APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 2004, p. 6525–6534 Vol. 70, No. 11
0099-2240/04/$08.00+0 DOI: 10.1128/AEM.70.11.6525–6534.2004

Coupling of Functional Gene Diversity and Geochemical Data from Environmental Samples

A. V. Palumbo,* J. C. Schryver, M. W. Fields,† C. E Bagwell,‡ J.-Z. Zhou, T. Yan, X. Liu,§ and C. C. Brandt

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

Received 29 August 2003/Accepted 28 June 2004

Genomic techniques commonly used for assessing distributions of microorganisms in the environment often produce small sample sizes. We investigated artificial neural networks for analyzing the distributions of nitrite reductase genes (*nirS* and *nirK*) and two sets of dissimilatory sulfite reductase genes ($dsrAB_1$ and $dsrAB_2$) in small sample sets. Data reduction (to reduce the number of input parameters), cross-validation (to measure the generalization error), weight decay (to adjust model parameters to reduce generalization error), and importance analysis (to determine which variables had the most influence) were useful in developing and interpreting neural network models that could be used to infer relationships between geochemistry and gene distributions. A robust relationship was observed between geochemistry and the frequencies of genes that were not closely related to known dissimilatory sulfite reductase genes ($dsrAB_2$). Uranium and sulfate appeared to be the most related to distribution of two groups of these unusual *dsrAB*-related genes. For the other three groups, the distributions appeared to be related to pH, nickel, nonpurgeable organic carbon, and total organic carbon. The models relating the geochemical parameters to the distributions of the *nirS*, *nirK*, and $dsrAB_1$ genes did not generalize as well as the models for $dsrAB_2$. The data also illustrate the danger (generating a model that has a high generalization error) of not using a validation approach in evaluating the meaningfulness of the fit of linear or nonlinear models to such small sample sizes.

One of the goals of microbial ecology is to understand which abiotic factors control the abundance and distribution of microorganisms in the environment. Environmental microbial ecology is beginning to achieve this goal in a wide range of habitats (6, 8, 30, 59) with the advent of molecular techniques that allow a significant part of the indigenous populations to be identified to some phylogenetic or functional level. For example, microbial distributions and diversity have been examined in relation to spatial factors (1), freshwater and ocean environments (51), and soil type (48, 50). Distribution or diversity has also been linked to dominant environmental characteristics or seasonal variations (29, 43, 57, 63, 68). To identify the critical factors that influence population distribution in complex environments, sophisticated data analysis techniques are needed to model the relationships between microbial distributions and environmental characteristics (14, 66).

Cloning and sequencing of functional genes from environmental samples are powerful methods for investigating the ecology of microorganisms. These techniques have advanced our understanding of the types of microorganisms and degradation capabilities found in various habitats (6, 12, 15, 43, 51). However, relating the population data generated by these techniques to environmental characteristics, such as geochemical measurements, can be challenging. One problem is the small sample size that is typical in these studies (66). The time and difficulty of generating and characterizing clone libraries that adequately cover the microbial populations often limit the sample size. Another characteristic is the large number of measurements for each sample. Finally, the underlying relationships between the microbial populations and their environment are often complex and nonlinear.

Various statistical and mathematical tools are available for relating the distributions of microorganisms to environmental characteristics. One powerful nonlinear approach that has been used to analyze such data is artificial neural networks (ANNs) (10, 17, 39, 45, 46, 50). ANNs are interconnected layers of simple computational units that map the relationships between predictor and target vectors. A computational unit sums its inputs and computes its present state from a nonlinear activation function based on this sum. Outputs from each layer are passed onto the next layer via weights that can be optimized to reflect the strength of the connection. Adjacent layers are typically fully interconnected. Bishop (7), Haykin (25), and Jain et al. (27) provide basic introductions to ANN theory.

ANN models are more general than linear methods. Hornik et al. (26) have shown that with a sufficiently complex architecture, an ANN is capable of approximating any continuous function. These approximations can be very precise if the training set is sufficiently large (64). ANNs are also more general than other classes of nonlinear statistical methods, such as general additive models, because the form of the nonlinear function does not have to be specified. They can also generalize to new data sets and degrade gracefully in the presence of noisy data. ANNs have demonstrated some of these potential advantages when applied to microbial data in studies compar-

* Corresponding author. Mailing address: Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831. Phone: (865) 576-8002. Fax: (865) 576-0524. E-mail: palumboav@ornl.gov.

† Present address: Department of Microbiology, Miami University, Oxford, OH 45056-1400.

‡ Present address: Savannah River National Laboratory, Aiken, SC 29808.

§ Present address: Central South University, Changsha, Hunan, People's Republic of China.

