Poster Number: R-043

Presenter: Chengyong Yang **E-mail:** cyang01@cs.fiu.edu

An Ecoinformatics Tool for Microbial Diversity Studies: Supervised Classification of ALH Profiles of 16S rRNA

C. Yang[1],Y. Wang[1], D. Mills[1], K. Mathee[1], K. Jayachandran[1],P. Gillevet[2], J. Entry[3], G. Narasimhan[1]

[1]Florida International University, Miami, FL; [2]George Mason University, Manassas, VA; [3]USDA Agricultural Research Service, North Irrigation and Soils Research Laboratory, Kimberly, ID.

A Abstract

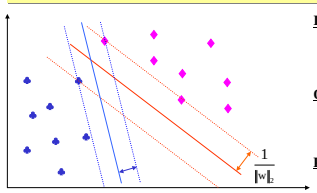
Soil microbial diversity is often regarded as an important index of soil ecosystem health. Direct amplification of bacterial 16S rRNA genes from extracted soil DNA provides the most comprehensive and flexible means of sampling bacterial communities. Amplicon length heterogeneity (ALH) profiles of the hypervariable regions from 16S rRNA have been used to rapidly analyze the relative diversity of complex microbial communities. The characterization capability of a combination of the ALH profiles obtained from four different hypervariable regions within the 16S rRNA gene is investigated. A machine-learning tool based on the technique of support vector machines (SVMs) has been developed. The software tool is provided a training set (collection of "labeled" samples) from which it "learns". Following the training session, the software has the capability of automatically "labeling" a test set (set of new samples to be classified). The software was trained and tested with a collection of ALH profiles of microbial communities obtained from samples of Idaho agricultural soils and sediments from the Chesapeake Bay marsh regions. The Idaho soils represented four different soil management practices (native sagebrush vegetation, irrigated moldboard plowed crops, irrigated conservation tilled crops, and irrigated pasture systems), collected at different depths; the barrier island fringe marsh in the Chesapeake Bay region was the source of samples from three distinct habitats (high dry Spartina marsh, low wet Spartina marsh, and adjacent mud flats), collected at different times of the year. Our experiments show that the total prediction accuracies for the software ranged from 90-100% for the Idaho soils, and between 80-85% for the Chesapeake Bay samples. Results from K-nearest-neighbor method, another supervised approach, are comparable to those using SVMs. Our methods are robust, i.e., relatively insensitive to errors. Computational tools proposed here can help in building a database of profiles of known samples, and in evaluating techniques that characterize microbial diversity.

B Introduction

B.1 *Amplicon Length Heterogeneity* (ALH) assays are based on the natural variations in sequence lengths of specific hypervariable regions of the ribosomal DNA (1, 2). Given a sample consisting of many organisms, performing PCR using a pair of primers will yield a profile of the domain lengths associated with the microorganisms in the sample. The heights of the peaks indicate the relative abundance of the organisms with the associated domain lengths. Different pairs of primers were used to target different variable regions of the 16S rRNA gene. A *combined profile* is simply a combination of the (normalized) ALH profiles obtained from using different pairs of primers on the same sample. To use machine learning methods, feature vectors were extracted from ALH combined profiles based on the length, i.e., each different length corresponding to a feature vector and the relative abundance corresponding to a value at the specific vector. Since there are differences in the length patterns for different soil samples, the total number of feature vectors were the combination of all lengths exhibited in different samples. The data was used to train a support vector machine.

B.2 *Support vector machines* (SVMs), which are supervised machine learning tools (3), have been shown to perform well in a variety of research areas including evaluating microarray expression data, detecting remote protein homologies and predicting subcellular localization prediction. SVMs have demonstrated the ability to not only correctly separate entities into appropriate classes, but also to identify instances whose classification is not supported by the data. SVMs are also well-suited for dealing with high-dimensional data.

B.3 *K-nearest-neighbor* (KNN) classifiers are memory-based, and require no model to be fit. Given a query point x_0, the k training points closest in distance to x_0 are used to classify using a majority vote among the k neighbors.



Input:

Training vectors: x_i, i=1,…,n

Class labels: $y_i = \pm 1$, i=1,…,n

Output:

$\text{minimize}_{w,b} \langle w \cdot w \rangle$

subject to $y_i(\langle w \cdot x_i \rangle + b) \geq 1$, i=1,…,n

Prediction:

$y = \text{sign}(\langle w \cdot x \rangle + b)$

Figure.1 Linear support vector machines (SVMs)

