Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Journal of Microbiological Methods 65 (2006) 49–62
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

An ecoinformatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA

Chengyong Yang [a], DeEtta Mills [b], Kalai Mathee [b], Yong Wang [a], Krish Jayachandran [c], Masoumeh Sikaroodi [d], Patrick Gillevet [d], Jim Entry [e], Giri Narasimhan [a,*]

[a] *Bioinformatics Research Group (BioRG), School of Computer Science, Florida International University, Miami, Florida, 33199, USA*
[b] *Department of Biological Sciences, Florida International University, Miami, Florida, USA*
[c] *Department of Environmental Sciences, Florida International University, Miami, Florida, USA*
[d] *Microbial and Environmental Biocomplexity, Department of Environmental Sciences and Policy, George Mason University, Manassas, Virginia, USA*
[e] *USDA Agricultural Research Service, Northwest Irrigation and Soils Research Laboratory, Kimberly, Idaho, USA*

Received 18 January 2005; received in revised form 22 April 2005; accepted 24 June 2005
Available online 27 July 2005

Abstract

Support vector machines (SVM) and *K-nearest neighbors* (KNN) are two computational machine learning tools that perform supervised classification. This paper presents a novel application of such supervised analytical tools for microbial community profiling and to distinguish patterning among ecosystems. Amplicon length heterogeneity (ALH) profiles from several hypervariable regions of *16S rRNA* gene of eubacterial communities from Idaho agricultural soil samples and from Chesapeake Bay marsh sediments were separately analyzed. The profiles from all available hypervariable regions were concatenated to obtain a *combined profile*, which was then provided to the SVM and KNN classifiers. Each profile was labeled with information about the location or time of its sampling. We hypothesized that after a learning phase using feature vectors from labeled ALH profiles, both these classifiers would have the capacity to predict the labels of previously unseen samples. The resulting classifiers were able to predict the labels of the Idaho soil samples with high accuracy. The classifiers were less accurate for the classification of the Chesapeake Bay sediments suggesting greater similarity within the Bay's microbial community patterns in the sampled sites. The profiles obtained from the V1+V2 region were more informative than that obtained from any other single region. However, combining them with profiles from the V1 region (with or without the profiles from the V3 region) resulted in the most accurate classification of the samples. The addition

* Corresponding author. Tel.: +1 305 348 3748; fax: +1 305 348 3549.
E-mail address: giri@cs.fiu.edu (G. Narasimhan).

0167-7012/$ - see front matter
doi:10.1016/j.mimet.2005.06.012

of profiles from the V9 region appeared to confound the classifiers. Our results show that SVM and KNN classifiers can be effectively applied to distinguish between eubacterial community patterns from different ecosystems based only on their ALH profiles.

Keywords: Ecoinformatics; Supervised classification; Machine learning; Amplicon length heterogeneity; Ecosystem community patterns; Support vector machines

1. Introduction

Microbial communities that occur in both natural and man-made environments can be complex, consisting of a large number of bacterial, archaeal, and fungal species. Thus, it is impractical to use culture-based microbiological methods for species identification. Understanding and analyzing at a whole-community level enables fast and efficient ways to provide a glimpse into the patterned diversity of such communities (Dunbar et al., 2002; Hill et al., 2002). Molecular methods based on amplification of DNA using polymerase chain reaction (PCR), cloning and sequencing of highly conserved prokaryotic target genes have played a central role in determining the extent of diversity. The predominant choice for a target gene has been the 16S small subunit ribosomal RNA (rRNA) (Olsen et al., 1986; Pace et al., 1986), resulting in the accumulation of extensive sequence information (e.g., the Ribosome Database Project (Maidak et al., 1999)).

Ribosomal RNA is essential for cellular growth, function, and survival of all organisms. Consequently, ribosomes have highly conserved functional domains that share high sequence identity. These conserved regions are interspersed with hypervariable sequence regions that are due to base substitutions, or insertions or deletions of short segments of nucleotides. These variations are phylogenetically relevant as they are related to the genetic makeup of each species (Ludwig and Schleifer, 1994). The natural variations and composition of *16S rRNA* have been exploited in molecular assays such as ***t***erminal ***r***estriction ***f***ragment ***l***ength ***p***olymorphism (TRFLP) and ***a***mplicon ***l***ength ***h***eterogeneity (ALH). These assays depend on the amplification of the variable regions of the *16S rRNA* (or any other appropriate gene) using sets of primers that are designed based on the highly conserved regions. The portion of the DNA sequence amplified by a pair of primers is referred to as an *amplicon*. Given a sample consisting of a community of microbes, PCR amplification using a pair of primers will yield a profile of amplicon lengths associated with the microorganisms in the sample, where the height (intensity) of the peak is proportional to the abundance of the amplicons associated with any given length (Dunbar et al., 2001; Suzuki et al., 1998). Different pairs of primers can be used to target different variable regions of the *16S rRNA* genes. We introduce the concept of a *combined profile*, which is simply a concatenation of the normalized ALH profiles obtained from using different pairs of primers on the same sample (analogous to multiple loci analysis). Thus, the ALH system profiles a community based on the patterns of lengths of amplified products (amplicons) providing a rapid and cost-effective way to distinguish among the communities without identifying individual species or genera. Length heterogeneity has been used to estimate bacterial diversity in a variety of ecosystems (Bernhard et al., 2005; Bernhard and Field, 2000; Litchfield and Gillevet, 2002; Mills et al., 2003; Ritchie et al., 2000; Suzuki et al., 1998; Tiirola et al., 2003).

Prior approaches to study soil microbial diversity and community dynamics include computing measures such as species richness and dominance or evenness indices (Hill et al., 2002). Theoretical models of microbial diversity based on the log-normal distributions have been studied (Dunbar et al., 2002). Clustering of soil samples using the UPGMA (unweighted pair-group method using arithmetic averages) algorithm based on the use of distance metrics (such as the Jaccards or Hellinger or Pearson distances) on length heterogeneity data has also been reported (Blackwood et al., 2003; Dunbar et al., 2000; Griffiths et al., 2000). Such unsupervised methods have been used to support claims that certain

relationships between communities can be discerned, that the groupings are natural, and that outliers can be identified.

