Computational Biology and Chemistry 35 (2011) 199–209

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Information-theoretic approaches to SVM feature selection for metagenome read classification

Elaine Garbarine[a], Joseph DePasquale[b], Vinay Gadia[a], Robi Polikar[b], Gail Rosen[a,*]

[a] Electrical and Computer Engineering Department, Drexel University, 3141 Chestnut St., Philadelphia, PA 19104, USA
[b] Electrical and Computer Engineering Department, Rowan University, 201 Mullhica Rd., Glassboro, NJ 08028, USA

ARTICLE INFO

Article history:
Received 14 July 2010
Received in revised form 25 April 2011
Accepted 25 April 2011

Keywords:
Metagenomics
Information theory
Support vector machines

ABSTRACT

Analysis of DNA sequences isolated directly from the environment, known as metagenomics, produces a large quantity of genome fragments that need to be classified into specific taxa. Most composition-based classification methods use all features instead of a subset of features that may maximize classifier accuracy. We show that feature selection methods can boost performance of taxonomic classifiers. This work proposes three different filter-based feature selection methods that stem from information theory: (1) a technique that combines Kullback–Leibler, Mutual Information, and distance information, (2) a text mining technique, TF-IDF, and (3) minimum redundancy-maximum-relevance (mRMR). The feature selection methods are compared by how well they improve support vector machine classification of genomic reads. Overall, the 6mer mRMR method performs well, especially on the phyla-level. If the number of total features is very large, feature selection becomes difficult because a small subset of features that captures a majority of the data variance is less likely to exist. Therefore, we conclude that there is a trade-off between feature set size and feature selection method to optimize classification performance. For larger feature set sizes, TF-IDF works better for finer-resolutions while mRMR performs the best out of any method for $N = 6$ for all taxonomic levels.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Traditional genomics studies have focused on culturing and sequencing a single microbe and studying its genome. This approach becomes problematic because over 99% (Handelsman, 2007) of microbes cannot be cultured and thus their genomes cannot be sequenced. The field of metagenomics has evolved to solve this problem. The goal of metagenomics is to take environmental samples containing many different organisms in their natural habitats, changing their focus from "how does one organism work?" to "how do many organisms interact with one another in their natural habitats?" (Konforti et al., 2008; Bohannon, 2008). These environmental mixtures of DNA are then sequenced by using high throughput sequencing techniques producing large quantities of genome sequence fragments that researchers wish to assemble into full genomes (Mardis, 2008; Pop and Salzberg, 2008).

The problem is that high throughput sequencing methods often produce small fragments of DNA, ranging from 35 base pairs (bp) to about 450bp (Venter et al., 2004; Wommack et al., 2008; Mardis, 2008). These short fragment sizes present significant challenges for the assembly, annotation and classification of genome fragments from multiple organisms.

In order to solve the classification problem, many automated classification algorithms, used to identify genomes based on the features derived from the DNA fragments, have been proposed. However, more powerful classifiers, such as support vector machines (SVM), cannot work with a large set of features (the so-called curse of dimensionality problem). Therefore, feature extraction and selection techniques are needed to determine the most useful and distinguishing features for classification. This work utilizes three different filtered feature selection methods that stem from information theory: (1) a technique that combines Kullback–Leibler, Mutual Information, and distance information (Garbarine and Rosen, 2008); (2) a text mining technique, TF-IDF (Gadia and Rosen, 2008); and (3) minimum-redundancy-maximum-relevance (Ding and Peng, 2003). Previously, we have developed (1) but have not shown its performance with a classifier, and we have only shown the performance of (2) only with a simple Euclidean-distance classifier. In this work, we compare the performance of mRMR and previously implemented feature selection methods using a support vector machine (SVM) classifier to assign taxonomic labels. SVM has been shown to be efficient at this task (McHardy et al., 2007), but it uses all possible features, which may not be optimal. By varying the feature size of the *N*-mers and

* Corresponding author.
E-mail address: gailr@ece.drexel.edu (G. Rosen).

1476-9271/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2011.04.007

the feature vector size (the number of *N*-mers), we assess the performance of the SVM, trained with features extracted by each of the aforementioned techniques, to determine the feature selection method that is best suited for genome discrimination.

1.1. Background on fragment identification for taxonomy

Metagenomics projects such as Venter Institute's Sorcerer II Global Ocean Expedition project (Rusch et al., 2007) and the MetaHIT project (Qin et al., 2010) (sequencing microbiomes from 124 people) are generating millions of reads. Large amounts of sequence fragments, created by these and similar studies, must then be assembled and annotated. To annotate unknown fragments, we wish to classify them to the closest taxa within a phylogeny and to the most specific taxa-level. The field of phylogenetics is focused on building clades which contain organisms that have derived traits. It is important to place fragments from previously uncultured organisms into a particular clade because doing so provides us information about traits that these unstudied organisms may have. This classification process can be a difficult problem due to the limited set of organisms that we have fully sequenced and the fact that the contents of the metagenomic mixtures are largely unknown.

Fragment classification for taxonomy can be broken down into two primary areas (Rosen et al., 2009): supervised (Huson et al., 2007; McHardy et al., 2007; Rosen et al., 2008) and unsupervised methods (Teeling et al., 2004; Chan et al., 2008). For our purposes, we are interested in supervised methods, because they leverage the existing knowledge of current databases to place metagenomic data of known and unknown origins into context.

One primary class of supervised methods are homology based approaches, which align sequence fragments to known genomes based on similarity. Homology-based methods such as BLAST (Madden, 2003), CARMA (Gerlach et al., 2009), and MEGAN (Huson et al., 2007) fall into this category. Studies of BLAST's performance (Wommack et al., 2008; Manichanh et al., 2008) have shown that its performance depends largely on whether close relatives of a given sequence are available for comparison. To solve this issue, MEGAN adds the use of a lowest common ancestor algorithm (LCA) to BLAST, which allows a fragment to generalize up to a higher branch in the tree instead of simply matching to the nearest neighbor (Wommack et al., 2008; Manichanh et al., 2008). CARMA implements the LCA algorithm and only matches the sequences that belong to particular protein families, which are acquired from the Pfam database (Finn et al., 2008).

Another category of supervised methods are composition-based approaches. These methods use *N*-length words or *N*-mers as features. The *N*-mers are used to build frequency profiles (how often each word occurs in a given sequence), which are then used to build models for classifiers. Methods such as Naive Bayes classifiers (NBC) (Sandberg et al., 2001; Rosen et al., 2008), TACOA (a k-nearest neighbor (k-NN) approach using a genomic feature vectors (GFVs) (Diaz et al., 2009)), and a support vector machine method, Phylopythia (McHardy et al., 2007), fall into this category. TACOA uses the word vector space model to form probabilities into the GFVs, which are then classified to database genomes using the k-NN algorithm. Support vector machines have been used with strong results for biological datasets and show significant promise for this type of application. Although Phylopythia also has good performance, TACOA has been shown to provide better classification performance (Diaz et al., 2009). On the other hand, while composition based approaches work very well, they suffer from *c* urse of dimensionality, as *N*-mers of sizes larger than 6 produce extremely large sets of features, which may produce too many dimensions (McHardy et al., 2007) and worsen performance. In order to address this issue, we propose an information theory-based feature selection method that extracts the most relevant information provided by the high dimensional feature set and represents such information in a (lower dimensional) subspace.

1.2. Composition-based fragment classification

As previously mentioned, *N*-mers are often used as features in supervised methods for fragment identification. An *N*-mer is simply a word composed of letters A, T, C, G (the four bases of DNA) of a particular length *N*. *N*-mers have been well established as usable classifier features. They have been used with Naive Bayes classifiers (Sandberg et al., 2001; Rosen et al., 2008) for fragment identification and more recently as features for an SVM classifier for the same purpose (McHardy et al., 2007). Both methods have shown strong results for fragment identification. The challenge of using *N*-mers as features for optimization-based classifiers, such as neural networks or SVMs, is that as *N* increases, the number of possible words increases exponentially. The total number of possible combinations can be represented by $words = 4^N$.