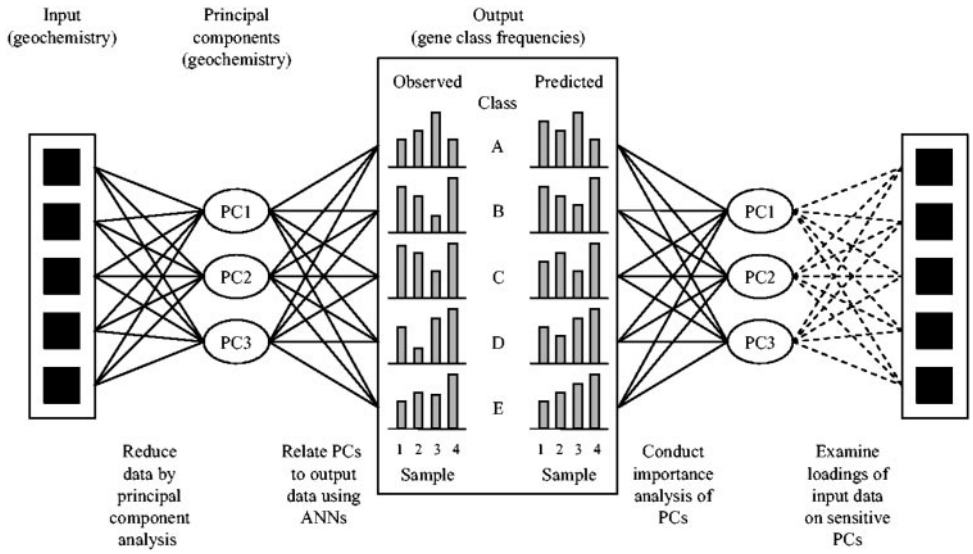

FIG. 1. Data analysis strategy.

ing linear methods with ANNs (10, 45). ANNs are becoming a primary modeling tool in biotechnology (2).

This paper describes the use of ANN models to analyze clone libraries for two sets of functional genes (dissimilatory sulfite reductase and nitrite reductase genes) isolated at the U.S. Department of Energy's Natural and Accelerated Bioremediation Research (NABIR) Program field site in Oak Ridge, Tenn. (55). Our general approach (Fig. 1) was to divide the clone libraries into phylogenetic groups based on sequence similarity and investigate whether the groups could be linked to the geochemistry of the samples. This process began with a reduction of the geochemical data by principal component analysis. Next, linear and nonlinear ANN models were developed to relate the reduced geochemical data to the distributions of the clone groups. We examined weight decay and leave-one-out cross-validation as methods for managing generalization error in the models. The influences of the geochemical principal components on the final model predictions were assessed using a normalized importance index that was computed by measuring the proportional increase in data misfit following effective removal of each principal component from the model. The results were used to identify the geochemical measurements that had the greatest influence on the distributions of the clone groups. The critical need for assessing generalization error and the utility of the weight decay method in constructing predictive models are discussed.

MATERIALS AND METHODS

Sampling sites. One background and five contaminated wells were sampled at the NABIR Field Research Center in Oak Ridge, Tenn. The contaminated samples (FW-010, FW-005, FW-015, FW-003, and TPB-16) were taken at varying distances from the former S-3 waste disposal ponds (Fig. 2), resulting in different levels of contamination in the samples. The background sample, FW-300, was from an uncontaminated area that has soil characteristics similar to those originally found near the S-3 waste ponds. Geochemical parameters that were measured include pH, dissolved oxygen (DO), total organic carbon (TOC), nonpurgeable organic carbon (NpOC), nitrate, nickel, technetium-99 (Tc-99), sulfate, and uranium. The geochemistry of the sampling sites has been well described (66). In general, the contaminated sites have low pH, high nitrate, high Tc-99, high sulfate, low DO, and variable NpOC.

Molecular methods. The extraction and purification of DNA, amplification of *nirS*, *nirK*, and *dsrAB* genes, cloning, restriction fragment length polymorphism (RFLP) analysis, and sequencing were done as described by Yan et al. (66). Briefly, bacteria were harvested by centrifugation (10,000 × *g*, 4°C for 30 min), and the biomass pellets were stored at −80°C until they were used for DNA extraction. DNA was extracted as previously described (67), and the precipitated DNA was further purified. The partial gene sequences were amplified in a 9700 Thermal Cycler (Perkin-Elmer, Wellesley, Mass.) with previously described primer pairs (9, 28, 66) and reaction conditions (66). The PCR parameters were selected to minimize artifacts as described by Qiu et al. (49). Cloning was done using a pCR2.1 vector from a TA cloning kit, and competent *Escherichia coli* cells were transformed according to the manufacturer's (Invitrogen, Carlsbad, Calif.) instructions. RFLP analysis of clone libraries was done as described by Yan et al. (66). Briefly, inserts from picked colonies were amplified with the TA primers specific for the pCR2.1 vector (TAF: 5′GCC GCC AGT GTG CTG GAA TT 3′ and TAR: 5′TAG ATG CAT GCT CGA GCG GC 3′). The inserts were visualized on a 1.5% agarose gel, and PCR products of the correct size were digested with 0.1 U of MspI and RsaI (Gibco-BRL, Carlsbad, Calif.) overnight at 37°C. Digested fragments were separated by electrophoresis (7 V/cm, 4 h) in 3.5% Metaphor agarose gels with 10 μl of 10-mg/liter ethidium bromide in 1× Tris-borate-EDTA buffer. The RFLP patterns were visualized with UV radiation, saved as TIFF images, and compared with Molecular Analyst version 1.6 software (Bio-Rad Laboratories, Hercules, Calif.).