In contrast to unsupervised methods, computational tools based on supervised classification methods from machine learning are not known to have been used for studying microbial diversity. Two well-known *supervised classification* tools include: (a) *Support Vector Machines* (SVM), and (b) *K-Nearest Neighbor Method* (KNN). These tools have the ability to "learn" to classify samples after being trained with a collection of known, labeled feature vectors obtained from the inputs. Both are computational machine-learning tools that treat the data as points or vectors in Euclidean space. These vectors are usually referred to as "feature vectors" because their coordinates correspond to quantified "features" of the data. These features are usually obtained after a feature extraction process. Given a new sample, it too is represented by a feature vector. In both methods, classification of the new sample is based on the location of its feature vector vis-à-vis the location of the labeled feature vectors. For further details, the reader is encouraged to consult the following references (Cristianini and Shawe-Taylor, 2000; Hastie et al., 2001; Michie et al., 1994; Noble, 2004). SVMs have been shown to perform well in a variety of research areas including pattern recognition (Burges, 1998), text categorization (Joachims, 1997), face recognition (Osuna et al., 1997), computer vision (Scholkopf et al., 1997), classifications based on microarray gene expression data (Brown et al., 2000; Furey et al., 2000; Lee and Lee, 2003; Sturn et al., 2002; Zheng et al., 2003), detecting remote protein homologies (Vert, 2002), classifying G-Protein coupled receptors (Karchin et al., 2002), predicting signal peptide cleavage site and predicting subcelluar localization prediction (Hua and Sun, 2001; Lin et al., 2002), and many more. In particular, SVMs are well suited for dealing with high-dimensional data (Cristianini and Shawe-Taylor, 2000; Noble, 2004). KNN classifiers have been successfully used in applications such as classification of handwritten digits and satellite image scenes (Michie et al., 1994).

In this paper, computational machine learning classifiers based on SVMs and KNNs were used to identify and compare different types of microbial communities. After a "learning" phase, the resulting classifiers were able to classify with high accuracy (according to pre-assigned labels): (1) a set of Idaho native sagebrush and agricultural soil samples, and (2) a set of Chesapeake Bay marsh sediments. Detailed studies using these tools revealed the limitations of the data and the minimum amount of information from ALH assays that were necessary to perform reliable classification in such soil samples.

2. Materials and methods

2.1. Data sets

Supervised classifications were performed on a collection of ALH combined profiles of eubacterial communities from Idaho agricultural soil samples and Chesapeake marsh sediment samples. The DNA extracted from the samples was PCR amplified as described previously (Mills et al., 2003) using four sets of fluorescently labeled universal eubacterial primers for the Idaho samples and one set for the Chesapeake samples. The *16S rRNA* gene primers for the four hypervariable regions were as follows: for region V1+V2, 6-FAM-27F and 355R (Suzuki et al., 1998); for region V1, 6-FAM-P1F and P1R (Cocolin et al., 2001); for region V3, HEX-338F and 518R (Cocolin et al., 2001); for region V9, NED-1055F and EC1392R (Cocolin et al., 2001).

2.1.1. Idaho soil samples

The soil samples from Idaho represented a (control) native sagebrush (NSB) soil and three different soil management practices (conservation tillage (CT), irrigated pasture (IP) and moldboard plowed (MP)). The NSB and CT samples were collected from depths between 0 and 5 cm, 5 and 15 cm and 15 and 30 cm. Due to the land use and tillage practice, the IP and MP soils tend to be homogeneous, and were therefore only sampled from depths between 0 and 30 cm. All samples were sieved and homogenized after collection. For each of the Idaho soil types, samples were collected from two or three different locations within each descriptive sample type. Finally, for each location, samples were divided into triplicates and ALH profiles were obtained on each individual replicate. For each replicate, the V1, V1+V2, V3, and V9 hypervariable regions were PCR amplified and analyzed by the ALH method.

The computational analyses were performed on two different sets of samples. The first set, referred to as *Idaho-top*, included soil samples from all NSB and CT locations obtained from depths of 0 to 5 cm (surface) and all IP and MP samples obtained from depths of 0 to 30 cm. The second set, referred to as *Idaho-deep*, included soil samples from all NSB and CT locations obtained from a depth of 15 to 30 cm (subsurface) and all IP and MP samples obtained from depths of 0 to 30 cm. To use the machine learning methods, feature vectors were extracted from the ALH combined profiles. These vectors contained one component for each possible length with the value of that component equal to the relative abundance (i.e., the intensity (amplitude) of each peak divided by the total intensity of all peaks). If the ALH profile of a particular sample had a peak missing (i.e., contained no amplicons of a specific length) when compared to others, then the corresponding component of its feature vector was set to zero. The samples were labeled according to the soil management practice used. The classifiers were designed to predict the labels for unknown samples.

2.1.2. Chesapeake Bay samples

Sediment samples from the barrier island fringe marsh in the Chesapeake Bay were separately analyzed. One data set consisted of samples from nine different locations within the coastal habitats (Chimney Pole, Cattle Shed, Hog Island Dry, Hog Island Intermediate, Hog Island Wet, Oyster Creek Bank, Oyster Creek Marsh, Red Bank, and Upper Phillips Creek), all collected at the same time of the year. Another data set consisted of samples from the Chesapeake Bay collected from a single location at seven different time points over a 14-month period (Sep 1999–Nov 2000). Two different classifiers were designed, one to optimally predict the location label of unknown test samples, and the other to predict the time of the year when the samples were collected.

As with the Idaho soil data, feature vectors were extracted from the ALH profiles. The data from the Chesapeake ALH profiles were only from the V1+V2 hypervariable region. As with the Idaho soil data analysis, if a particular ALH profile had a peak missing when compared to others, then the corresponding component of the feature vector was set to zero.