In our previous work with Naive Bayes classifiers, all possible *N*-mers were used as features for the classifier with varying sizes (up to $N = 15$); in general, larger *N* performed as well or better for the classification (Rosen et al., 2008). In the case of an SVM, or any other optimization-based classifier, however, using large *N* is not currently computationally tractable. The (SVM) classifier, when trained with *N*-mer frequency profiles, is a supervised classifier which constructs a 4^N-dimensional hyperplane to optimally separate data into categories. The 4^N-dimensions, chosen by the information theoretic measures discussed in Section 2, are the features input into the support vector machine. The SVM then seeks to classify an incoming, unknown fragment into one of the genomes in the training dataset by using the training data's *N*-mer features. Without feature selection the number of *N*-mers for larger word sizes would be overwhelming for the SVM, especially for $N > 6$.

A 6mer ($N = 6$) produces 4096 total possible words (or features) for the classifier; 7mers increase the number of potential features to 16,384. This means that 16,384 features describes each of 635 genomes, resulting in a $16,384 \times 635$ dimensional space, which is computationally challenging for current desktop computers. On the other hand, a generative classifier, the naïve Bayes approach, shows clearly that longer *N*-mers produce better classification (generalization) (Rosen et al., 2008), but this is simply due to the fact that as *N* increases, the number of possible words increases, and consequently the uniqueness of each individual word increases (Robin and Schbath, 2002). Conversely, we can think of this as the variance of a word occurrence decreasing across example fragment instances of a genome, which in turn makes the identification easier. In order to take advantage of the learning capabilities of SVMs, as well as the classification benefits that comes with the uniqueness of longer *N*-mers as features, some form of feature selection is necessary.

The uniqueness of each word increases with *N*. For example, for $N = 3$, there are only $4^3 = 64$ possible words (AAA, AAG, AAC,..., TTT), and each of these 64 3-letter words can and usually do appear many times in the genome. For $N = 15$, however, there are $4^{15} = 1,073,741,824$ (15-letter) words. It is much less likely for each one of these 15-letter words (e.g., AGTGGCTACGTACGTA) to occur many times compared to three-letter words.

1.3. Review of feature selection and information-theoretic approaches

In any pattern recognition problem, of utmost importance, perhaps more so than choosing the right classification algorithm, is

choosing the correct features (predictors) to train the classification algorithm; after all, it is the features where the information lies. In the presence of a large number of predictors, it may be difficult to determine which of these predictors are really relevant to the problem, and which ones are irrelevant. Irrelevant features effectively amount to noise, making the classifiers job all the more difficult. On the other hand, in the absence of prior knowledge on which features are relevant, it is tempting to use all of them to train the classifier and let the classifier to determine how to weight the features. This "brute-force approach" is not only inefficient, but also counterproductive, as using irrelevant features can – and usually do – deteriorate the classifiers performance. As described very elegantly in their aptly titled work, Greiner et al. showed that "knowing what does not matter" matters (Greinera et al., 1997) and classifier performance can be improved significantly by removing irrelevant features.

Several approaches have been developed over the years to determine the most relevant features for a given classification algorithm. These approaches are generally divided into two groups: filter approaches and wrapper approaches. Wrapper approaches are so-called because the feature selection is "wrapped" around the classifier being used. Essentially, wrapper approaches use a guided search such as forward or backward selection, to methodically add or eliminate features one a time, and trying each resulting combination of features to determine which subset of features provide the best classification performance when used with the chosen classifier. Random subspace methods, which use resampling to choose a random subset of features to generate a large number of sub-classifiers that are then combined using voting procedures, constitutes another subcategory of wrapper approaches (Ho, 1998). Necessarily, then, wrapper approaches are computationally expensive, as they require the classifier to be trained and evaluated a large number of times with different subsets of features. It is therefore important to choose a classifier that has a large enough predictive capacity but low enough computational complexity. Support vector machines (SVMs), which create linear hyperplanes in high dimensional space through kernel trick based computations in low dimensional space, are therefore good choices, and have been routinely used in many such applications including those involving genomic data (Guyon et al., 2002; Bi et al., 2003).

On the other hand, filter approaches apply a "filter" to the entire feature set to remove a subset of them deemed irrelevant by the chosen filter approach. Filter approaches typically use some transformation function, such as computing the frequency response of time-series data and choosing those spectral coefficients with the highest amplitudes. Information theoretic feature selection algorithms constitute a different subcategory of filter approaches, where the features are determined based on the amount of "information" they carry for the given classification problem, as measured by the joint probability of the features and the correct labels. The features that maximize an objective function, such as the mutual information, are then chosen as the most relevant features (Torkkola, 2003; Nenadic, 2007). Information theoretic feature selection approaches have been successfully used particularly in extremely large dimensional domains, such as text categorization (Dhillon et al., 2003; Novovicova and Malik, 2005), as in such domains with thousands or tens of thousands of features, wrapper approaches become computationally prohibitive. More recently, efforts in combining filter and wrapper based approaches have also been fruitful, as shown by Francoisa et al. (2007).

An overview of feature selection approaches can be found in Guyon et al.'s review article (Guyon and Elisseeff, 2003), whereas a more thorough treatment of information theoretic approaches can be found in Principe's recent text (Principe, 2010).

2. Information-theoretic approach for *N*-mer feature selection

The goal of traditional information theory is to maximize channel capacity by preserving the parts of the signal which hold the most information while simultaneously ignoring the non informative signal parts. This fundamental concept is vital to our ability to classify fragments to specific genomes or taxa. In this scenario of fragment classification the goal is to choose features which maximize our ability to distinguish between genomes while simultaneously ignoring the parts of the genomes which hold little information for this purpose.

2.1. Kullback–Leibler, Mutual Information and Difference method for N-mer feature selection

In the case of smaller *N*-mer sizes, such as 3 and 6, there are a tractable number of words for computing the genome frequency profiles, therefore we can use all of the *N*-mer frequency profiles as available features for a classifier. As *N* gets larger, the number of available words far exceeds the computational capabilities of an SVM classifier. Preliminary work has been conducted in this area to investigate which *N*-mers from a large set can distinguish between two organisms in a mixture (Garbarine and Rosen, 2008). In this feature selection method, three primary measures were used to select *N*-mers: Kullback–Leibler divergence, mutual information, and the frequency difference between *N*-mer counts in the two genomes.

Given two organisms, genome **A** and genome **B**, we can compute the Kullback–Leibler divergence for the *m*th *N*-mer between each genome in the set $\mathbf{C} = \{\mathbf{A}, \mathbf{B}\}$. Denoting the probability of the *m*th *N*-mer in genome set **C** represented by $p_C(m)$:

$$\mathbf{D_{KL}}(A(m), B(m)) = p_A(m)\log_2\frac{p_A(m)}{p_B(m)} + p_B(m)\log_2\frac{p_B(m)}{p_A(m)} \quad (1)$$

where $\mathbf{D_{KL}}$ is the symmetric Kullback–Leibler distance. $p_C(m)$ is calculated by the number of the *m*th *N*-mer in one of the genomes from the set $\mathbf{C} = \{A, B\}$, $X_C(m)$, divided by the total number of *N*-mers in that genome. The microbial strains genome lengths range from 160 k(bp) for *Candidatus Carsonella* to 13 Mil(bp) for *Sorangium Cellulosum*, so $p_C(m)$ estimates vary depending on the genome size. The genome sizes range from 300 kbp to 10 Mbp, This provides us with a measure of the divergence between genome **A** and genome **B** due to the *m*th *N*-mer.

Mutual information (MI) is a metric that provides a measure of the information between the *N*-mer and the set of genomes. To derive the mutual information between an *N*-mer and associated genomes, we have the genomes, **C**, and an *M*-dimensional feature vector $\mathbf{X_C} = \{X_C(1), X_C(2), X_C(3), \ldots, X_C(M)\}$, where each $X_C(m)$ represents the number of the *m*th *N*-mer present in a genome **C**. In our problem, we want to find the *N*-mers with the maximum MI between the *N*-mer and the set of genomes, $I(X(m), \mathbf{C})$ which are the $X(m)$'s that satisfy:

$$arg\max_{X(m)} I(X(m), \mathbf{C}) = arg\max \sum_{c \in \mathbf{C}} p(X_C(m), c)\log_2\frac{p(X_C(m), c)}{p(X_C(m))p(c)} \quad (2)$$

MI can be rewritten as $I(X(m), \mathbf{C}) = H(\mathbf{C}) - H(\mathbf{C}\,|\,X(m))$, where *H* is the entropy. Since the conditional entropy of the genomes given an *N*-mer, $H(\mathbf{C}\,|\,X(m))$, is always negative, $I(X(m), \mathbf{C})$ is maximized by maximizing $H(\mathbf{C}\,|\,X(m))$ as the marginal entropy of the genomes, $H(\mathbf{C})$, is constant. Therefore, the best *N*-mers that discriminate between genomes can be interpreted as the ones with the highest conditional entropy.