Unique *nirS*, *nirK*, and *dsrAB* clones from each site were selected for further sequence analysis based on differences in their RFLP patterns. PCR products were amplified (66), and DNA sequences were determined with a BigDye Terminator kit (Applied Biosystems, Foster City, Calif.) and a 3700 DNA analyzer

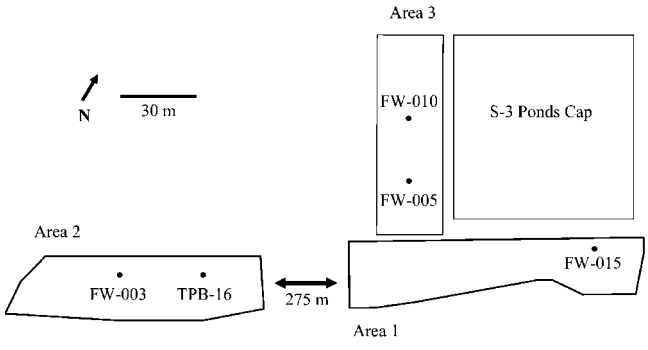

FIG. 2. Sample locations at the contaminated field site. The uncontaminated sample was taken about 5 km to the west of this location.

(Perkin-Elmer) according to the manufacturer's instructions. The sequences obtained were compared with *nirS*, *nirK*, and *dsr* sequences from GenBank, translated into amino acid sequences, and aligned using techniques described by Yan et al. (66). Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 2.1 (36), and phylogenetic trees were constructed from distance matrices using the neighbor-joining method. Trees constructed with maximum-likelihood methods were not significantly different.

Five gene groups within the *nirS* and *nirK* groups were defined based on sequence similarity as described by Yan et al. (66). However, many of the *dsrAB* clones appeared to be consistent with those found in a variety of environments (16, 21, 61) but very different from those found in confirmed sulfate-reducing groups. Since the function of these outlying sequences is in some doubt, we divided the *dsrAB* clones into two subgroups for data analysis. The subgroup designated $dsrAB_1$ included those clones that appear to be most similar to known sulfate reducers, such as *Desulfosporosinus*, *Desulfococcus*, and *Desulfosarcina* (16). The subgroup designated $dsrAB_2$ contained much more diversity and represented sequences that were relatively dissimilar to confirmed sulfate reducers. The $dsrAB_1$ subgroup contained four classes, and the $dsrAB_2$ subgroup consisted of five classes.

Data preparation. Relative to the small number of samples, the nine geochemical analytes constituted a large set of potential predictors. Principal components (PCs) analysis (PCA) was used to reduce the original geochemical variables to a smaller set of predictors. Several investigators have recently employed PCA (5, 62) to orthogonally transform input variables for predictive ANN analysis. The main advantage of this approach is that the complexity of the model is substantially reduced without sacrificing much important information. On the other hand, the PCs are surrogates for the original variables, and the relationships among the original analytes and output variables (class frequencies) could be obscured.

Since the distributions of the nitrate, sulfate, uranium, Tc-99, and nickel data were skewed, they were transformed using the function $\log(x + 1)$ to approximately normalize the data prior to PCA. The first three PCs, which cumulatively accounted for 92% of the total variance, were selected as inputs for subsequent data analyses. The preprocessing step resulted in a reduction of the number of inputs from nine to three.

Both the geochemical PCs (inputs) and gene group frequencies (outputs) were normalized to values between 0 and 1 to maximize the performance of the models (7). A sigmoid transformation $\{y = 1/[1 + \exp(ax + b)]\}$ was applied to the geochemical PC scores because they were centered about zero but unbounded. Nonzero gene frequencies were normalized using the function $y = x^b/(x^b + a)$, which has a lower bound of zero and an upper bound of 1.

Model selection. Models were implemented using a combination of custom-designed MATLAB functions (The MathWorks, Natick, Mass.) and the Netlab (42) toolkit for MATLAB. A standard multilayer perceptron architecture with three fully interconnected layers (input, hidden, and output) was employed. The hyperbolic tangent transform was the nonlinear activation function in the hidden layer, and the logistic function was selected as the nonlinear activation function in the output layer. All of the ANNs were trained with the scaled conjugate gradient algorithm. For comparison, the same MATLAB functions were used to train a simpler class of generalized linear models (GLMs) on all of the data sets. Each GLM was computed by employing only a single node in the hidden layer with a linear activation function. This network computed a logistic function, which is linear in the input variables with nonlinear output. GLMs offer a natural, simpler alternative to ANNs for predictive modeling.

Overfitting is a concern when using ANNs to analyze small data sets such as the one described here. We used weight decay, one of the most popular and theoretically motivated methods, to improve the generalization performance of ANNs by introducing a controlled amount of regularization to the objective function (40). The form of the modified objective function was $M(w) = E_D(w) + \alpha E_W(w)$, where the vector w includes both weights and unit biases. The term $E_D(w)$ is the measure of model-data misfit. The additional term $E_W(w)$ limits the model complexity by imposing a penalty on large weights, where large weights are associated with more complex functions. The measure of model complexity can take many forms (7), but the simplest and most frequently used formulation is $E_W(w) = 0.5 \Sigma w_i^2$. The hyperparameter α is a positive constant that determines the penalty for model complexity. Unit biases were needed to adapt to the output range, so only the weights received a positive value for α.