2.2. Supervised classification methods

The task of classification consisted of constructing a method that could automatically "label" the sample from combined ALH profile patterns. For every sample, this pattern was given as a vector of relative abundance at different lengths. Given a set of training examples, $X=\{x^i : x^i \in R^n\}$, with known labels, $Y=\{y^i : y^i \in \{\text{possible types}\}\}$, a discriminant function, $f: R^n \rightarrow \{\text{possible types}\}$, where n is the number of possible lengths, has to be learned. The number of misclassifications of f on the training set $\{X,Y\}$ is minimized by the learning machine during the training phase. The practical interest of these methods is their capacity to predict the class of previously unseen samples (test set), i.e., the so-called "generalization" performance. The data samples in any given data set were divided into a *training set* and a *test set*. This was done so that no repeats from the same location or sampling time were present in both the training and the test set. Otherwise, the SVM classifier would have been trained with a very similar training sample and it would be easy to build highly reliable classifiers for the test samples. Such a strategy for dividing input samples into training and test sets is used in *k-fold cross validation* techniques and is, therefore, statistically sound (Efron and Tibshirani, 1993), allowing us to train and test on different samples without the need for unknown environmental samples whose labels may be uncertain.

The major problem of training a learning machine to perform supervised classification is to find a function that not only captures the essential properties of the data distribution, but also avoids over-fitting the data. The support vector machine (SVM) tries to construct a (linear) discriminant function for the data points in feature space in such a way that the feature vectors of the training samples are separated into classes, while simultaneously maximizing the distance of the discriminant function from the nearest training set feature vector. SVM classifiers also allow for non-linear discriminant functions. This is achieved by mapping the input vectors into a different feature space using a mapping function, $\Phi: x_i \rightarrow \Phi(x_i)$, and using the vectors, $\Phi(x_i), x_i \in X$, as the feature vectors. The corresponding *kernel function* used by the SVM algorithm is $K(x_i, x_k)=<\Phi(x_i) \cdot \Phi(x_k)>$. Standard kernel functions include: (a) the *polynomial kernel function* of degree

d given by $K(X, Y)=(X\cdot Y+1)^d$, which for $d=1$ is the linear kernel function, (b) the *radial basis function* (RBF) kernel with parameter γ, given by $K(X, Y)=\exp(-\gamma\|X-Y\|^2)$, and (c) the *sigmoid* kernel given by $K(X, Y)=\tanh(\gamma(X\cdot Y)+\theta)$. Default parameters for each kernel function were applied for the learning and testing phase of the SVM classifier. However, kernels with default parameters did not perform well for some of the data analyzed here. In such cases, model selection is recommended (Chang and Lin, 2002), which requires performing a "grid-search" on exponentially growing sequences of values of C and γ, and picking the one with the minimum k-fold cross validation error. The penalty parameter, C, is part of the error term in the SVM and represents the rate at which the SVM "learns" from the misclassifications. Varying the parameter, γ, which is relevant only for the sigmoid and the radial basis function (RBF) kernels, helps in trying a range of different kernel functions. For the model selection, we performed a grid search with log C and log γ taking values in the range −25 through 25. As recommended (Chang and Lin, 2002), we started with a coarse grid, searching for the optimal values of log C and log γ in the range from −25 through 25 with a step size of 5, after which the step size was reduced to 1. A final search was conducted with a step size of 0.25. The pair of values of C and γ with the highest average cross-validation accuracies were selected and used to train the whole training set and to generate the final model.

KNN classifiers are memory-based, and do not require optimizing any of the parameters. Given a query point x_0, the k training points x_r, $r=1,\ldots,k$, closest in distance to x_0 are used to classify using a majority vote among the k neighbors (ties are broken at random). Euclidean distance was used as a measure of distance.

2.3. Design and implementation of the classifiers

Many implementations of SVMs are currently available, including mySVM (Rüping, 2002), svmTorch (Collobert and Bengio, 2001), SVMLight (Joachims, 1999), Gist (Pavlidis et al., 2004), and LibSVM (Chang and Lin, 2001). We used the LibSVM package, available from http://www.csie.ntu.edu.tw/~cjlin/libsvm (free for academic use). The core optimization method in LibSVM is based on a decomposition method (Joachims, 1999). Once the SVM classifier is built, classification of unknown test samples is efficient and rapid since the software only calculates the inner products between the test sample and a small subset of feature vectors known as the support vectors. The multi-class classification was implemented by the "one-against-one" approach (Knerr et al., 1990) in which $k(k-1)/2$ pair-wise classifiers (assuming a total of k classes) were constructed and each classifier was used to train samples from a pair of classes. A voting strategy was used, in which each pair-wise classification gave a vote to the winning class. The final classification was the class with the maximum number of votes. Ties were broken by picking the class with the smaller index (Hsu and Lin, 2002). The KNN classifier was implemented using the Java programming language.

2.4. Evaluating the accuracy of the classifiers

For the testing phase, the prediction performance was evaluated using the *jackknife test* (Efron and Tibshirani, 1993); each sample (including all its replicates) was singled out in turn as test samples, and the remaining samples were used to train the classifiers. All replicates of the sample were pooled together for testing in order to avoid biasing the training set. All tests and their outputs were run through an independent "batch" program (written in Java) that invoked the LibSVM software package. To evaluate the predictive ability of the classifiers, the following measures were calculated: (a) the total prediction accuracy (TPA), given by $\mathrm{TPA}=\sum_{i=1}^{k} p(i)/N$, (b) the prediction accuracy (PA), given by $\mathrm{PA}(i)=p(i)/\mathrm{obs}(i)$, and (c) the Matthew's Correlation Coefficient (MCC) (Matthews, 1975), given by

$$\mathrm{MCC}(i)=\frac{p(i)n(i)-u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}}.$$

Here, N is the total number of amplicons; k is the number of classes; obs(i) is the number of amplicons observed in location i; $p(i)$ is the number of correctly predicted samples of class i; $n(i)$ is the number of

correctly predicted samples not of class location i; $u(i)$ is the number of false negatives; and $o(i)$ is the number of false positives.

The accuracy value is a measure of the number of correct classifications. The value PA(i) measures the accuracy for a specific class i, while TPA measures the quantity for all the classes and is therefore a measure of the accuracy of the classification for the whole data set. Note that PA(i) will be 100% if all the samples in set i are correctly classified. However, the MCC value for that set could be less than the optimal value of 1.0 even if the accuracy is 100%. The MCC value also takes into account samples from outside this set that are misclassified as belonging to this set, and is, therefore, a rough measure of selectivity.

