wish to train on a subset of the example strains in a species and then classify test-strain fragments to the closest training-set species. If strains classify to a strain within their same species, it is marked as correct. As reported before, 66 species contain more than one strain, and many classes contain 2 example strains.

Cross-validation involves K partitions. In many cases, the rule of thumb for cross-validation is to use 10 training/test sets [32]. One of the many reasons for $K = 10$ is to uniformly train on 90% of the data at a time in order to obtain a better estimate. This poses a difficulty for our sparse data set because only 4 species have 10 or more strains. 9 species have 5 or more example strains, and therefore we determine 5-fold cross-validation to be sufficient for this small data set. The 9 species classes, containing 77 strains, are selected. For each 5-fold cross-validation set, about 62 strains are trained on while about 15 strains are left out (approximately 1/5 of each class).

4.3.1. The NBC species cross-validation results

In Figure 10, the performance of the classifier using 5-fold cross-validation is shown. Each fragment size can be classified to over 90% accuracy. An interesting note is that while the maximum performance is for 15 mers for 500 bp and 100 bp fragments, 14 mers yield the maximum performance for the 25 bp fragments. The accuracy and standard deviation, respectively, for each fragment size is $97.3 \pm 1.0\%$ for 500 bp fragments, $95.3\% \pm 1.3\%$ for 100 bp fragments, and $90.2\% \pm 1.2\%$ for 25 bp fragments.

4.3.2. Comparison Against 25 bp BLAST Cross-Validation

A BLAST database is built using each ~60-strain training set, and the ~15 strains are left out at a time to form a validation set. 25 bp validation fragments are BLASTed, and if the top matches contain only those strains belonging to fragment's species, the results are considered correct. We generate a list

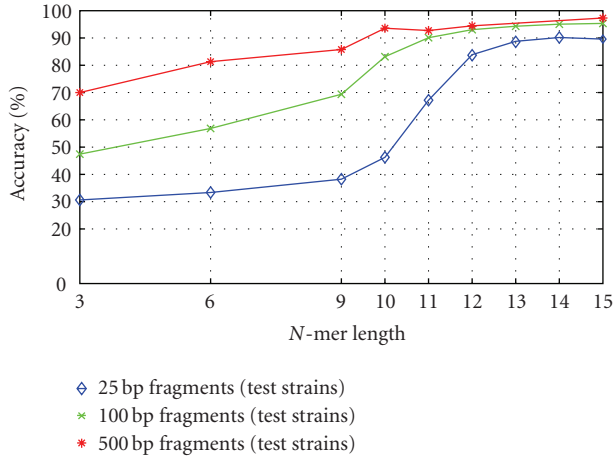

FIGURE 10: 5-fold cross-validation performance for species-accuracy on 9 species classes (77 strains), training on ~62 strains and testing on ~15 strains at a time. The maximum performance for each fragment size is 97.3% ± 1.0% for 500 bp, 95.3% ± 1.3% for 100 bp, and 90.2% ± 1.2% for 25 bp fragments, demonstrating that the NBC has good classification species accuracy for never seen before strains.

of the size of all possible 4^N combinations of N-mers. On the other hand, if the top-scoring correct species is tied with an incorrect species, the classification is marked as incorrect. Both no hits cases and purely incorrect hits are marked as incorrect as well. 25 bp fragments are scored with an accuracy of 89.2% ± 1.9%, which is comparable performance to NBC's 90.2% ± 1.2%. There is also slightly less variance for NBC showing that it has the potential to be more stable classifier for species classification. As shown in the strain-level BLAST comparison, NBC performs at least as well as BLAST with no augmentations, and this holds true for species-level accuracy using never seen before strains in cross-validation.