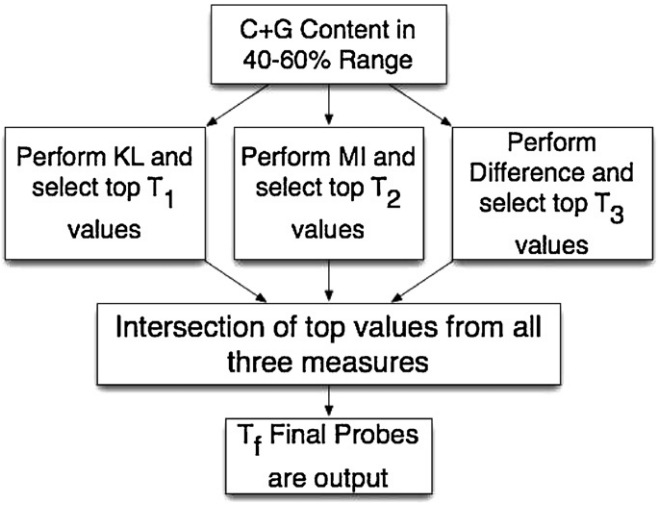

Fig. 1. Algorithm flow of the three measures used are Kullback–Leibler (KL), Mutual Information (MI) and Difference Measure (difference). T1, T2, and T3 represent the number of top features taken from each method. For the purposes in the paper, when we say we select the top-50 features, T1 = T2 = T3 = 50.

The final measure is the difference in frequency count of the mth N-mer between genome **A** and genome **B**, and is defined as:

$$D(n) = |X_A(m) - X_B(m)| \tag{3}$$

We tested this measure with two other measures, the Kullback–Leibler divergence and difference method on two bacteria. The goal was to select N-mers using the three measures which would strongly distinguish between the two organisms in the mixture. The KL, MI and difference measures were computed, and the results were intersected to find the top measures across all three measures seen in Fig. 1.

Fig. 1 illustrates the feature selection process with each of the steps of the process detailed in Garbarine and Rosen (2008). It is important to note that all measures are performed on a pair of genomes such that one genome can be identified from another genome. In order to extend these pairwise measures to the 100 genomes, we intersect all the T_f feature sets and choose the top-scoring 50, 100, 150, etc. features across all genomes to be used for training the SVM classifier.

2.2. TF-IDF method for N-mer feature selection

Text mining approaches such as Term Frequency-Inverse Document Frequency (TF-IDF) can also be applied to N-mer feature selection in a genome classification problem (Gadia and Rosen, 2008). TF-IDF can be viewed as an information theoretic measure, the amount of information of a term weighted by it's occurrence probability. More specifically, the IDF term represents a change in the amount of information after observing a specific term, while the TF term expresses the probability estimation that the term is actually observed (Aizawa, 2003).

The TF-IDF measure is broken into two components. First, we calculate the term frequency, which is usually defined as the word frequency divided by the total number of words in the document, l. In our case, the term frequency is the N-mer count divided by the total number of N-mers in the genome. The TF measure is then computed as:

$$tf_{ml} = \frac{d_{ml}}{\sum_{m=1}^{M} d_{ml}} \tag{4}$$

The inverse document frequency can be computed as:

$$idf_m = \log\left(\frac{L}{\#\ \text{of documents that have word}\ \ m}\right) \tag{5}$$

where L is the total number of genomes.

The IDF measure generally assumes that the M terms are sparse and only exist in a few documents. To adjust for the fact that with small N-mer sizes, there is a possibility of all words existing in all genomes we adjust IDF to:

$$idf_m = \log\left(\frac{L}{\sum_{l=1}^{L} d_{ml}}\right) \tag{6}$$

Therefore, our proposed TF-IDF measure is:

$$TF - IDF_m = \sum_{l=1}^{L}\left(\frac{d_{ml}}{\sum_{m=1}^{M} d_{ml}} \log\left(\frac{L}{\sum_{l=1}^{L} d_{ml}}\right)\right) \tag{7}$$

2.3. Preliminary results for TF-IDF and Mutual Information/Kullback–Leibler methods

Both the KL/MI/frequency difference and the TF-IDF methods described in the previous section have been applied to genome classification (Garbarine and Rosen, 2008). In the case of the MI, KL and difference measures, the test case was that of a mixture of two bacteria. The goal was to select N-mers using the three measures which would strongly distinguish between the two organisms in the mixture. The KL, MI and difference measures were computed, and the results were intersected to find the top measures across all three measures seen in Fig. 1.

In addition to demonstrating the KL/MI/frequency difference method on pairwise genomes, we also apply the TF-IDF method to hundreds of genomes to reduce the feature set needed for fragment classification. Our preliminary results for TF-IDF are based on a database of 635 microbes that belong to 470 distinct species and 260 distinct genera. One hundred 500bp fragments are chosen from each of the genomes and classified using only those features chosen by the TF-IDF (and variants). The overall accuracy is computed as the read being assigned the correct taxa. The TF-IDF measure is employed to sort the N-mers in the order of importance of the words. The order of importance is selected in three ways: (1) the "highest frequency of words in the least number of genomes (documents in the text-mining literature is called the *tfidf sort*, (2) the lowest frequency of words in most number of documents is called the *inverse-tfidf sort*, and (3) a random order to compare its performance with the others. Subsets of the important words are chosen, and the fragments are classified using the Euclidean classifier. Subset intervals in increments of 10% are chosen to compare the performance of selection of different cardinalities of features against selection of all features. Figs. 2 and 3 compare the identification accuracy performance for $N = 6$ and $N = 9$, respectively. For $N = 6$, randomly selected words perform better than the measures up until 40% of the words are chosen. The inverted TF-IDF using 90% of words performs 4% above the full-word performance (16% improvement). For $N = 9$, randomly selected words perform better than the TF-IDF measure. Inverse TF-IDF is again the best performer, particularly when more then 70% of the words are used for classification. We hypothesize that this could be due to the genetic structure of the genomes. In TF-IDF, the most frequent N-mers in a genome are chosen that are the least common across genomes;

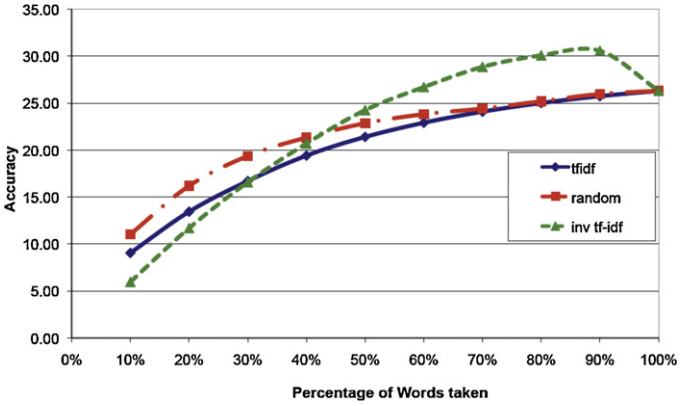

Fig. 2. Percentage of words taken vs. accuracy of a Euclidean classifier for TF-IDF measure with $N = 6$. If a genome were chosen by chance, the accuracy rate would be, 1/635 or 0.2%, so while 20–30% accuracy may seem low, it is 100-fold higher than chance.

the idea is to choose features that characterize a few genomes that can be used to discern those from others. But in biology, genomes that share the same genes have frequent *N*-mers, and frequently, with certain genes being prevalent within a phyla class! Therefore, inverse TF-IDF works better, since it is essentially choosing *N*-mers that are frequent and frequently found across all genomes. The inverted TF-IDF using 80% of words performs 2% above the full-word performance (a 10% improvement), showing that classifier performance can be improved by intelligently choosing a subset of words (Gadia and Rosen, 2008).