3. Results

3.1. Prediction accuracy for the Idaho soil samples

The prediction accuracies and MCC values were calculated for Idaho top and deep soil samples (Table 1). The total accuracy for the top soil samples using an SVM classifier was 96.67% (Table 1A) even with the simplest linear kernel function. The accuracy was kept constant when applying the more complex non-linear kernel function. The total accuracy for the deep soil samples was 88.57% (Table 1B). One sample point from the MP data set was consistently misclassified as a CT sample, suggesting that it may be an outlier. The misclassifications with the deep soil data were not consistent in any manner (one NSB sample classified as CT, one CT sample classified as IP, and one MP sample classified as CT). With more misclassifications, prediction accuracies were lower for deep soil samples, suggesting that deep soil samples were less distinguishable than the top soil samples.

Although the accuracy and MCC values for each classifier were strongly correlated, they were not identical. For example, an accuracy of 88.89% for MP (Table 1A) indicates that one out of the nine MP samples was misclassified as belonging to some other class. For CT samples tested with an SVM classifier using a linear kernel (Table 1A), the accuracy was 100%, while the MCC value was 0.91, implying that while all the CT samples were correctly classified, some other sample was incorrectly

Table 1
Prediction accuracies and MCC values for Idaho top (A) and deep (B) soils with K-Nearest Neighbor Method (KNN) and Support Vector Machines (SVM) classifiers using amplicon length heterogeneity (ALH) profiles from four 16S rRNA hypervariable regions (V1, V1+V2, V3 and V9)

Location	Number of samples	KNN	SVM Classifiers					
			Linear kernel		RBF kernel		Sigmoid kernel	
			Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
A: Idaho top soils								
NSB	6	100	100	1.00	100	1.00	100	1.00
CT	7	100	100	0.91	100	0.91	100	0.91
IP	8	100	100	1.00	100	1.00	100	1.00
MP	9	88.89	88.89	0.92	88.89	0.92	88.89	0.92
Overall accuracy	30	96.67	96.67		96.67		96.67	
B: Idaho deep soils								
NSB	9	88.89	77.77	0.85	77.77	0.85	77.77	0.85
CT	9	100	88.89	0.78	88.89	0.78	88.89	0.78
IP	8	100	100	0.92	100	0.92	100	0.92
MP	9	100	88.89	0.85	88.89	0.85	88.89	0.85
Overall accuracy	35	97.14	88.57		88.57		88.57	

The KNN classifier was implemented with $k=1$. For the SVM classifier, three different kernel functions, the linear function when $d=1$, the *radial basis function* (RBF) kernel with parameter γ, and a *sigmoid* kernel were tested. See Materials and methods for details. The soil samples from Idaho represented the following four soil management types: pristine natural sagebrush (NSB), conservation tillage (CT), irrigated pasture (IP) and moldboard plough (MP).

labeled as CT (in this case, one of the MP samples). The errors in the classification are small, yet non-trivial. An improved analysis that achieves 100% accuracy for the SVM classifiers for the same data set is presented below. The KNN classifiers outperformed the SVM classifiers for the deep soil samples, while they were evenly matched for the top soil samples.

3.2. Prediction accuracies for Chesapeake Bay samples

The procedure described above for the Idaho soil samples was independently applied to the Chesapeake Bay samples with the location-based and time-based labels (Table 2). The performance of the classifiers for the Chesapeake Bay data was clearly inferior to that for the Idaho samples. The average accuracy was only about 83%, with the accuracy for the individual classes ranging from 55% to 100% (Table 2). The MCC values ranged from 0.62 to 0.91. The performance of the SVM and KNN classifiers were comparable.

3.3. Optimization of the SVM classifier for Chesapeake Bay samples using model selection

Performance by the SVM classifier on the Chesapeake Bay sediments was less than satisfactory, requiring further optimization. Optimization of the SVM classifier was done using *model selection* for the various kernel function parameters and the penalty parameters (Chang and Lin, 2002). The penalty parameter, C, is part of the error term in the SVM that represents the rate at which the SVM "learns" from the misclassifications. A range of different kernel functions for the sigmoid and the RBF kernels can be explored by varying the parameter γ. Model selection suggested the use of a SVM classifier using a RBF kernel function with $\log_2 C=9$ and $\log_2 \gamma=2.15$ for Chesapeake Bay location-based classification, and

Table 2
Prediction accuracies and MCC values for location-based and time-based classifications of Chesapeake Bay samples with KNN and SVM classifiers using amplicon length heterogeneity (ALH) profiles from a single 16S rRNA hypervariable region (V1+V2)

	Sample labels	Number of samples	KNN	SVM with RBF kernel	
				Accuracy (%)	MCC
Samples from different locations	CP	23	86.95	91.30	0.85
	CS	60	90.00	91.67	0.85
	HD	52	84.62	90.38	0.86
	HI	38	76.31	84.21	0.82
	HW	42	83.33	83.33	0.80
	OC	23	82.60	78.26	0.76
	OM	9	55.56	55.56	0.62
	RB	30	70.00	76.67	0.84
	UP	5	60.00	60.00	0.77
	Overall accuracy (location)	282	81.56	84.75	
Samples from different times of year	Sep 99	76	85.52	88.16	0.82
	Dec 99	58	79.31	75.86	0.68
	Feb 00	50	80.00	80.00	0.81
	Mar 00	33	78.78	81.82	0.78
	May 00	15	86.66	80.00	0.89
	July 00	15	100.00	100.00	0.91
	Nov 00	35	85.71	80.00	0.77
	Overall accuracy (time)	282	83.33	82.62	

The KNN classifiers were implemented with $k=1$. For the SVM classifier, only the *radial basis function* (RBF) kernel was used. After model selection, parameters $\log_2 C=9$ and $\log_2 \gamma=2.15$ were used for the location experiments, while $\log_2 C=7.25$ and $\log_2 \gamma=0.75$ were used for the time experiments. The soil samples from Chesapeake Bay represented the following locations: CP=Chimney Pole, CS=Cattle Shed, HD=Hog Island Dry, HI=Hog Island Intermediate, HW=Hog Island Wet, OC=Oyster Creek Bank, OM=Oyster Creek Marsh, RB=Red Bank, UP=Upper Phillips Creek.

with $\log_2 C = 7.25$ and $\log_2 \gamma = 0.75$ for Chesapeake Bay time-based classification.