4.4. 10 K Reads from the Sargasso Sea Set

The Sargasso Sea data set was published in 2004 [33] by Venter et al. Four geographic sampling sites' microbial cells were shotgun sequenced yielding ~1.66 million reads of average length 818 bp. For our analysis, we selected the first 10 000 reads from Sample 1 for analysis which Huson et al. have also analyzed in their MEGAN analysis [6]. In this section, we wish to show how our classifier can be used to analyze this data and compare it to Huson's results which uses BLAST and the NCBI taxonomy database. In metagenomic applications, scientists seek the overall taxonomic content, or the evolutionary relationship of all the microorganisms in the sample. The first step is to identify different strains, or just to identify what phyla/genera an organism is from. In our results, we do an exact strain-matching test on the set (where species/genera can be inferred, such as the example of Yersinia pestis/pseudotuberculosis in Section 4.1.2. We evaluate the 10 K fragments through our classifier for $N = 9$ mers and $N = 15$ mers to see how different N performed for strain recognition to our database and compared it with MEGAN's BLAST-based results.

A comparison of the results can be seen in Table 3. Venter's analysis of the Burkholderia genera in the Sargasso Sea sample 1 is around 38.5%. With the exact same first 10 K reads of sample 1, MEGAN found Burkholderia to be 25.2% of the sample. In our top 10 analysis, we find Burkholderia is 21% for 9 mers and 24.6% for 15 mers). Venter et al. estimated 14.4% for the Shewanella genera in Sample 1. MEGAN specifically finds 17.4% In our top 10 analysis, Shewanella composes 11.4% with 9 mers, and 17.4% with 15 mers.

As explained above, the NBC is able to find the classification rate comparable to BLAST methods of a genera within the top 10 content of the sample for 15-mer analysis. This leads us to a question: do higher N-mer models overfit the unknown data? for example, Burkholderia 383 is shown to have a substantially greater percentage in the sample in the 15-mer set (20.4%) over the 9-mer set (6.93%). The same phenomenon occurs with Shewanella ANA-3.

5. Discussion

While the naive Bayes classifier works well on our training data set, is comparable to BLAST, and is able to classify some genomes in an environmental sample, it needs further refinement. For example, in Figure 5, one can see that the 9 mers have consistently poor accuracy for 25 bp fragments, but for 15 mers, the accuracy performs well. Although, one can see that the 15 mer histogram is approaching a binomial distribution, because most strains perform near 100% but some strains never able to resolve and perform poorly near 10%. These fragments should be investigated further.

We compare our work to that of Sandberg et al. [31]. Sandberg used parts of 28 eubacterial and archaeal genomes to train a naive Bayes classifier that would classify segments into 25 species classes. The performance worked quite well and obtained >85% accuracy for more fragment sizes of more than 400 bp, and a promising result is that 35 bp reached 35% accuracy. An unintuitive result in the work of Sandberg et al. was that there seemed to be an upper threshold on how much the N-mer (motif in the paper's terminology) size could help in the naive Bayes computation. In our computations, we show that for a large data set, the optimal N-mer size increases as the length of the fragment decreases. Also, the Nmer length needed is larger than what Sandberg et al. needed due to the larger size of our database. On the training data, we show we can achieve 89% strain accuracy and 99.8% genus accuracy for 500 bp fragments. And a great result is that the NBC can resolve training data 25 bp fragments with 76% accuracy for strains and 98% for genera. Training on multistrain species, we show that this method can obtain over 90% for all fragment sizes on unseen strains, and we obtain comparable results to BLAST. In fact, there has been little analysis on the performance of BLAST for general organism recognition, and this paper opens the opportunity for further study of BLAST to metagenomic applications. The results demonstrate great promise for use of this classifier in metagenomic applications.

Table 3: Comparison of the top 10 reads from the naive Bayes analysis of the Sargasso Sea set for 9 mers and 15 mers and a side-by-side comparison with MEGAN results. There are 7 common strains between the naive Bayes sets substantiating their presence in the sample. Not all NBC "best matches" are found in MEGAN (indicated by "None"), and this can be due to "no hits" or to not having that strain in the database. An interesting NBC find is that *Trichodesmium erythraeum* has been found to compose 0.6% of the sample. It has been extensively found in the Sargasso Sea, but no prior methods show this presence in the Sargasso Sea data set.