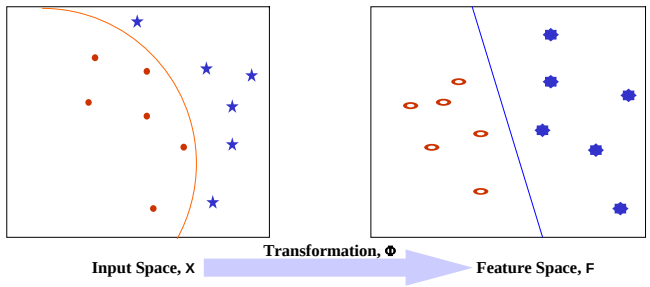


Figure 2. Nonlinear Case: Three different Kernel functions

- **Linear**: K(X,Y) = (X•Y + 1)d, d=1
- **RBF**: K(X,Y) = exp(-||X-Y||2/2α2)
- **Sigmoid**: K(X,Y) = tanh(ϖ(X•Y)+θ)

C-1 Experiments

The two classification tools(3) were applied to two different datasets:

- Idaho soil data sets: Data from four different soil management types were used
- Chesapeake Bay soil data sets: Data from nine different sites were used

C-2 Measures of Accuracy

$$\text{accuracy} = \frac{p(i)}{obs(i)}, \text{ total accuracy} = \frac{\sum_{i=1}^{k} p(i)}{N}$$

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

Jackknife: Singling out one for test and training with the remaining

N, the total number of sequences; *k*, the class number; obs*(i)*, the number of sequences observed in location *I; p(i)*, the number of correctly predicted samples of class *i,; n(i)*, the number of correctly predicted samples not of class location *I; u(i)*, the number of under-predicted samples; and *o(i)*, the number of over-predicted samples.

C-3 Prediction accuracy – Idaho Dataset

Location	Number of samples	KNN Accuracy (%)	Linear		RBF		Sigmoid	
			Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
NSB	6	100.00	100.00	1.00	100.00	1.00	100.00	1.00
CT	7	100.00	100.00	0.91	100.00	0.91	100.00	0.91
IP	8	100.00	100.00	1.00	100.00	1.00	100.00	1.00
MP	9	88.89	88.89	0.92	88.89	0.92	88.89	0.92
Total accuracy	30	96.67	96.67		96.67		96.67	

Table 1. Prediction accuracies for Idaho top soils with different type of kernel functions
NSB = natural sage brush, CT = conservation tillage, IP = irrigated pasture and MP = moldboard plowed.

C-4 Prediction accuracy-Chesapeake Bay Dataset

Location	Number of samples	KNN Accuracy(%)	RBF*	
			Accuracy (%)	MCC
CP	23	86.95	78.26	0.80
CS	60	90.00	88.33	0.84
HD	52	84.62	90.38	0.88
HI	38	76.31	86.84	0.81
HW	42	83.33	80.95	0.83
OC	23	82.60	82.61	0.74
OM	9	55.56	55.56	0.48
RB	30	70.00	73.33	0.75
UP	5	60.00	60.00	0.67
Total accuracy	282	81.56	82.97	

Table 3. Prediction accuracies for location classification of Chesapeake Bay samples with optimized RBF function.
* C = 512, ϒ = 1. The results were given by the jackknife test.
CP = Chimney Pole, CS = Cattle Shed, HD = Hog Is. Dry, HI = Hog Island, HW = Hog Is. Wet, OC = Oyster Creek, OM = Oyster Creek Marsh, RB = Red Bank, UP = Upper Phillips Creek

D Discussions and Conclusions

- SVM tool has been effectively used to accurately classify ecogenomics data.
- SVM training includes the selection of the proper kernel, function parameters and the regulation parameter C.
- SVM has a few tunable parameters which needed to be optimized. The results by 5-fold cross validation were used to select the appropriate parameters.
- K-Nearest Neighbor method, another classification tool, was also used to classify ecogenomics data.
- The performance of the K-Nearest Neighbor method is comparable to that of SVMs.
- It is anticipated that the above prediction methods would be useful tools for the large-scale analysis of ecogenomics data.
- As an application, ALH profiles of known samples of soils from different regions were used to train a program & then accurately predict.

Acknowledgements

The presenter's graduate study is supported by the School of Computer Science with Graduate Assistantship. Travel is supported by funds from Graduate Student Association and College of Arts & Sciences of FIU.

References

1. Suzuki, M., M. S. Rappe, S. J. Giovannoni. (1998). Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA Gene PCR amplicon length heterogeneity. Applied Environ. Microbiology 64:4522-4529.
2. Mills, D. K., P. M. Gillevet, K. Fitzgerald, C. D. Litchfield. (2003). A comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum contaminated soils, J Microbiol Meth 54: 57-74.
3. Joachims, T. (1999). Making large-scale SVM learning practical. http://svmlight.joachims.org
4. Chang, C.-C., and C.-J. Lin. (2002). LIBSVM: a Library for Support Vector Machines. http://www.csie.ntu.edu/tw/~cjlin/libsvm