The overall accuracy (after optimization) for the location-based samples was 84.75%. The 49 misclassifications did not appear to have any perceivable pattern to them. This would suggest that the eubacterial community patterns in the Chesapeake Bay sediments are spatially similar at the resolution of the ALH profile. Alternatively, the fact that we had only one hypervariable region in the input may have made it less distinguishable. We also tried alternative ways to group the location-based samples—by dividing them into three coastal habitats, high dry *Spartina* marsh, low wet *Spartina* marsh, and adjacent mud flats. However, this grouping did not significantly change the performance of the classifiers (data not shown), suggesting that the tidal flux in the system is high enough to eliminate any distinguishing features in the eubacterial communities.

The overall accuracy (after optimization) for the grouped time-based samples was 82.62%, and was therefore comparable to that of the location-based samples (84.75%). However, the misclassifications of the time-based samples were mostly between adjacent time periods. In fact, 40 out of the 51 misclassifications were to time labels that were within three months of the correct label. This led us to question whether a more accurate classifier could be built to distinguish samples that were sufficiently far apart (seasonal differences) in their time labels. For example, when we built classifiers trained only with samples from July and December, the resulting SVM classifier was accurate with 94.52% of the test samples with those time labels. Similar results were observed for samples from March and September (90.83% accuracy), and from May and November (92% accuracy).

3.4. Significance of 16S rRNA hypervariable regions

The poorer performance with the Chesapeake Bay data, which interrogated only one hypervariable region, raised several questions about the relative significance of the ALH data from the different hypervariable regions of 16S rRNA. Since ALH profiles from all the four regions were available for the Idaho soil samples (Table 1), we sought to determine the combinations of regions that would provide the most amount of information in terms of the ability to distinguish soil samples. This question was addressed by determining the accuracy of the resulting classifiers when trained with Idaho soil ALH profile data from every possible combination of the four regions. Since the number of regions for which ALH assays are done determines the cost of the experiments, this analysis could also shed light on the tradeoff between cost and accuracy.

The prediction accuracies were calculated when data from different combinations of regions were used to train both, a KNN classifier, and a SVM classifier with RBF kernel function and a KNN classifier after optimized model selection (Tables 3–5). When the profiles from only one region were utilized to design classifiers (Table 3), the performance was best when the V1+V2 region was used. For Idaho soil samples, prediction accuracies were equally good using the V1 regions, with the exception of the deep soil samples, where the SVM classifier performance on the V1+V2 region was marginally better than that for V1 region. For both top and deep soils, the worst results were obtained with the V9 region. The accuracies varied considerably with soil management types. For example, the NSB soil samples were best distinguished by the SVM and KNN classifiers using the profiles from the V1+V2 region, or by the KNN classifier using the profiles from the V1 region. The classifiers using the V9 region profiles were only successful in distinguishing the CT and IP top soil samples, and performed poorly otherwise. In fact, the classifiers using profiles from only the V9 region had an overall accuracy of about 80%. Interestingly, the SVM classifiers using any of the four regions made no misclassifications for the IP soil samples.

When two variable regions were used to design classifiers (Table 4), using a combination of V1 and V1+V2 regions improved the performance (for both top and deep soil samples) over the classifiers using only one of the regions. For Idaho top soil samples, prediction accuracies were equally good with a combination of V1 and V3 regions. The performance of classifiers that included the V9 region was markedly worse than when this region was excluded (Tables 3–5). All other classifiers had reasonably high accuracies. When three variable regions (V1, V1+V2 and V3) were used (Table 5), the performance was good for all samples.

The data seems to imply that region V9 generated data that tends to confound both the classifiers espe-

Table 3
Prediction accuracies for Idaho top (A) and deep (B) soil samples using KNN and SVM classifiers with the radial basis kernel function (with model selection) using ALH profiles from single 16S rRNA hypervariable regions

Location	Number of samples	16S rRNA hypervariable region utilized							
		V1		V1+V2		V3		V9	
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
A: Idaho top soils									
NSB	6	100.00	100.00	100.00	100.00	100.00	100.00	33.33	33.33
CT	7	85.71	85.71	85.71	85.71	85.71	85.71	100.00	100.00
IP	8	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MP	9	100.00	100.00	100.00	100.00	88.89	88.89	77.78	66.67
Overall accuracy	30	96.67	96.67	96.67	96.67	93.33	93.33	80.00	76.67
B: Idaho deep soils									
NSB	9	100.00	88.89	100.00	100.00	88.89	77.78	77.78	66.67
CT	9	100.00	100.00	100.00	100.00	77.78	88.89	77.78	88.89
IP	8	100.00	100.00	100.00	100.00	87.50	100.00	75.00	100.00
MP	9	100.00	88.89	100.00	88.89	100.00	88.89	88.89	66.67
Overall accuracy	35	100.00	94.29	100.00	97.14	88.57	88.57	80.00	80.00

The soil samples from Idaho represented the following four soil management types: pristine natural sagebrush (NSB), conservation tillage (CT), irrigated pasture (IP) and moldboard plough (MP). Sizes of the feature vectors for the four regions for the Idaho top soils were as follows: V1: 23; V1+V2: 31; V3: 11 and V9: 5. Sizes of the feature vectors for the four regions for the Idaho deep soils were as follows: V1: 24; V1+V2: 34; V3: 14 and V9: 7.

cially when used by itself (Table 3) or in combination with one of the other regions (Tables 4 and 5). However, when three regions were combined, the inclusion of V9 region had no influence on the top soil samples when combined with V1 and V1+V2 (Table 5). Interestingly, the worst performance of the classifiers

Table 4
Prediction accuracies for Idaho top (A) and deep soil (B) samples using KNN and SVM classifiers with linear kernel function using ALH profiles from combination of pairs of 16S rRNA hypervariable regions