9 mers			15 mers		
High-strain content in sample (genome size of both sides)	No. of reads	No. of MEGAN reads	High-strain content in sample	No. of reads	No. of MEGAN reads
Burkholderia 383 (9.3 M)	693	514	Burkholderia 383 (9.3 M)	2044	514
Burkholderia Cenocepacia AU 1054 (14.6 M)	684	13	Clostridium Beijerinckii NCIMB 8052 (12 M)	1698	2
Clostridium beijerinckii NCIMB 8052 (12 M)	623	2	Shewanella ANA-3 (10.3 M)	989	186
Shewanella ANA-3 (10.3 M)	562	186	Trichodesmium erythraeum IMS101 (15.6 M)	584	2
Trichodesmium erythraeum IMS101 (15.6 M)	533	2	Flavobacterium johnsoniae UW101 (12.2 M)	481	10
Burholderia xenovorans LB400 (19.6 M)	404	None	Sorangium cellulosum So Ce 56 (26 M)	309	None
Shewanella MR-4 (9.4 M)	329	14	Shewanella oneidensis MR-1 (10.4 M)	297	78
Burholderia ambifaria/cepacia AMMD (15 M)	265	91	Shewanella MR-4 (9.4 M)	245	14
Alkaliphilius metalliredigens QYMF (9.8 M)	261	None	Burkholderia cenocepacia HI2424 (15.5 M)	219	102
Shewanella MR-7 (9.6 M)	250	26	Shewanella MR-7 (9.6 M)	206	26
Acidobacteria bacterium Ellin345 (11.6 M)	187	None	Burkholderia xenovorans LB400 (19.6 M)	198	None

Our results are comparable to Huson et al.'s work [6] for metagenomic samples, and for comparison, Table 4 lists the top 10 of MEGAN and our method's side-by-side comparison. There are a few surprising differences. While MEGAN finds *Candidatus pelagibacter* as the second most abundant organism, the NBC finds it as a less common sequence. It has been shown in the literature to be a prolific organism and common in the Sargasso Sea [34]. However, about 20% of the reads that gave *Candidatus pelagibacter* in MEGAN correspond to *Trichodesmium erythraeum* in the naive Bayes method. While 20%, 50% (9 mers, 15 mers) of the *pelagibacter* reads end up being *Clostridium beijerinckii*. In addition, a surprising difference from MEGAN is that more reads, 533/584 ($N = 9/12$), are assigned to *Trichodesmium erythraeum* IMS101. This organism has been found in the Sargasso sea through gene expression studies [35], but MEGAN only shows 3 reads for this organism. The naive Bayes classifier finds this organism consistently in the top 10 organisms present. The NBC could signal some of these organisms that BLAST-like methods do not find, but further analysis should be conducted.

The differences of our Sargasso sea findings from the BLAST findings cause concern, especially since it has been shown that *Candidatus pelagibacter* is arguably the most abundant prokaryote in the ocean [36]. With further analysis, we find that the NBC gives preference to longer genomes

Table 4: MEGAN's top-ten strains for the Sargasso Sea dataset, their respective reads, and comparison to the NBC 9 mer and 15 mer methods. N/A means the strain is not in our training set (it is unfinished so it cannot be found. Burkholderia and Shewanella which were also found by Venter et al. [33] also have high matches in the NBC. The NBC's detection of Candidatus Pelagibacter drastically changes from $N = 9$ to $N = 15$.