The text mining method demonstrates that the application of TF-IDF to genome classification and its ability to successfully reduce the number of features necessary to perform classification (Gadia and Rosen, 2008). While the reduction in features is only 10%, it bears noting that this reduction in features was observed for a very simple Euclidean classifier. As we show in our experimental results later in this paper, TF-IDF can be a particularly effective feature selection technique when paired with a stronger classifier.

2.4. mRMR method for N-mer feature selection

The minimum-Redundancy-Maximum-Relevance (mRMR) method of feature selection seeks to choose features that best characterize the statistical property of the target classification variable under the constraint that the chosen features are as mutually dissimilar to each other while still being as marginally similar to the classification variable as possible. Essentially the method chooses features as maximally relevant to the classification variable as possible while still being minimally redundant (Ding and Peng, 2003). The ideas of "relevance" and "redundancy" can be based on mutual information, statistical *t*-tests/*F*-tests, correlation, or distances.

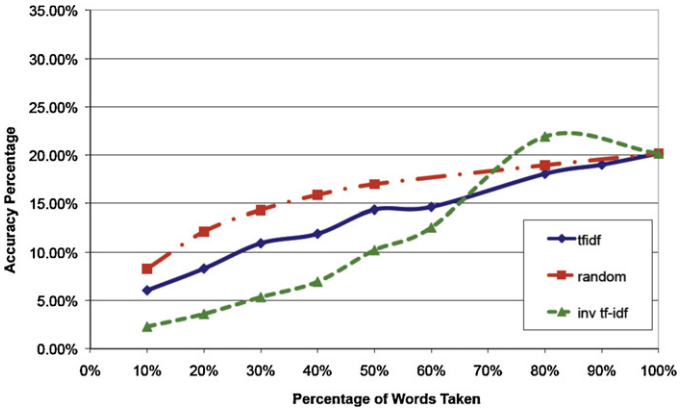

Fig. 3. Percentage of words taken vs. accuracy of Euclidean classifier for TF-IDF measure with $N = 9$.

Type	Acronym	Full Name	Formula
Discrete	MID	Mutual information difference	$\max_{i\in\Omega_S}[I(i,h) - \frac{1}{\lvert S\rvert}\sum_{j\in S} I(i,j)]$
	MIQ	Mutual information quotient	$\max_{i\in\Omega_S}\{I(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S} I(i,j)]\}$
Continuous	FCD	*F*-test correlation difference	$\max_{i\in\Omega_S}[F(i,h) - \frac{1}{\lvert S\rvert}\sum_{j\in S} \lvert c(i,j)\rvert]$
	FCQ	*F*-test correlation quotient	$\max_{i\in\Omega_S}\{F(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S} \lvert c(i,j)\rvert]\}$
	FDM	*F*-test distance multiplicative	$\max_{i\in\Omega_S}[F(i,h)\cdot\frac{1}{\lvert S\rvert}\sum_{j\in S} d(i,j)]$
	FSQ	*F*-test similarity quotient	$\max_{i\in\Omega_S}\{F(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S}\frac{1}{d(i,j)}]\}$

Fig. 4. Equations for mRMR method of feature selection for continuous and discrete variables.

For discrete variables, we wish to minimize redundancy by minimizing:

$$W_I = \frac{1}{|S|^2}\sum_{i,j\in S} I(i,j) \tag{8}$$

Additionally we want to maximize relevancy by maximizing:

$$V_I = \frac{1}{|S|}\sum_{i\in S} I(h,i) \tag{9}$$

where S is the set of features, $I(i, j)$ is the mutual information between features i and j and h is the target class.

The mRMR equations can also be defined for continuous variables simply by replacing mutual information with an *F*-statistic or correlation. Relevance and redundancy for continuous variables can also be defined using mutual information of hybrid variables (Peng et al., 2005). There are two methods to combine the ideas of maximum relevance and minimum redundancy, namely – additive combination max $(V - W)$ and multiplicative combination, max (V/W). Fig. 4 displays the final equation for both continuous and discrete variables for the mRMR method. For our mRMR implementation, we use MID (Fig. 4), the mutual information difference equation between discrete variables (since the *N*-mer distributions are discrete representations).

The mRMR feature selection method has been used in microarray data related to gene expression in cancer samples (Ding and Peng, 2003, 2005), and more recently for recognition and annotation of gene expression patterns in fly embryos (Zhou and Peng, 2007). The method has not been applied to feature selection for classification of bacterial genomes, which chooses features from the entire genome instead of simply using genes.

3. Materials and methods

To investigate feature selection with an SVM, 100 bacterial genomes were chosen for the training database. While 100 strains is nowhere near the complexity of a soil sample, it is representative of a diverse sample (e.g. in Mavromatis et al., 2007, where 113 genomes were used to construct high complexity samples). The list of organisms used in our work can be found in Tables 3 and 4 in Appendix. The 100 strains belong to 3 phyla, 14 genera and 64

Table 1
SVM classification results for 6mers (in %).

Taxonomic level	Method	50 features	100 features	150 features	500 features	1000 features	All features	CARMA	TACOA
	TF-IDF	3.96	2.53	2.78	1.54	1.00			
Strain	mRMR	1.39	6.73	**7.13**	2.03	1.01	1.01	N/A	N/A
	MI/KL	4.19	3.53	2.68	1.00	1.00			
	Variance	5.54	5.39	3.87	1.00	1.00			
	TF-IDF	7.23	3.86	3.63	2.09	1.00			
Species	mRMR	2.3	13.68	14.78	7.98	1.04	3.00	**17.40**	N/A
	MI/KL	7.92	6.41	6.04	3.00	3.00			
	Variance	10.60	10.41	7.59	3.00	3.00			
	TF-IDF	20.93	10.97	15.68	10.89	8.00			
Genus	mRMR	10.92	**29.57**	29.21	12.06	7.06	7.00	25.10	10.40
	MI/KL	21.53	17.39	13.76	7.00	7.00			
	Variance	24.30	21.69	16.25	7.00	7.00			
	TF-IDF	20.04	8.83	13.81	10.54	8.00			
Family	mRMR	10.90	**28.31**	27.87	11.89	7.06	7.00	26.40	N/A
	MI/KL	19.58	14.87	12.41	7.00	7.00			
	Variance	22.57	19.43	14.14	7.00	7.00			
	TF-IDF	22.67	11.21	15.68	10.89	8.00			
Order	mRMR	17.86	32.86	**32.87**	18.13	7.06	7.00	27.00	12.8
	MI/KL	22.77	17.90	14.14	7.00	7.00			
	Variance	26.39	22.99	17.01	7.00	7.00			
	TF-IDF	39.80	26.18	23.55	12.29	8.00			
Phyla	mRMR	40.87	**53.34**	51.94	33.63	36.06	42.00	29.00	22.20
	MI/KL	36.16	25.77	41.21	42.00	42.00			
	Variance	44.27	42.23	41.07	42.02	42.00			

species, according to NCBI Taxonomy (Benson et al., 2011), though it is of note that NCBI Taxonomy does not necessarily reflect the true taxonomy in nature (as described by Bergey's manual) (Wang et al., 2007). Of the 14 genera, 6 belong to the firmicutes phyla, 6 belong to the proteobacteria phyla and 2 belong to the cyanobacteria phyla. Most genera have 7 example strains with the exception of lactobacillus and mycoplasma which have 8 strains. Further, 50 species have one example strain, 7 species have 2 example strains, 2 species have 3 example strains, 1 species has four example strains, 2 species have 6 example strains and finally 2 species have seven example strains.

The training dataset consists of the frequency profiles of the *N*-mers obtained from 50,000 long fragments. These frequency profiles include the number of times each possible combination of

Table 2
SVM classification results for 9mers (in %).