Type	Number of samples	16S rRNA hypervariable region utilized											
		[V1, V1+V2]		[V1, V3]		[V1, V9]		[V1+V2, V3]		[V1+V2, V9]		[V3, V9]	
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
A: Idaho top soils													
NSB	6	100.00	100.00	100.00	100.00	83.33	100.00	100.00	100.00	83.33	100.00	83.33	83.33
CT	7	85.71	100.00	100.00	100.00	100.00	85.71	85.71	85.71	100.00	100.00	100.00	85.71
IP	8	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MP	9	100.00	100.00	100.00	100.00	100.00	88.89	88.89	88.89	100.00	66.67	88.89	66.67
Overall accuracy	30	96.67	100.00	100.00	100.00	96.67	93.33	93.33	93.33	96.67	90.00	93.33	83.33
B: Idaho deep soils													
NSB	9	100.00	100.00	100.00	88.89	88.89	88.89	100.00	88.89	77.78	100.00	77.78	66.67
CT	9	100.00	100.00	100.00	100.00	88.89	100.00	100.00	100.00	88.89	88.89	77.78	88.89
IP	8	100.00	100.00	100.00	100.00	87.50	100.00	100.00	100.00	100.00	100.00	87.50	87.50
MP	9	100.00	100.00	100.00	88.89	88.89	88.89	100.00	100.00	88.89	88.89	88.89	88.89
Overall accuracy	35	100.00	100.00	100.00	94.29	88.57	94.29	100.00	97.14	88.57	94.29	82.86	82.86

The soil samples from Idaho represented the following four soil management types: pristine natural sagebrush (NSB), conservation tillage (CT), irrigated pasture (IP) and moldboard plough (MP). Sizes of the feature vectors for the four regions for the Idaho top soils were as follows: V1: 23; V1+V2: 31; V3: 11 and V9: 5. Sizes of the feature vectors for the four regions for the Idaho deep soils were as follows: V1: 24; V1+V2: 34; V3: 14 and V9: 7.

Table 5
Prediction accuracies for Idaho top (A) and deep soil (B) samples using KNN and SVM classifiers with linear kernel function using ALH profiles from combination of triples of 16S rRNA hypervariable regions

Type	Number of samples	16S rRNA hypervariable region utilized							
		[V1, V1+V2, V3]		[V1, V1+V2, V9]		[V1+V2, V3, V9]		[V1, V3, V9]	
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
A: Idaho top soils									
NSB	6	100.00	100.00	100.00	100.00	83.33	100.00	83.33	100.00
CT	7	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
IP	8	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MP	9	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Overall accuracy	30	100.00	100.00	100.00	100.00	96.67	100.00	96.67	100.00
B: Idaho deep soils									
NSB	9	100.00	100.00	88.89	100.00	77.78	88.89	88.89	88.89
CT	9	100.00	100.00	77.79	100.00	88.89	100.00	100.00	100.00
IP	8	100.00	100.00	87.50	100.00	100.00	100.00	100.00	100.00
MP	9	100.00	100.00	88.89	100.00	88.89	100.00	88.89	88.89
Overall accuracy	35	100.00	100.00	85.71	100.00	88.57	97.14	94.29	94.29

The soil samples from Idaho represented the following four soil management types: pristine natural sagebrush (NSB), conservation tillage (CT), irrigated pasture (IP) and moldboard plough (MP). Sizes of the feature vectors for the four regions for the Idaho top soils were as follows: V1: 23; V1+V2: 31; V3: 11 and V9: 5. Sizes of the feature vectors for the four regions for the Idaho deep soils were as follows: V1: 24; V1+V2: 34; V3: 14 and V9: 7.

were observed when all four regions (Table 1) or when single regions (Table 3) were used. The KNN classifier performed better (96.69%) than the SVM classifier (92.62%) when all four regions used. However, when a combination of three regions was used, KNN performed relatively poorly with deep soil samples, especially when region V9 was included.

4. Discussion

4.1. Effective classification of ALH profiles using computational tools

One attractive property of SVMs is that it condenses information in the training samples to provide a sparse representation using a linear combination of a small number of samples, referred to as the support vectors, and only these vectors are used in the subsequent classification. The number of support vectors is typically small compared to the total number of training samples. This makes the classification task very efficient even when analyzing large datasets containing many uninformative data points. The training and optimization phases for SVMs includes the selection of an appropriate kernel function, selection of function parameters and the regulation parameter, C. The function parameters implicitly define the structure of the mapped feature space, while C controls the learning rate, thereby affecting the training speed. The results show that comparable accuracies were obtained with different types of kernels (Table 1). Large variations of the parameters including γ for the RBF kernel had little influence on the classification performance.

Both the classifiers (SVM and KNN) performed well for Idaho soil samples. In particular, the SVMs exhibited flawless performance for the Idaho soil samples. Our results suggest that the top soil samples are more clearly distinguishable than deep soil samples, confirming the conclusions of other researchers (Griffiths et al., 2000). This is not surprising since the surface soil would tend to be more heterogeneous due to soil mixing from wind and/or rain erosion. Furthermore, the importation of new community members from allochthonous sources would be more likely to impact the top soil layers than the deeper ones.

Although there was no perceptible difference in the performance of the SVM and KNN classifiers, when we looked at the aggregation of all the results, we found that the SVM classifier exhibited a marginally superior performance with an average overall accuracy of about 92%, as compared to 91% for the

KNN classifier. The standard deviation was also smaller with the SVM classifier, suggesting a more consistent performance than the KNN classifier. However, note that the KNN tool can be implemented more easily than the more sophisticated SVM tool.

It may be argued that the computational tools of the type presented here assume that the majority of ALH amplicons are common and detectable across a wide range of samples of the same type. It is not clear if such an assumption is justified. However, the strong performance of these predictors on at least some of the soil types (e.g., natural sage brush, or irrigated pastures), even though the sampling was done at two or three different locations, lends support to such a hypothesis. Since both the SVM and the KNN classifiers can easily deal with high dimensional data, it is possible to extend the analyses to incorporate other useful features that may improve the prediction accuracy (i.e., physical and chemical parameters of the samples such as pH, salinity, temperature, mineral and nutrient concentrations).