High strain content in sample (genome size—bothsides)	MEGAN # of Reads	NBC 9 mer # of reads	NBC 15 mer # of reads
Burkholderia 383 (9.3 M)	514	693	2044
Candidatus Pelagibacter ubique HTCC1062 (2.6 M)	323	13	111
Shewanella ANA-3 (10.3 M)	186	484	989
Procholorococcus marinus MIT 9312 (3.4 M)	125	28	24
Psychroflexus torquis ATCC 700755 (8.6 M)	119	N/A	N/A
Burkholderia cenoecapacia HI2424 (15.52 M)	102	106	219
Burholderia vietanamiensis G4 (16.8 M)	101	93	92
Burkholderia ambifaria/cepacia AMMD (15.06 M)	91	265	127
Shewanella oneidensis MR-1 (10.32 M)	78	79	297
Synechoccus sp. WH8102 (4.86 M)	75	68	82

for long fragments and high N. Comparing Tables 4 and 3, we can see that *pelagibacter* is the 2nd most common taxa found from the reads in BLAST, but the NBC does not find it in its top 10. Instead, genomes that have 10–14 million bases show up high on the list. For example, when $N = 15$, there are 1 billion possible words, but all genome sizes are between 320 K and 26 million nucleotides (both sides). With those genome sizes, the 15 mers that exist in them are usually singletons (one occurrence). Therefore, a long genome that is probabilistically more likely to have a 15 mers from a fragment, is more likely to get a "hit" and have a higher score than a small genome. This is especially the case when a fragment is not from a genome in the database. Therefore, the scoring vector needs more intelligence for classifying unknown fragments in order to not penalize smaller genomes.

The analysis of n-gram models may yield insight into ways to distribute the probability mass in a more effective manner. Overall, while the accuracy is quite good for fragments existing in our database, the method will need to be improved for unseen species and even genera, and how to assess if the fragment is from an unseen genome.

6. Conclusion

Our approach differs from sequence alignment-based methods because word composition of the sequences is taken into account instead of string matching and alignment. Counting the word-frequencies present in a genome represents global features of the genome as opposed to the local similarities and differences scored by alignment-based methods. More than ever, a method is needed to classify all fragments resulting from high-throughput sequencing technology. It is shown that a global classifier that utilizes N-mer frequencies is able to achieve good results (90% for cross-validation species-resolution accuracy) and has great potential to be used in metagenomic applications. In our work, we demonstrate that this approach is viable for any fragment and is scalable to hundreds of genomes. It also performs well for strain and higher-class identifications. It also has the advantage of resolution despite classifying ubiquitous genomic fragments.

In conclusion, global N-mer frequency-based profiling based on NBC is a general method for classifying organisms and their genomic content. It can be used for a broad range of applications for analyzing all data from a metagenomic set that will be generated through large-scale projects in ecology, agriculture, and human health. Given that the Human Genome Project is still at an early stage, these new kinds of massive data sets will require innovative informatics approaches for their analysis and translating them into useful knowledge.

Acknowledgment

The authors would like to thank Christopher Pearson for the N-mer counting code.

References

[1] H.-M. Müller and S. E. Koonin, "Vector space classification of DNA sequences," *Journal of Theoretical Biology*, vol. 223, no. 2, pp. 161–169, 2003.

[2] G. Yeo and C. B. Burge, "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals," in *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB '03)*, pp. 322–331, Berlin, Germany, April 2003.

[3] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe, "Naïve Bayes for microRNA target predictions—machine learning for microRNA targets," *Bioinformatics*, vol. 23, no. 22, pp. 2987–2992, 2007.

[4] R. S. Gupta and E. Griffiths, "Critical issues in bacterial phylogeny," *Theoretical Population Biology*, vol. 61, no. 4, pp. 423–434, 2002.

[5] B. B. Ward, "How many species of prokaryotes are there?" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 16, pp. 10234–10236, 2002.

[6] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377–386, 2007.

[7] K. E. Wommack, J. Bhavsar, and J. Ravel, "Metagenomics: read length matters," *Applied and Environmental Microbiology*, vol. 74, no. 5, pp. 1453–1463, 2008.