We examined the effect of weight decay on generalization performance by plotting values of α versus the corresponding validation error E_D. If an ANN is overfitting, the validation E_D typically begins at a relatively high value, gradually decreases to a minimum, and then starts to increase again as α increases. For a large α value, E_D converges to a value that corresponds to the performance obtained by the simplest possible model, i.e., prediction of the mean output. Thus, the plot defines a smooth path from a complex ANN to the simplest model. The minimum of the weight decay function identifies the value of α that produces the lowest generalization error. For this study, the values of α selected for *nirS*, *nirK*, $dsrAB_1$, and $dsrAB_2$ were 1.0, 0.35, 1.0, and 0.001, respectively.

K-fold cross-validation is a well-established method of using an entire data set for both training and testing (7, 54, 58). We performed onefold, also known as leave-one-out, cross-validation in which one sample was withheld from training and used to test the model fitted to the remaining data. This procedure was repeated six times, each time withholding a different sample for testing. The final ANN models selected were those that possessed the smallest generalization error over a wide range of values for α. The final architectures (input $\times$ hidden $\times$ output nodes) selected for each data set were $3 \times 4 \times 4$ for $dsrAB_1$, $3 \times 4 \times 5$ for $dsrAB_2$, $3 \times 3 \times 5$ for *nirS*, and $3 \times 3 \times 5$ for *nirK*.

Importance analysis. It is critical to assess the relative importance or salience of each input variable with respect to the model predictions. We used an importance index that is based on a sensitivity concept proposed by Moody (41). His approach evaluated the change in training error that occurs when an input is effectively removed from the network. The most unbiased method of removing the influence of an input is to substitute its mean value (computed over the entire data set) in each sample. The sensitivity index is the average change in the mean squared error (MSE) that occurs after removing the input's influence and without retraining the network. Instead of measuring the difference, we computed the ratio of corrected MSE to MSE and normalized the ratio such that the importance indices sum to unity over all input variables. The resulting index describes the importance of the specified input variable relative to that of the other input variables.

RESULTS

Data reduction. The principal component analysis reduced the initial geochemical data set from nine measured parameters to three PCs that cumulatively accounted for 92% of the variability in the original data. The first component (PC1) accounted for most of the variability (60%), with the second (PC2) and third (PC3) components explaining less variance (22 and 9%, respectively). Variables that loaded heavily on the first component were nitrate, pH, Tc-99, nickel, and NpOC (Fig. 3). None of these variables loaded heavily on the other components. TOC loaded to a lesser degree (0.728) on PC1, and all others loaded at values between -0.55 and 0.55 (Fig. 3). Variables that loaded particularly heavily on the second component included uranium (-0.940) and sulfate (-0.765). Only dissolved oxygen (-0.819) loaded heavily on the third component (Fig. 3).

Model results and validation. The ANN models were always equal to or better than the GLMs in predicting the *nirS*, *nirK*, $dsrAB_1$, and $dsrAB_2$ frequencies in the entire data set (Table 1). Both the ANN and GLM (Table 1) were able to predict the $dsrAB_2$ frequencies with a higher explained variance (EV) than the other classes. The ANN method was also able to fit *nirS* and $dsrAB_1$ frequencies with EV values greater than 95% and *nirK* frequencies with EV values greater than 85%. The additional parameters and flexibility of the ANN appeared to permit a better fit of the geochemical PCs to the gene frequencies. However, examination of the validation data indicated potential overfitting problems.

Addition of weight decay to the ANN and GLM was designed to reduce generalization error at the expense of increased training error, and it was evident that the magnitude of the weight decay term did have a substantial influence on the training and generalization error (Fig. 4). Two patterns were observed when examining changes in E_D for training and generalization with increasing levels of weight decay. When weight decay was added to the models for $dsrAB_2$, the minimum train-