Taxonomic level	Method	50 features	100 features	150 features	500 features	1000 features
	TF-IDF	1.92	**1.96**	**1.96**	1.91	1.69
	mRMR	1.04	1.15	1.22	1.46	1.53
Strain	MI/KL	1.43	1.85	1.77	1.41	1.12
	MI One vs. All	1.42	1.75	1.70	1.78	1.79
	Variance	1.74	1.85	1.85	1.54	1.43
	TF-IDF	1.92	1.96	1.96	1.91	1.69
	mRMR	1.04	4.16	4.63	6.46	**7.40**
Species	MI/KL	2.14	6.89	5.31	2.26	2.47
	MI One vs. All	2.54	3.40	3.35	3.55	3.56
	Variance	6.44	4.88	4.67	3.2 0	2.85
	TF-IDF	13.24	13.96	**14.08**	13.13	11.90
	mRMR	7.59	8.12	8.57	10.25	10.40
Genus	MI/KL	10.42	10.19	9.38	8.61	8.31
	MI One vs. All	9.22	11.24	11.25	12.17	12.25
	Variance	8.86	9.53	10.03	9.08	8.70
	TF-IDF	13.24	13.96	**14.10**	8.78	8.61
	mRMR	7.59	8.12	8.57	10.25	10.40
Family	MI/KL	10.59	10.30	9.38	8.63	8.31
	MI One vs. All	9.22	11.24	11.30	12.17	12.25
	Variance	8.88	9.48	10.03	9.08	8.70
	TF-IDF	13.24	13.96	14.08	13.13	11.90
	mRMR	8.23	9.59	10.69	14.47	**16.47**
Order	MI/KL	10.59	10.33	9.38	8.64	8.31
	MI One vs. All	10.15	11.28	11.37	12.17	12.25
	Variance	9.07	9.56	10.03	9.08	8.70
	TF-IDF	38.50	35.94	34.39	14.40	12.82
	mRMR	42.74	43.68	44.25	**45.18**	42.14
Phyla	MI/KL	16.60	20.78	14.17	10.89	8.92
	MI One vs. All	43.93	40.06	39.44	37.68	37.60
	Variance	23.68	20.15	16.93	13.95	13.90

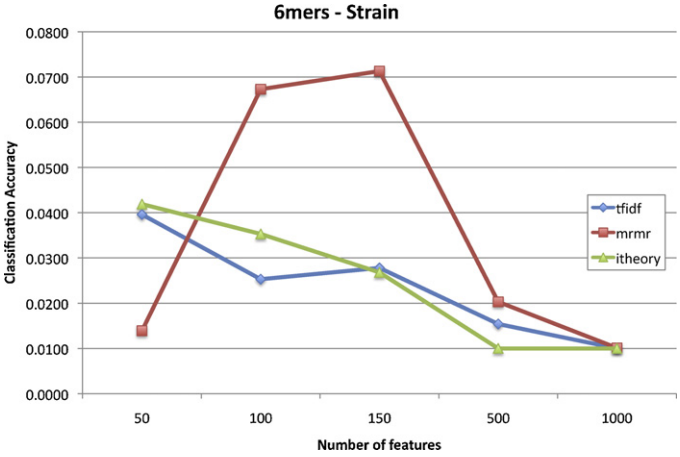

Fig. 5. Feature Set Size vs. Classification Accuracy at the Strain level, $N = 6$.

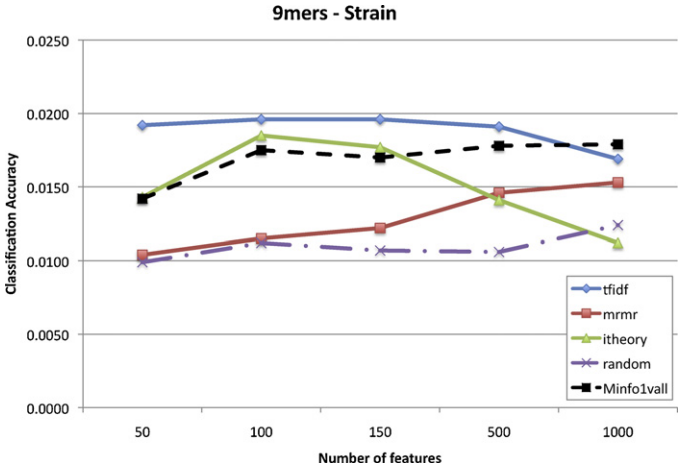

Fig. 7. Feature Set Size vs. Classification Accuracy at the Strain level, $N = 9$.

A,T,C,G of the given N-mer length appears in each genome. A large matrix is created with the frequency of all possible words in each genome. The word, or N-mer, sizes used to create these frequency profiles are 6 and 9. One-hundred 50,000bp-long fragments are chosen randomly, and the process is repeated 100 times, once for each of the 100 genomes. All results are therefore averages of 100 trials. Once the SVM classifiers are trained (for each of the 100 trials), fragments of 500 base pairs were randomly selected from each genome and used as the test data for the classifier. While the test data comes from the training data, it is still a difficult problem since there is great intra-genome variation among the short fragments.

From this set, the entire genomes were used for the TF-IDF, information theory and mRMR methods to select sets of features. TF-IDF was applied on the 100 genome frequency tables, and the results were sorted using the TF-IDF sort method discussed in the previous work (Gadia and Rosen, 2008). This generates a list of words with the highest frequency that are in the least number of genomes. From this sorted list, features were chosen to pass to the SVM classifiers for training.

The mRMR method was also applied to the 100 genome data set. In order to use mRMR, the frequency profiles were discretized. While mRMR does not require discretized data, it has been shown that discretizing the data produces much better results (Ding and Peng, 2005). Given this information, we chose to discretize the frequency profiles into 3 states using the method suggested by the researchers who developed mRMR (Ding and Peng, 2005). We found the mean and standard deviation of all word frequencies. Then, states were created that discretize frequencies that fall below $mean - std$ as -2, frequencies that fall above $mean + std$ as +2, and frequencies that fall in between these two values as 0. The mRMR algorithm, using mutual information difference (MID), was then applied to the now discretized data to select features.

Finally, the Information-Theoretic approach combining Mutual Information and Kullback–Leibler Divergence was applied to the frequency profiles. This method must be applied to the data in a pairwise fashion. In other words, MI/KL must be applied to the set genome1/genome 2 then again to genome 1/genome 3, etc. all the way through until genome 99/genome 100. This process creates a large matrix of results for 4950 possible pairs, which must now be combined into one set of features. To do this, a thresholding method was used. First, the results for all pairs were sorted. Next, the top 25% of the sorted matrix is considered and the number of occurrences of each word in this top 25% for all pairs are counted and the results are sorted. From this final, sorted vector features were chosen to pass to the SVM classifier.

To compare our methods to a non-information theoretic method, we chose the features with the highest variance and compared our methods to those results. For each feature, m, we computed $\sigma_m^2 = (1/N)\sum_{i=1}^{N} X(m)_i - \mu$ where μ is the average counts for all features and $X(m)$ is the count for feature, m. Features were rank-ordered from the highest σ^2 to lowest, similar to the rank-ordering procedure of the information-theoretic selection.

The feature set sizes selected by all of the methods were 50,100,150, 500, and 1000. Once features were selected, fragments of 50,000 base pairs are randomly selected from the genomes and used in conjunction with the sets of features to train an ensemble of SVM classifiers. 4950 SVM classifiers were trained, one for each possible pair of genomes from the original set of 100. For example, SVM1 is a classifier for genomes 1 and 2, SVM2 for 1 and 3, etc. The same selected features were used to train all classifiers.

Once the SVM classifiers were trained, fragments of 500 base pairs – not included in the training data – were selected from each genome and used as the test data for the classifier. In our implementation, each fragment is passed into a one-vs.-one SVM, which generates a vote on which genome the fragment most likely to belongs to. The genome that receives the majority of the votes is then chosen by the classifier as the most likely source of the fragment. It is important to note that with this ensemble of classifiers an SVM which does not contain the genome in question is equally likely to vote for either of the genomes it was originally trained on. For example, if a fragment comes into the system from genome 9

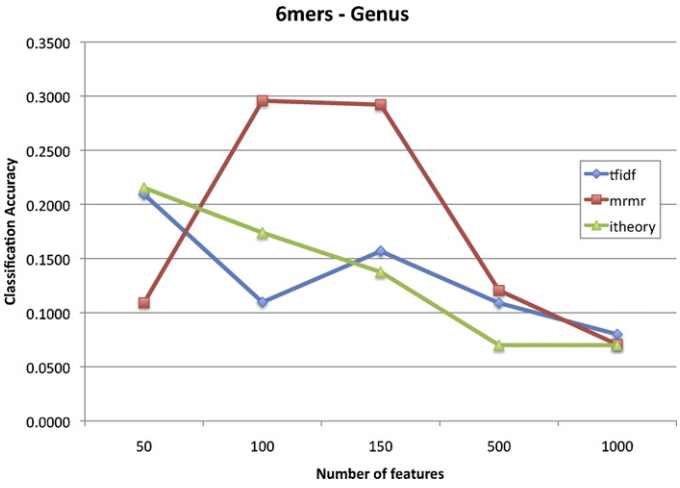

Fig. 6. Feature Set Size vs. Classification Accuracy at the Genus level, $N = 6$.