[8] 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 Research*, vol. 36, no. 16, pp. 5180–5188, 2008.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[10] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Applied and Environmental Microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[11] L. Krause, N. N. Diaz, A. Goesmann, et al., "Phylogenetic classification of short environmental DNA fragments," *Nucleic Acids Research*, vol. 36, no. 7, pp. 2230–2239, 2008.

[12] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[13] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[14] D. G. Higgins and P. M. Sharp, "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer," *Gene*, vol. 73, no. 1, pp. 237–244, 1988.

[15] S. Abby and V. Daubin, "Comparative genomics and the evolution of prokaryotes," *Trends in Microbiology*, vol. 15, no. 3, pp. 135–141, 2007.

[16] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: quantification and classification," *Annual Review of Microbiology*, vol. 55, pp. 709–742, 2001.

[17] S. Neph and M. Tompa, "MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes," *Nucleic Acids Research*, vol. 34, pp. W366–W368, 2006.

[18] J. L. Sebat, F. S. Colwell, and R. L. Crawford, "Metagenomic profiling: microarray analysis of an environmental genomic library," *Applied and Environmental Microbiology*, vol. 69, no. 8, pp. 4927–4934, 2003.

[19] E. A. Galbraith, D. A. Antonopoulos, and B. A. White, "Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model," *Environmental Microbiology*, vol. 6, no. 9, pp. 928–937, 2004.

[20] J. J. Dunn, S. R. McCorkle, L. A. Praissman, et al., "Genomic signature tags (GSTs): a system for profiling genomic DNA," *Genome Research*, vol. 12, no. 11, pp. 1756–1765, 2002.

[21] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, article 163, pp. 1–7, 2004.

[22] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser, "Evolutionary implications of microbial genome tetranucleotide frequency biases," *Genome Research*, vol. 13, no. 2, pp. 145–158, 2003.

[23] 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 Research*, vol. 33, supplement 2, pp. W512–W515, 2005.

[24] M. Ganapathiraju, J. Klein-Seetharaman, R. Rosenfeld, et al., "Comparative n-gram analysis of whole-genome sequences," in *Proceedings of the Human Language Technologies Conference (HLT '02)*, San Diego, Calif, USA, March 2002.

[25] A. Apostolico, M. E. Bock, and S. Lonardi, "Monotony of surprise and large-scale quest for unusual words," in *Proceedings of the 6th Annual International Conference on Computational Molecular Biology (RECOMB '02)*, pp. 22–31, Washington, DC, USA, April 2002.

[26] 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, no. 1, pp. 63–72, 2007.

[27] I. Rish, "An empirical study of the naive bayes classifier," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI '01)*, pp. 41–46, Seattle, Wash, USA, August 2001.

[28] G. K. Zipf, *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge, Mass, USA, 1949.

[29] G. Hampikian and T. Andersen, "Absent sequences: nullomers and primes," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 12, pp. 355–366, Boise, Idaho, USA, January 2007.

[30] V. Y. Fofanov, C. Putonti, S. Chumakov, B. M. Pettitt, and Y. Fofanov, "Fast algorithm for the analysis of the presence of short oligonucleotide sequences in genomic sequences," Tech. Rep. #UH-CS-05-11, University of Houston, Houston, Tex, USA, May 2005.

[31] 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.

[32] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[33] J. C. Venter, K. Remington, J. F. Heidelberg, et al., "Environmental genome shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[34] S. J. Giovannoni, H. J. Tripp, S. Givan, et al., "Genetics: genome streamlining in a cosmopolitan oceanic bacterium," *Science*, vol. 309, no. 5738, pp. 1242–1245, 2005.

[35] S. T. Dyhrman, P. D. Chappell, S. T. Haley, et al., "Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*," *Nature*, vol. 439, no. 7072, pp. 68–71, 2006.

[36] S. M. Sowell, A. D. Norbeck, M. S. Lipton, et al., "Proteomic analysis of stationary phase in the marine bacterium "*Candidatus* pelagibacter ubique"," *Applied and Environmental Microbiology*, vol. 74, no. 13, pp. 4091–4100, 2008.