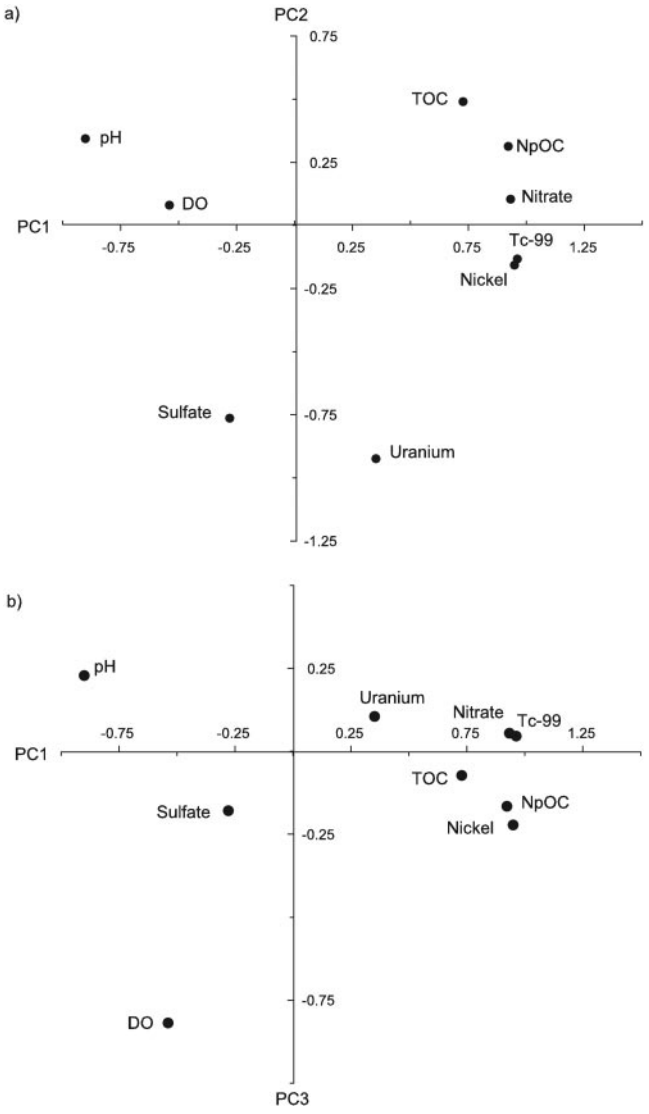

FIG. 3. Loadings of the geochemical variables on (a) PC1 and PC2 and (b) PC1 and PC3. The loading measures the correlation between the PC and the variable.

ing error was observed at a low value of α (10^{-3}). For *nirK*, *nirS*, and $dsrAB_1$, the validation E_D decreased monotonically as α increased with an accompanying increase in the training error.

Cross-validation, based on the leave-one-out method, indicated that with weight decay the $dsrAB_2$ ANN model was the most general since it had the lowest validation mean squared error (Table 2). The validation error was also small for the $dsrAB_2$ GLM (Table 2). For all other gene groups, the validation errors of both the ANN model and the GLM were much larger (Table 2). Although the EV was quite high for the *nirS* and $dsrAB_1$ ANN models (Table 1), the validation errors were also quite high (Table 2) due to a poor fit of predicted to observed gene frequencies in the validation data (e.g., for $dsrAB_1$ [Fig. 5b]). Although the ANNs were able to better fit the *nirS*, *nirK*, and $dsrAB_1$ data than the GLMs, the validity of both model types was questionable for these genes given the high validation error. For $dsrAB_2$ the relationships between geochemical PCs and the gene class frequencies appeared to be more robust since the validation error was low.

Importance analysis. The validation results did not lend confidence to the models produced for $dsrAB_1$, *nirS*, and *nirK*. Due to this poor performance, an importance or sensitivity analysis was not done for these gene groups. The importance analysis for $dsrAB_2$ indicated that the first two PCs were much more important than the third (Fig. 6). Overall, PC2 was the most important (mean importance index over all gene classes = 47.4%), while PC1 was close to it in importance (42.9%) and PC3 was much less important (9.68%). These results suggested that the different classes of $dsrAB_2$ were responding to different geochemical site characteristics. The A, D, and E classes responded most to PC1, and for the A and E classes, PC2 also had moderate importance. The B and C classes responded more to PC2. PC3 was the least important in predicting the five classes (Fig. 6).

The loadings on the PCs indicated the important geochemical parameters for the different gene classes. For example, nitrate, pH, Tc-99, nickel, and NpOC loaded heavily on PC1 (Fig. 3) and hence were linked to changes in $dsrAB_2$ gene classes A, D, and E. Uranium and sulfate loaded heavily on PC2 and were most strongly linked to changes in gene classes C and D. However, there was a lesser but fairly high sensitivity of gene classes A and E to the second component. Because none of the distributions of the gene classes were highly sensitive to the third component (PC3), it appears that the one geochemical variable that loaded heavily on this component (dissolved oxygen) did not have a large influence on any of the gene class distributions in these samples.

The highest importance value observed was PC1 for class D in $dsrAB_2$ (Fig. 6), and it appeared that the influence of PC1 on this class was monotonic (Fig. 7). The highest frequency for class D was found at the stations with the lowest values for PC1 (e.g., FW-300), and the frequency in class D generally decreased with increasing values of PC1 (Fig. 7). The highest level of class D was associated with low nitrate, nickel, Tc-99, and NpOC (positive loadings on PC1) and high pH (negative loading on PC1). Class D frequencies also decreased with decreasing values of PC2 (Fig. 7), thus indicating that the class was associated with low sulfate and uranium. The importance of PC3 for this group was extremely small (Fig. 6).

Near-monotonic relationships with PC1 and PC2 were also seen for classes A and E (Fig. 7). For class A, the importance values indicated that both PC1 and PC2 were meaningful (Fig. 6). The largest frequency of class A was seen at the maximum values of PC1 (i.e., low pH, high nitrate, high nickel, high TOC) and PC2 (i.e., low sulfate and uranium). The frequency generally dropped with decreasing values for PC1 and PC2.