4.2. Analysis of Chesapeake Bay samples

The ALH profile data from the sediment samples from Chesapeake Bay were only from the V1+V2 hypervariable region of the *16S rRNA* gene. The resulting classifiers did not perform as well as the ones with the data from the Idaho soils, which had ALH profiles from four hypervariable regions of the *16S rRNA* gene. It is likely that ALH profiles from only one hypervariable region (V1+V2) is not sufficient for good classifications. Many other factors could have contributed to the difficulty in classification. The lower performance may be due to the imbalance in the size of the data sets in the sense that the ratio of the size of the largest class to the size of the smallest class is 60/5=12 for the location-based labels and 76/15=5 for the time-based labels. It may also be due to the fact that the community patterns of the Chesapeake Bay samples were less distinguishable from each other than the corresponding Idaho samples. Since the Bay samples came from undisturbed and similar *Spartina*-dominated marsh sediments compared to the range of plant (native sagebrushes to crops like potatoes or alfalfa) and management systems (pasture to moldboard plowed) in the Idaho soil, it is not surprising that the community patterns were similar between sites. Sediments tend to be saturated most of the time driving the community structure to those members that can best survive or adapt to fluctuating anoxic conditions. Another reason could be that a dense cover of *Spartina* marsh grasses found at most of the sampling sites may be driving the structure of the eubacterial communities associated with the life cycle of the plants. In a related study, Hines and coworkers showed that seasonal changes in the biogeochemical parameters in *Spartina* marsh sediments of New Hampshire were aligned to the growth phases of the marsh grasses (Rooney-Varga et al., 1997). While the relative abundance of the sulfate-reducing bacterial community members fluctuated over time, members of the Desulfobacteriaceae were found throughout the year. The dynamics of the marsh community appears to be driven by the growth cycle and physiology of the *Spartina* rather than by sediment temperature (Rooney-Varga et al., 1997). Therefore, it is possible that the Chesapeake Bay eubacterial communities were reflecting similar trends toward structural homogeneity. Recent analyses from the Gillevet laboratory of clone libraries from representative samples used in this study indicate significant overlap in eubacterial communities in all sample sites in their Chesapeake Bay study (personal communications).

Several interesting observations are possible from the analysis of the performance of the location-based and time-based classifiers. Even though samples obtained within a short span of time were not very distinguishable, samples that were obtained about six months apart were sufficiently distinguishable. Thus, spatial differences across sites were not as pronounced as temporal changes within a site, which could impact how sampling of sediments should be performed for a reliable study of changes in eubacterial diversity. This suggests that some environmental factor other than sediment saturation, tidal washing, or anoxia may be driving the community structure. Temperature is unlikely to be a factor since the overall sediment temperatures did not fluctuate greatly. The driving force could be the growth of the *Spartina* marsh grasses in the summer and their death and decay in the winter. The carbon and nutrient influx into this ecosystem is largely due to the decay of the *Spartina* plants, which may influence the resulting nutrient status, and subsequently the eubacterial community composition.

Since the Chesapeake Bay study was focused on biodiversity within the microbial communities, no environmental parameters were included in the analyses (P. Gillevet, personal communication).

4.3. Which combinations of 16S rRNA hypervariable regions are most informative?

The data suggests that a combination of profiles from the different regions complement each other and help to produce better classifiers for the whole community profiles (Tables 3–5). While it may seem intuitive that including more regions should improve the accuracy, this was not always true. Our results suggest that combining profiles from V1 and V1+V2 regions gave the most accurate classifiers (100% for both top and deep soils). Adding the profiles from region V3 to those from V1 and V1+V2 also gave 100% accuracy (Table 5), while adding the data from region V9 lowered the accuracy considerably (Table 1). We propose that profiles from different regions may be useful in different applications. For example, to identify the location of a soil sample based on its ALH pattern, the combined profiles from V1 and V1+V2 appear to be sufficient. On the other hand, to understand the eubacterial diversity of the whole community, the V3 region profiles should be included since they showed high variations even within an ecosystem. It is not clear whether the V9 region profiles provide any added value to understand eubacterial diversity.

5. Conclusions

Microbial community profiling and their utilization to distinguish patterning among microbial ecosystems is a novel application for supervised learning techniques such as SVMs and KNNs. Classification tools based on these machine learning techniques worked well for classifying or distinguishing soil samples based on their ALH patterns. For best results it seemed necessary to combine the profiles from several hypervariable regions of *16S rRNA* genes. In particular, the profiles from region V1+V2 were sufficient to distinguish between samples sampled at different times of the year. However, they were inadequate in distinguishing sediment communities sampled at the same time but at different locations in the Chesapeake Bay marshes. For the Idaho soil samples, a combination of profiles from V1 and V1+V2 regions provided the best results (100% accuracy). Profiles from region V3 may be included without any loss of accuracy. Including region V9 seemed to decrease the accuracy of the resulting classifiers.

The fact that two different software tools were able to learn from the data and successfully classify and discriminate between length heterogeneity profiles of new soil samples, indicates that there are hidden patterns in these profiles that can be discerned by these mathematical-based tools. This work paves the way for other classification tools to be tried on similar microbial ecology data. It is also anticipated that the computational tools developed here will be useful for large-scale and comparative analyses of ecogenomic data. Potential applications also exist in forensic science (Horswell et al., 2002) and environmental studies (Litchfield and Gillevet, 2002; Mills et al., 2003; Ritchie et al., 2000; Suzuki et al., 1998; Tiirola et al., 2003). The field of microbial ecology could benefit enormously by the development of classification tools of the type described in this paper.

Acknowledgements

We thank Dr. Chih-Jen Lin for help with the model selection. We also thank Ramakrishna Ruthala, Sasha Miller, and Sheria King for assistance.

References

Bernhard, A.E., Field, K.G., 2000. Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic markers from fecal anaerobes. Appl. Environ. Microbiol. 66 (4), 1587–1594.

Bernhard, A.E., Colbert, D., McManus, J., Field, K.G., 2005. Microbial community dynamics based on *16S rRNA* gene profiles in a Pacific Northwest estuary and its tributaries. FEMS Microbiol. Ecol. 52, 115–128.