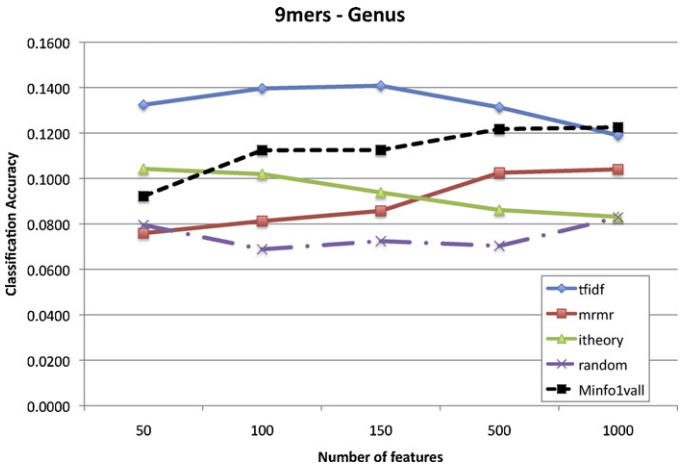

Fig. 8. Feature Set Size vs. Classification Accuracy at the Genus level, $N = 9$.

and it goes to SVM1 (genome 1 and genome 2), that SVM has an equal chance of classifying the fragment as genome 1 or genome 2.

We compared our methods to two methods, CARMA and TACOA, neither of which is based on information theoretic features. Internal default parameters were used for both. CARMA was executed by using WebCARMA 1.0 (Gerlach et al., 2009) in March 2010. The reads were uploaded to the server. For TACOA and CARMA, each program requires no input parameters.

4. Results

The accuracy of the SVM classifier for the features selected using TF-IDF, MI/KL and mRMR were recorded for all feature set sizes for each of the two *N*-mer sizes. The accuracy of the classifier was tracked for all levels of taxonomy from lowest to highest. These are ordered as follows: Strain, Species, Genus, Family, Order, Class, Phyla. Table 1 shows the classification accuracy for all 3 methods for feature set sizes 50, 100, 150, 500, 1000, and 4096 (all) for 6mers while Table 2 displays the 3 methods plus MI One vs. All for feature set sizes 50, 100, 150, 500, 1000 for the 9mer tests, since it was computationally intractable to use all features in the 9mer case. In Table 1, we compare the results to two other methods, CARMA and TACOA. The highest performance accuracy is bolded. CARMA was chosen since it is the most intricate homology-based method, which uses a last-common ancestor algorithm, and selectively chooses protein families. TACOA was chosen since it is a composition-based method, and the authors show that it obtains better performance than Phylopythia, another SVM-based approach (McHardy et al., 2007). We show that feature selection greatly improves the performance of an SVM classifier compared to these current methods.

Fig. 5 displays the 6mer results for TF-IDF, mRMR, and MI/KL at the Strain taxonomy level. Fig. 6 shows the results of the three methods at the Genus level. The same two plots can be seen for 9mers in Figs. 7 and 8. The 9mer plots have an additional set of data points for sets of features that were randomly selected.

Given the results from the MI/KL, TF-IDF and mRMR methods on 9mers further runs were attempted to improve the classification accuracy. A mutual information only approach was employed to find the mutual information of a word in one genome vs. the average mutual information of the same word in all other genomes. This "one vs. all" mutual information computation was repeated for all 100 genomes generating a large matrix of results for each of the 100 genomes vs. the other 99 genomes in the set. This matrix was then combined by again sorting the results from each genome, thresholding the sorted matrix and counting the number of occurrences of each word in the sorted and thresholded matrix. The results of the mutual information "one vs. all" method compared to other 9mer results can be found in Table 2 and Figs. 7 and 8.

To explore how similar top-percentages affect the SVM classification, we selected the top-0.37% of the 6mer features (top-15) and compared them to the top-0.38% (top-1000) of the 9mer features. A direct comparison for mRMR is (top-15 6mers/top-1000 9mers) 1.9%/1.5% for strain, 5.0%/7.4% for species, 12.5%/10.4% for genera, 11.7%/10.4% for family, 14.0%/16.5% for order, and 45.5%/42.1% for phyla. For ML/KI, top-15 6mers/top-1000 9mers performance is 1.3%/1.1% for strain, 4.0%/2.5% for species, 9.7%/8.3% for genera, 8.4%/8.3% for family, 11.2%/8.3% for order, and 32.4%/8.9% for phyla. This shows that the mRMR performance with the same top-% of features is comparable between 6mers and 9mers, while the MI/KL method decreases in performance for 9mers. Therefore, we expect the top-6400 mRMR using 9mers to achieve similar performance to the top-100 mRMR using 6mers, but using only 100 features is less computationally complex. Our goal of feature selection is to use less features while obtaining better accuracy, so it is more feasible to use 6mers.

5. Discussion

The classification results indicate that for 6mers with feature sets of 100 or 150 at the Strain level, mRMR is clearly the best method achieving much higher classification accuracy than other methods for genus-level and above (i.e.: genus, family, order, phylum). Also, mRMR performs well for larger feature sets, especially at the higher taxonomic levels like order and phylum for feature sets of 500 and 1000. The same observations hold true for 6mers at the genus-level. Table 1 demonstrates that as we move up through the taxonomic levels, mRMR and MI/KL both outperform TF-IDF especially at the Class and Phyla levels.

In Table 1, we compare the feature selection methods for SVM with a homology-based method, CARMA and a composition-based method, TACOA. Neither CARMA nor TACOA provide strain-level classifications. TACOA also only classifies at the genus-level and above, and the results were also missing the family-level classification. An "N/A" is placed in that column when there are missing taxonomic levels. Only for the species-level, CARMA outperforms the MI/KL feature selection method by about 3% accuracy. Classification accuracy improves when classifying to higher levels on the taxonomic tree, with mRMR yielding a 24% increase over CARMA at the phyla-level. This demonstrates that the SVM is able to capture higher-level taxonomies better. Also, using 6mer feature selection for 100 to 150- features, boosts performance by over 17% than not using any feature selection at all.

The 9mer results demonstrate decreased classification accuracy as compared to 6mers for all methods. For the family-level and finer-resolution, 100-150 features that we selected still yielded the best results for 9mers, similar to the number of features needed to yield the best results for 6mers. Promisingly, we see that 1000-long feature sets performs the best for order-level classification, and 500-long feature sets perform the best for phyla-level classification. This shows that more features may be necessary to capture the variance of the data when such features are available. But it still remains that SVM using 6mers performs better; the SVM performs better if each feature of a small set represents more data variation. This is probably due to the discriminative classification nature of SVM, compared to the generative classification of NBC which causes NBC to "memorize" the data.

TF-IDF outperforms all other methods at the strain and genus levels with the exception of the Mutual Information one vs. all method (applied to attempt to improve on the results of the MI/KL method). The mRMR method performs well even for 9-mers. It

bears noting here that for feature set sizes of 50, 100, 150 and 500, mRMR is capable of selecting features from the very large set of 262,144 in approximately 8 min on a 2.5 GHz Intel Core 2 Duo processor with 4 GB of RAM. However, the algorithm is not optimized for over 500 features and the computation time increases to almost 40 h on the same machine to choose 1000 features with very minimal benefits. In this situation the Mutual Information one vs. all method or TF-IDF would be much more desirable being able to run in approximately 10 min and 1 h, respectively regardless of the feature set size to be chosen. Another important note is that Mutual Information one vs. all and mRMR outperform other methods in terms of classification accuracy at the Class and Phyla levels and Mutual Information one vs. all consistently has high classification accuracy at the Phyla level while the mRMR method accuracy decreases with feature set size as can be seen in Table 2.

The low classification accuracy, seen for 9mers as compared to 6mers, is due largely to the significant difference in number of possible features. For 6mers there are only 4096 possible combinations of A, T, C, G that are length 6, while at length 9, the number of possible combinations is 262,144. Thus, there is lower variance for each 9mer feature, and it is harder to capture the variance of the data with few selected features. This means that for 6mers, 500 features represents approximately 12% of the total number of features, while for 9mers it only represents 0.2% of the total number of features. While we show that mRMR performs similarly with the same top-percentage of features, MI/KL does not, and this is due to

Table 3
Bacterial genomes.