TABLE 1. Percent variance explained in the entire data set by GLM and ANN models with and without weight decay

Method	Weight decay	Variance (%) for indicated gene group			
		nirS	*nirK*	$dsrAB_1$	$dsrAB_2$
GLM	No	41.45	57.46	66.55	95.02
ANN	No	97.41	86.93	99.99	99.82
GLM	Yes	16.67	21.87	16.67	94.97
ANN	Yes	16.67	25.62	16.67	99.66

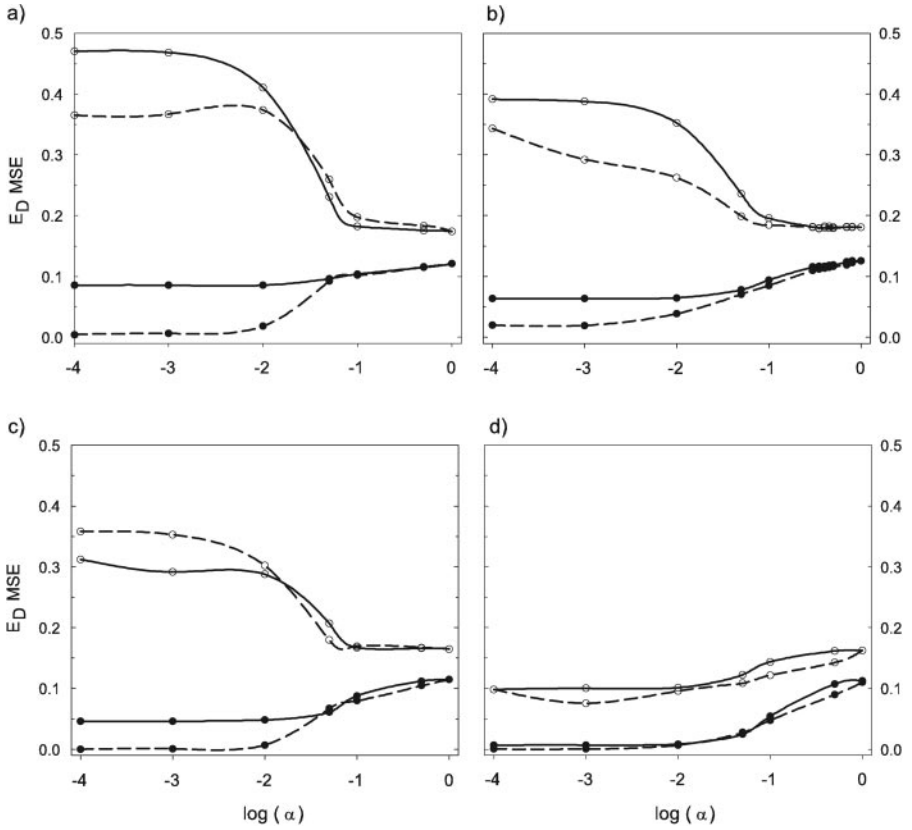

FIG. 4. Mean squared training error (E_D MSE) for GLM and ANN models using all and validation data for increasing (left to right) levels of weight decay (α) for the (a) *nirS*, (b) *nirK*, (c) $dsrAB_1$, and (d) $dsrAB_2$ gene groups. A solid line denotes a GLM, a dashed line indicates an ANN model, the filled circles represent all data, and the open circles indicate validation data.

The distribution of class E was somewhat different, with the highest frequencies found at low values of PC1 and high values of PC2 (Fig. 7). The frequency of class E generally decreased as both PC1 and PC2 decreased.

Apparent nonlinear relationships were observed between the geochemistry and the frequencies of classes B and C (Fig. 7). The highest importance values for PC2 were found for these classes (Fig. 6). The frequencies of these two groups appeared to peak at intermediate levels of PC2 and decreased at lower and higher values. The two classes differed in that PC1 was more important than PC2 for class B, but for class C PC2 was more important. Class C was the only case for which PC2 was more important than PC1. The frequencies of the two classes differed primarily at FW-300, where class B was high and class C was relatively low (Fig. 7).

DISCUSSION

Model results. Of the four gene groups examined in this study, only the frequencies of the $dsrAB_2$ classes could be linked to geochemical parameters. The negative results for $dsrAB_1$, *nirS*, and *nirK* could have been due to noise in the small samples overwhelming the statistical signals. Another more general problem was that gene transfer could obscure the relationships between functional groupings of bacteria and environmental characteristics. For example, multiple transfers of *dsrAB* genes are believed to have occurred (33), and similar *nirK* or *nirS* genes have been observed in distinctly different microorganisms (66). In addition, the diversity of the genes within the functional group may play a role. On a purely technical basis, the failure to observe environment-gene class relationships may be partially due to weaknesses in the training method used to find a global (or at least very good) minimum. We used a local search technique because it was more tractable when combined with cross-validation, which is computationally intensive.

Only the first two geochemical PCs were important in pre-

TABLE 2. MSE of GLM and ANN models using entire and validation data sets with and without weight decay for *nirS*, *nirK*, $dsrAB_1$, and $dsrAB_2$

Method	Weight decay	Data set	MSE[a] for indicated gene group			
			nirS	*nirK*	$dsrAB_1$	$dsrAB_2$
GLM	No	Entire	0.0848	0.0641	0.0460	0.0067
		Validation	0.4855	0.4055	0.3496	0.0994
ANN	No	Entire	0.0038	0.0197	<0.0001	0.0002
		Validation	0.3646	0.3270	0.3613	0.0956
GLM	Yes	Entire	0.1206	0.1178	0.1145	0.0068
		Validation	0.1737	0.1790	0.1649	0.0982
ANN	Yes	Entire	0.1206	0.1787	0.1145	0.0004
		Validation	0.1737	0.1787	0.1649	0.0757

[a]MSE is the mean of the squared differences between the observed and predicted transformed gene group frequencies.