Blackwood, C.B., Marsh, T., Kim, S.H., Paul, E.A., 2003. Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. Appl. Environ. Microbiol. 69 (2), 926–932.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., et al., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. U. S. A. 97 (1), 262–267.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discov. 2 (2), 121–167.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chang, C.-C., Lin, C.-J., 2002. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Cocolin, L., Manzano, M., Cantoni, C., Comi, G., 2001. Denaturing gradient gel electrophoresis analysis of the *16S rRNA* gene V1 region to monitor dynamic changes in the bacterial population during fermentation of Italian sausages. Appl. Environ. Microbiol. 67 (11), 5113–5121.

Collobert, R., Bengio, S., 2001. Svmtorch: support vector machines for large-scale regression problems. J. Mach. Learn. Res. 1, 143–160.

Cristianini, N., Shawe-Taylor, J., 2000. Support Vector Machines. Cambridge University Press.

Dunbar, J., Ticknor, L.O., Kuske, C.R., 2000. Assessment of microbial diversity in four Southwestern United States soils by *16S rRNA* gene terminal restriction fragment analysis. Appl. Environ. Microbiol. 66 (7), 2943–2950.

Dunbar, J., Ticknor, L.O., Kuske, C.R., 2001. Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of *16S rRNA* genes from bacterial communities. Appl. Environ. Microbiol. 67 (1), 190–197.

Dunbar, J., Barns, S.M., Ticknor, L.O., Kuske, C.R., 2002. Empirical and theoretical bacterial diversity in four Arizona soils. Appl. Environ. Microbiol. 68 (6), 3035–3045.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York, London.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16 (10), 906–914.

Griffiths, R.I., Whiteley, A.S., O'Donnell, A.G., Bailey, M.J., 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. Appl. Environ. Microbiol. 66 (12), 5488–5491.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hill, J.E., Seipp, R.P., Betts, M., Hawkins, L., Van Kessel, A.G., Crosby, W.L., et al., 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. Appl. Environ. Microbiol. 68 (6), 3055–3066.

Horswell, J., Cordiner, S.J., Maas, E.W., Martin, T.M., Sutherland, K.B.W., Speir, T., et al., 2002. Forensic comparison of soils by bacterial community DNA profiling. J. Forensic Sci. 47 (2), 350–353.

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425.

Hua, S., Sun, Z., 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17 (8), 721–728.

Joachims, T., 1997. Text categorization with support vector machines: learning with many relevant features. Proc of the 10th European Conference on Machine Learning, ECML-98.

Joachims, T., 1999. Making Large-Scale SVM Learning Practical. MIT-Press, Cambridge, MA.

Karchin, R., Karplus, K., Haussler, D., 2002. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18 (1), 147–159.

Knerr, S., Personnaz, L., Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for binding and training a neural network. Proc of Neurocomputing Algorithms, Architectures and Applications. Springer-Verlag.

Lee, Y., Lee, C.K., 2003. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19 (9), 1132–1139.

Lin, K., Kuang, Y., Joseph, J.S., Kolatkar, P.R., 2002. Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics. Nucleic Acids Res. 30 (11), 2599–2607.

Litchfield, C.D., Gillevet, P.M., 2002. Microbial diversity and complexity in hypersaline environments: a preliminary assessment. J. Ind. Microbiol. 28, 48–55.

Ludwig, W., Schleifer, K.H., 1994. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol. Rev. 15 (2-3), 155–173.

Maidak, B.L., Cole, J.R., Parker Jr., C.T., Garrity, G.M., Larsen, N., Li, B., et al., 1999. A new version of the RDP (Ribosomal Database Project). Nucleic Acids Res. 27 (1), 171–173.

Matthews, B.W., 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

Michie, D., Spiegelhalter, D., Taylor, C. (Eds.), 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.

Mills, D., Fitzgerald, K., Litchfield, C., Gillevet, P., 2003. A comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum-contaminated soils. J. Microbiol. Methods 54 (1), 57–74.

Noble, W.S., 2004. Kernel Methods in Computational Biology. MIT Press, Cambridge, MA.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., Stahl, D.A., 1986. Microbial ecology and evolution: a ribosomal RNA approach. Annu. Rev. Microbiol. 40, 337–365.

Osuna, E., Freund, R., Girosi, F., 1997. Training support vector machines: an application to face detection. Proc of the IEEE Conf. Computer Vision and Pattern Recognition.

Pace, N.R., Olsen, G.J., Woese, C.R., 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. Cell 45 (3), 325–326.

Pavlidis, P., Wapinski, I., Noble, W.S., 2004. Support vector machine classification on the web. Bioinformatics 20 (4), 586–587.

Ritchie, N.J., Schutter, M.E., Dick, R.P., Myrold, D.D., 2000. Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil. Appl. Environ. Microbiol. 66 (4), 1668–1675.

Rooney-Varga, J.N., Devereux, R., Evans, R.S., Hines, M.E., 1997. Seasonal changes in the relative abundance of uncultivated sulfate-reducing bacteria in a salt marsh sediment and in the rhizosphere of *Spartina alterniflora*. Appl. Environ. Microbiol. 63 (10), 3895–3901.

Rüping, S., 2002. mySVM. http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/.

Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., et al., 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans. Signal Process. 45 (11), 2758–2765.

Sturn, A., Quackenbush, J., Trajanoski, Z., 2002. Genesis: cluster analysis of microarray data. Bioinformatics 18 (1), 207–208.

Suzuki, M., Rappe, M.S., Giovannoni, S.J., 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit *rRNA* gene PCR amplicon length heterogeneity. Appl. Environ. Microbiol. 64 (11), 4522–4529.

Tiirola, M.A., Suvilampi, J.E., Kulomaa, M.S., Rintala, J.A., 2003. Microbial diversity in a thermophilic aerobic biofilm process: analysis by length heterogeneity PCR (LH-PCR). Water Res. 37, 2259–2269.

Vert, J.-P., 2002. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. Proc Pacific Symp Biocomputing.

Zheng, G., George, E.O., Narasimhan, G., 2003. Neural network classifiers and gene selection methods for microarray data on human lung adenocarcinoma. Proc Critical Assessment of Microarray Data (CAMDA), Raleigh, NC.