Strain	Species	Genera	Phyla
Bacillus amyloliquefaciens FZB42	Bacillus amyloliquefaciens	Bacillus	Firmicutes
Bacillus anthracis str. Sterne	Bacillus anthracis	Bacillus	Firmicutes
Bacillus clausii KSM-K16	Bacillus clausii	Bacillus	Firmicutes
Bacillus licheniformis ATCC 14580	Bacillus licheniformis	Bacillus	Firmicutes
Bacillus subtili	Bacillus subtili	Bacillus	Firmicutes
Bacillus thuringiensis str. Al Hakam	Bacillus thuringiensis	Bacillus	Firmicutes
Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis	Bacillus	Firmicutes
Burkholderia cenocepacia AU 1054	Burkholderia cenocepacia	Burkholderia	Proteobacteria
Burkholderia cenocepacia HI2424	Burkholderia cenocepacia	Burkholderia	Proteobacteria
Burkholderia mallei ATCC 23344	Burkholderia mallei	Burkholderia	Proteobacteria
Burkholderia pseudomallei 1106a	Burkholderia pseudomallei	Burkholderia	Proteobacteria
Burkholderia pseudomallei 1710b	Burkholderia pseudomallei	Burkholderia	Proteobacteria
Burkholderia pseudomallei K96243	Burkholderia pseudomallei	Burkholderia	Proteobacteria
Burkholderia xenovorans LB400	Burkholderia xenovorans	Burkholderia	Proteobacteria
Clostridium beijerinckii NCIMB 8052	Clostridium beijerinckii	Clostridium	Firmicutes
Clostridium botulinum A	Clostridium botulinum	Clostridium	Firmicutes
Clostridium botulinum A Hall	Clostridium botulinum	Clostridium	Firmicutes
Clostridium perfringens	Clostridium perfringens	Clostridium	Firmicutes
Clostridium perfringens ATCC 13124	Clostridium perfringens	Clostridium	Firmicutes
Clostridium perfringens SM101	Clostridium perfringens	Clostridium	Firmicutes
Clostridium phytofermentans ISDg	Clostridium phytofermentans	Clostridium	Firmicutes
Escherichia coli 536	Escherichia coli	Escherichia	Proteobacteria
Escherichia coli APEC O1	Escherichia coli	Escherichia	Proteobacteria
Escherichia coli E24377A	Escherichia coli	Escherichia	Proteobacteria
Escherichia coli HS	Escherichia coli	Escherichia	Proteobacteria
Escherichia coli O157H7	Escherichia coli	Escherichia	Proteobacteria
Escherichia coli O157H7 EDL933	Escherichia coli	Escherichia	Proteobacteria
Escherichia coli W3110	Escherichia coli	Escherichia	Proteobacteria
Lactobacillus acidophilus NCFM	Lactobacillus acidophilus	Lactobacillus	Firmicutes
Lactobacillus brevis ATCC 367	Lactobacillus brevis	Lactobacillus	Firmicutes
Lactobacillus casei ATCC 334	Lactobacillus casei	Lactobacillus	Firmicutes
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	Lactobacillus delbrueckii	Lactobacillus	Firmicutes
Lactobacillus gasseri ATCC 33323	Lactobacillus gasseri	Lactobacillus	Firmicutes
Lactobacillus helveticus DPC 4571	Lactobacillus helveticus	Lactobacillus	Firmicutes
Lactobacillus plantarum subsp. plantarum	Lactobacillus plantarum	Lactobacillus	Firmicutes
Lactobacillus sakei 23K	Lactobacillus sakei	Lactobacillus	Firmicutes
Mycoplasma gallisepticum	Mycoplasma gallisepticum	Mycoplasma	Tenericutes
Mycoplasma genitalium	Mycoplasma genitalium	Mycoplasma	Tenericutes
Mycoplasma hyopneumoniae 232	Mycoplasma hyopneumoniae	Mycoplasma	Tenericutes
Mycoplasma hyopneumoniae J	Mycoplasma hyopneumoniae	Mycoplasma	Tenericutes
Mycoplasma mobile 163K	Mycoplasma mobile	Mycoplasma	Tenericutes
Mycoplasma pneumoniae	Mycoplasma pneumoniae	Mycoplasma	Tenericutes
Mycoplasma pulmonis	Mycoplasma pulmonis	Mycoplasma	Tenericutes
Mycoplasma synoviae 53	Mycoplasma synoviae	Mycoplasma	Tenericutes
Prochlorococcus marinus AS9601	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Prochlorococcus marinus CCMP1375	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Prochlorococcus marinus MED4	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Prochlorococcus marinus MIT 9313	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Prochlorococcus marinus MIT 9303	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Prochlorococcus marinus MIT 9312	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Prochlorococcus marinus MIT 9515	Prochlorococcus marinus	Prochlorococcus	Cyanobacteria
Pseudomonas aeruginosa	Pseudomonas aeruginosa	Pseudomonas	Proteobacteria
Pseudomonas aeruginosa PA7	Pseudomonas aeruginosa	Pseudomonas	Proteobacteria
Pseudomonas fluorescens Pf-5	Pseudomonas fluorescens	Pseudomonas	Proteobacteria
Pseudomonas fluorescens PfO-1	Pseudomonas fluorescens	Pseudomonas	Proteobacteria
Pseudomonas mendocina ymp	Pseudomonas mendocina	Pseudomonas	Proteobacteria
Pseudomonas putida GB 1	Pseudomonas putida	Pseudomonas	Proteobacteria
Pseudomonas stutzeri A1501	Pseudomonas stutzeri	Pseudomonas	Proteobacteria

Table 4
Bacterial genomes (cont.).

Strain	Species	Genera	Phyla
Rickettsia bellii OSU 85-389	Rickettsia bellii	Rickettsia	Proteobacteria
Rickettsia canadensis McKiel	Rickettsia canadensis	Rickettsia	Proteobacteria
Rickettsia conorii	Rickettsia conorii	Rickettsia	Proteobacteria
Rickettsia felis URRWXCal2	Rickettsia felis	Rickettsia	Proteobacteria
Rickettsia prowazekii	Rickettsia prowazekii	Rickettsia	Proteobacteria
Rickettsia rickettsii Iowa	Rickettsia rickettsii	Rickettsia	Proteobacteria
Rickettsia rickettsii Sheila Smith	Rickettsia rickettsii	Rickettsia	Proteobacteria
Shewanella sp. ANA-3	Shewanella	Shewanella	Proteobacteria
Shewanella sp. MR-7	Shewanella	Shewanella	Proteobacteria
Shewanella amazonensis SB2B	Shewanella amazonensis	Shewanella	Proteobacteria
Shewanella denitrificans OS217	Shewanella denitrificans	Shewanella	Proteobacteria
Shewanella frigidimarina NCIMB 400	Shewanella frigidimarina	Shewanella	Proteobacteria
Shewanella halifaxensis HAW EB4	Shewanella halifaxensis	Shewanella	Proteobacteria
Shewanella putrefaciens CN-32	Shewanella putrefaciens	Shewanella	Proteobacteria
Staphylococcus aureus JH1	Staphylococcus aureus	Staphylococcus	Firmicutes
Staphylococcus aureus JH9	Staphylococcus aureus	Staphylococcus	Firmicutes
Staphylococcus aureus N315	Staphylococcus aureus	Staphylococcus	Firmicutes
Staphylococcus aureus NCTC 8325	Staphylococcus aureus	Staphylococcus	Firmicutes
Staphylococcus aureus USA300	Staphylococcus aureus	Staphylococcus	Firmicutes
Staphylococcus aureus USA300 TCH1516	Staphylococcus aureus	Staphylococcus	Firmicutes
Staphylococcus saprophyticus	Staphylococcus saprophyticus	Staphylococcus	Firmicutes
Streptococcus mutans	Streptococcus mutans	Streptococcus	Firmicutes
Streptococcus pyogenes MGAS10270	Streptococcus pyogenes	Streptococcus	Firmicutes
Streptococcus pyogenes MGAS10750	Streptococcus pyogenes	Streptococcus	Firmicutes
Streptococcus pyogenes MGAS2096	Streptococcus pyogenes	Streptococcus	Firmicutes
Streptococcus pyogenes MGAS315	Streptococcus pyogenes	Streptococcus	Firmicutes
Streptococcus pyogenes MGAS5005	Streptococcus pyogenes	Streptococcus	Firmicutes
Streptococcus pyogenes MGAS8232	Streptococcus pyogenes	Streptococcus	Firmicutes
Synechococcus CC9311	Synechococcus sp. CC9311	Synechococcus	Cyanobacteria
Synechococcus CC9605	Synechococcus sp. CC9605	Synechococcus	Cyanobacteria
Synechococcus CC9902	Synechococcus sp. CC9902	Synechococcus	Cyanobacteria
Synechococcus JA-3-3Ab	Synechococcus sp. JA-3-3Ab	Synechococcus	Cyanobacteria
Synechococcus WH 7803	Synechococcus sp. WH 7803	Synechococcus	Cyanobacteria
Synechococcus elongatus PCC 6301	Synechococcus elongatus	Synechococcus	Cyanobacteria
Synechococcus WH8102	Synechococcus sp. WH8102	Synechococcus	Cyanobacteria
Yersinia enterocolitica 8081	Yersinia enterocolitica	Yersinia	Proteobacteria
Yersinia pestis Antiqua	Yersinia pestis	Yersinia	Proteobacteria
Yersinia pestis CO92	Yersinia pestis	Yersinia	Proteobacteria
Yersinia pestis KIM	Yersinia pestis	Yersinia	Proteobacteria
Yersinia pestis Pestoides F	Yersinia pestis	Yersinia	Proteobacteria
Yersinia pseudotuberculosis IP32953	Yersinia pseudotuberculosis	Yersinia	Proteobacteria
Yersinia pseudotuberculosis IP31758	Yersinia pseudotuberculosis	Yersinia	Proteobacteria