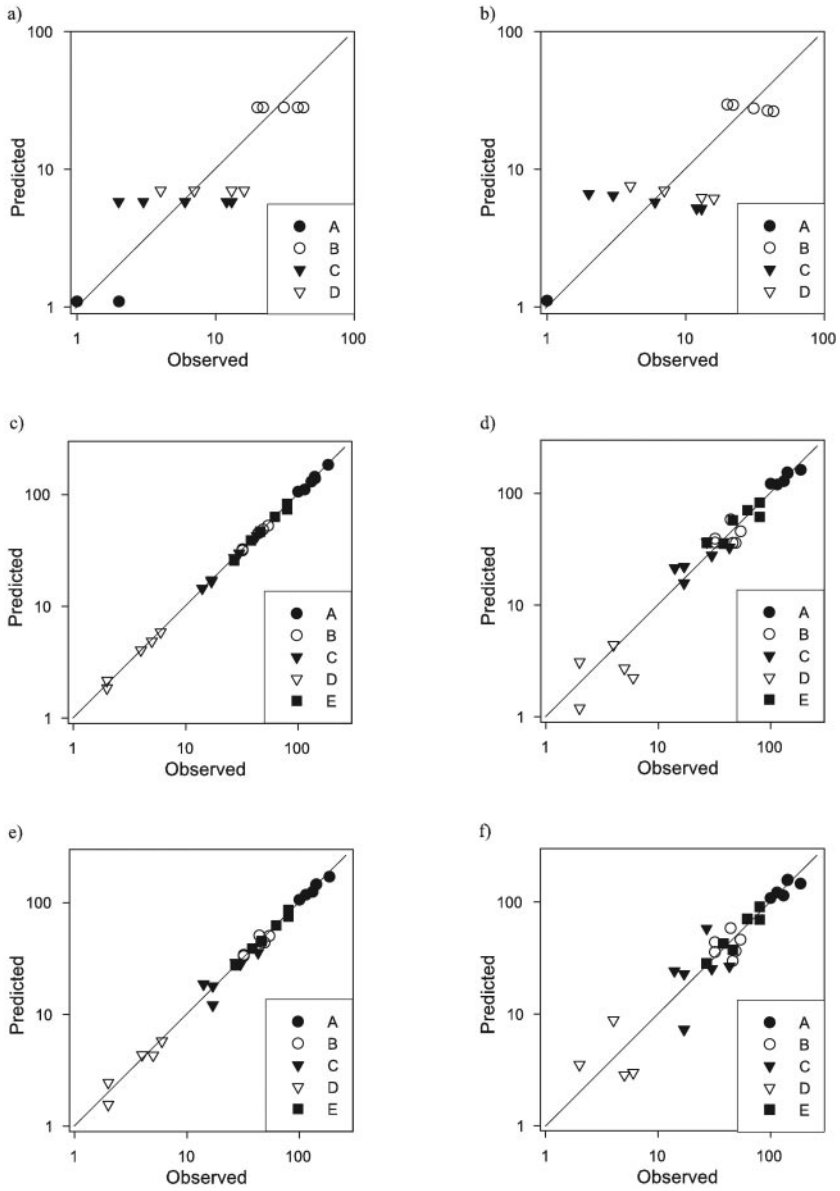

FIG. 5. Observed and predicted gene frequencies for (left column) training and (right column) validation of $dsrAB_1$ with (a and b) ANN, (c and d) $dsrAB_2$ with ANN, and (e and f) $dsrAB_2$ with GLM with weight decay.

dicting the frequency of different classes of $dsrAB_2$. The loadings on PC1 indicated that pH, Tc-99, nitrate, nickel, and NpOC were important in predicting the gene frequencies for classes A, B, D, and E. However, uranium and sulfate were more important in predicting the frequencies for class C. Interestingly, Chang et al. (14) found that uranium had an influence on one dominant cluster of *dsr* genes at a uranium mill tailings site. Examination of the loadings of the original geochemical parameters (Fig. 3) indicated that only DO loaded heavily on PC3. Thus, it appears that the DO concentration had little influence on the gene frequencies.

We found that cross-validation and weight decay were useful methods for measuring and increasing the generalizability of the ANN models. Due to the importance of the weight decay hyperparameter, we used a systematic method to assist in selection of α. The goal was to find a region of weight decay values that minimized generalization error. For $dsrAB_2$ this minimal region was observed when the weight decay hyperparameter was relatively small and had not resulted in a noticeable increase in training error (Fig. 4). For the other clone groups there was little improvement in the generalization error until the training error had increased substantially. The generalization error did not reach a minimum until the weight decay parameter was so high that the result was equivalent to guessing the mean for each sample and not using the geochemical data in the prediction at all. Thus, it appeared that the inferred relationships between the geochemical parameters and the $dsrAB_2$ frequencies were much more robust than those for the other gene groups.