each 9mer feature capturing less variation in the data. Some of the methods for 9mers such as mRMR and Mutual Information one vs. all show an upward trend of classification accuracy as the number of features increases and indicates that a larger percentage of features could lead to better classification results for these methods. However, the current classification scheme is limited in that using more features with an SVM can often lead to much longer computation times. Therefore, we deduce that 6mer feature selection is a sufficient trade-off between time and accuracy.

6. Conclusions

In this paper we present an information-theoretic approach to feature selection that improves SVM genome classification. Most composition-based methods use all features and do not use an intelligent feature selection, and we show that feature selection methods can boost performance of these methods. We also show that feature selection may not work as well if the number of features is too large, where there may not be a small set of features that capture most of the data variance. There are trade-offs between feature set sizes and methods. Therefore, we conclude that $N = 6$ yields better results than both $N = 3$ or $N = 9$ showing that there is a trade-off between feature set size and performance; although, TF-IDF works better on the $N = 9$ level for fine-resolutions. Overall, mRMR using $N = 6$ performs well in our benchmark study in most cases, and especially performs well on the phyla-level.

Acknowledgements

This work was supported by the National Science Foundation CAREER award #0845827 and DOE award DE-SC0004335.

Appendix A.

See Tables 3 and 4.

References

Aizawa, A., 2003. An information-theoretic perspective of TF-IDF measures. Information Processing & Management 39 (1), 45–65, doi:10.1016/S0306-4573(02)00021-3.
Benson, D.A., et al., 2011. Genbank. Nucleic Acids Research 39 (1), D27–D32.
Bi, J., Bennett, K.P., Embrechts, M., Breneman, C.M., 2003. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research 3, 1229–1243.
Bohannon, J., 2008. Confusing kinships. Science Magazine 320 (5879), 1031–1033.
Chan, C.-K., et al., 2008. Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. Journal of Biomedicine and Biotechnology.
Dhillon, I.S., Mallela, S., Kumar, R., 2003. A divisive information-theoretic feature clustering algorithm for text classification. Journal of Machine Learning Research 3, 1265–1287.
Diaz, N.N., et al., 2009. Tacoa – taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10 (56), doi:10.1186/1471-2105-10-56.
Ding, C., Peng, H., 2003. Minimum redundancy feature selection from microarray gene expression data. Computational Systems Bioinformatics Conference, International IEEE Computer Society 0, 523.

Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3 (2), 185–205.
Finn, R.D., et al., 2008. The pfam protein families database. Nucleic Acids Research 36, 281–288.
Francoisa, D., Rossib, F., Wertza, V., Verleysen, M., 2007. Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing 70 (7–9), 1276–1288.
Gadia, V., Rosen, G. L., 2008. A text-mining approach for classification of genomic fragments. In: IEEE International Workshop on Biomedical and Health Informatics.
Garbarine, E., Rosen, G., 2008. An information-theoretic method of microarray probe design for genome classification. In: Engineering in Medicine and Biology Conference.
Gerlach, W., et al., 2009. Webcarma: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics 10 (430), doi:10.1186/1471-2105-10-430.
Greinera, R., Grove, A.J., Kogand, A., 1997. Knowing what doesn't matter: exploiting the omission of irrelevant data. Artificial Intelligence 97 (1–2), 345–380.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.
Guyon, I., Weston, J., Barnhil, l.S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.
Handelsman, J., 2007. Committee on Metagenomics: Challenges and Functional Applications. The National Academies Press.
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 832–844.
Huson, D.E., Auch, A.F., Qi, J., Schuster, S.C., 2007. Megan analysis of metagenomic data. Genome Research 17, 377–386.
Konforti, B., et al., 2008. Sequencing the microbial soup. Nature Structural and Molecular Biology 15 (115).
Madden, T., 2003. The NCBI Handbook. NIH, Ch. Chapter 16: The BLAST Sequence Analysis Tool.
Manichanh, C., et al., 2008. A comparison of random sequence reads versus 16s rDNA sequences for estimating the biodiversity of a metagenomic library. Nucleic Acids Research 36 (16), 5180–5188.
Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Elsevier Trends in Genetics 24 (3), 142–149.
Mavromatis, K., et al., 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods 4, 495–500.
McHardy, A.C., et al., 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods 4, 63–72.
Nenadic, Z., 2007. Information discriminant analysis: feature extraction with an information-theoretic objective. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (8), 1394–1407.
Novovicova, J., Malik, A., 2005. Information-theoretic feature selection algorithms for text classification. In: Proceedings of IEEE International Joint Conference on Neural Networks, vol. 5, pp. 3272–3277.
Peng, H., Long, F., Ding, C., August 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27 (8), 1226–1238, doi:10.1109/TPAMI.2005.159.
Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Elsevier Trends in Genetics 24 (3), 142–149.
Principe, J., 2010. Information Theoretic Learning – Renyi's Entropy and Kernel Perspectives. Springer.
Qin, J., Li, R., Raes, J., Arumuga, M., Burgdorf, K., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Lian, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S., Wang, J., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65.
Robin, S., Schbath, S., 2002. Numerical comparison of several approximations of the word count distribution in random sequences. Journal of Computational Biology 8 (4), 349–359.
Rosen, G.L., et al., November 2009. Signal processing for metagenomics: extracting information from the soup. Current Genomics 10, 493–510.
Rosen, G.L., Garbarine, E.M., Caseiro, D.A., Polikar, R., Sokhansanj, B.A., September 2008. Metagenome fragment classification using *N*-mer frequency profiles. Advances in Bioinformatics.
Rusch, D.B., et al., March 2007. The sorcerer ii global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biology 5 (3), 77.
Sandberg, R., et al., 2001. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. Genome Research 11 (8), 1404–1409.
Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glockner, F.O., 2004. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5 (163), doi:10.1186/1471-2105-5-163.
Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research 3, 1415–1438.
Venter, et al., 2004. Environmental genome shotgun sequencing of the Sargasso sea. Science 304 (5667), 66–74.
Wang, Q., Garrity, G., Tiedje, J.M., Cole, J.R., 2007. Naive Bayes classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied Environmental Microbiology, 5261–5267.
Wommack, K.E., Bhavsar, J., Ravel, J., 2008. Metagenomics: read length matters. Applied Environmental Microbiology 74 (5), 1453–1463.
Zhou, J., Peng, H., 2007. Automatic recognition and annotation of gene expression patterns of fly embryos. Bioinformatics 23 (5), 589–596.