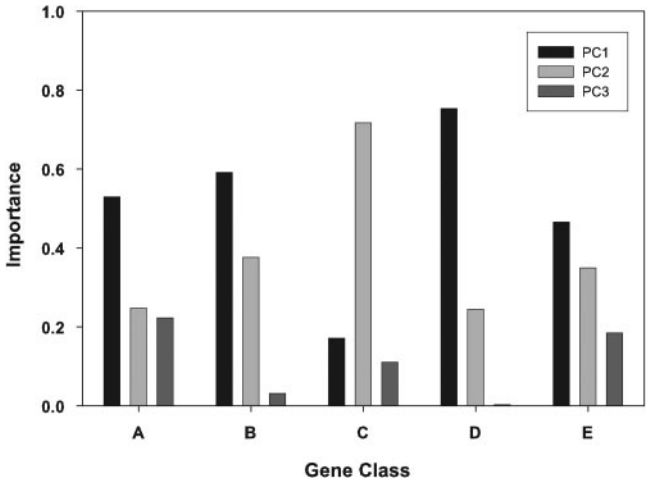

FIG. 6. Importance analysis of the three principal components for each $dsrAB_2$ gene class using the ANN model with optimized weight decay and leave-one-out cross-validation.

Modeling strategy. Modeling can help identify what environmental parameters control microbial community structure (17, 22, 34, 39, 44). However, if there are a large number of model parameters and a small sample size, a model can often be fit to the data even in the absence of truly meaningful relationships (19). This overfitting problem has been implicitly or explicitly recognized for environmental questions (18) and addressed by using techniques such as cross-validation to test for the generality of the model (39). Almeida (2) points out that some ANN packages can succeed in relating two sets of random numbers, thus indicating the importance of cross-validation. The modeling strategy should include techniques for assessing and addressing overfitting.

Overfitting is a potential problem in the application of ANNs, other parameterized predictive models, and classifiers such as GLMs to analysis of clone libraries and expression data (18, 31). A data set from a typical study might contain a few to a few dozen samples and a much greater number of measurements per sample. An ANN trained to predict distributions of clones from environmental measurements could possess numerous parameters (weights and biases) that must be estimated from the data. Since the number of parameters can greatly exceed the number of samples, the ANN is often able to provide a close fit to the data, even though the data contain measurement errors. The ANN attempts to fit all of the data's features, including the measurement error. The "true" model, on the other hand, is assumed to be a smooth function that describes a simpler relationship among variables but ignores the measurement error in the data. The result of overfitting is that the ANN generalizes poorly to new data that were not contained in the training set.

A useful approach for addressing overfitting is to reduce the number of inputs in the model by using a technique such as principal component analysis (57). In this study, we reduced nine geochemical characteristics to three principal components which still accounted for most of the variability in the original measurements (91%). An alternative approach is to use a technique that eliminates the variables that contribute the least either before (11) or after (50) specifying the final model. However, this approach is not as powerful as a data reduction technique, such as principal component analysis.

An important concern is the potential for generating models that may not generalize well (i.e., have a large generalization error). Weight decay is one method for addressing this problem. With weight decay, an additional term is added to the error function that is proportional to the sizes of the weights associated with each factor entering the models. Early stopping is a popular alternative to weight decay that is often employed when the number of parameters/number of samples ratio is significantly greater than unity (3, 20, 47, 54, 60). Early stopping is a nonconvergent technique that terminates training before the ANN is finished fitting the training data. Sarle (53) performed computer simulations which showed that early stopping can improve generalization. However, several investigators have noted problems with this technique (13, 56). In addition, examination of early stopping results with these small data sets (data not shown) showed that the optimal stopping point is highly sensitive to the variability in the validation set. Thus, we chose to focus this analysis on weight decay.

There is always the possibility that a model fits the training data well but is useless when applied to a new data set. Thus, a tool that allows for an assessment of the ability to generalize the model is necessary. A straightforward way to estimate generalization error is to test the model with a subset of data that was excluded from training (39, 50). If the percentage withheld is too high, not enough data remain for training. Too low a percentage can result in a validation set that does not resemble the entire data set if a few outliers are included by chance or if the subset contains only data in a narrow range. Some simulation studies suggest that the test set should be in the range of 5 to 25% of the total number of samples. Examples within most of this range (e.g., 10 to 25%) can be found in the application of ANNs to microbial data (35, 44, 46).

Cross-validation is a method that uses the entire data set for both training and testing (23). For example, suppose there are six (N) samples in a data set and we select one (k) for the test set. The procedure is repeated six (N/k) times such that all the data are eventually used in mutually exclusive test sets. Combining the error estimates from all N/k iterations provides a less biased estimate of the generalization error. Alternatives to cross-validation include early stopping and various complexity measures, such as the Akaike information criterion. In our experience, the usefulness of early stopping is severely limited by small sample sizes.

To make the final connection between the site geochemistry and the gene distributions, a method is needed to identify the model inputs that have the greatest influence on the predicted values. Traditionally, sensitivity analysis has been employed for measuring importance. However, there are several different definitions of sensitivity in general usage (32, 52). Sensitivity may be defined as a local measure that assumes a large value when a small perturbation about a specific input value produces a large change in the output. Other definitions are global measures that assess sensitivity as the mean of absolute sensitivities over all samples and outputs for each input variable. Different methods given identical input data may yield radically different results because they are based on different concepts. We propose an importance method that indicates which of the inputs had the greatest impact on predicting the distri-

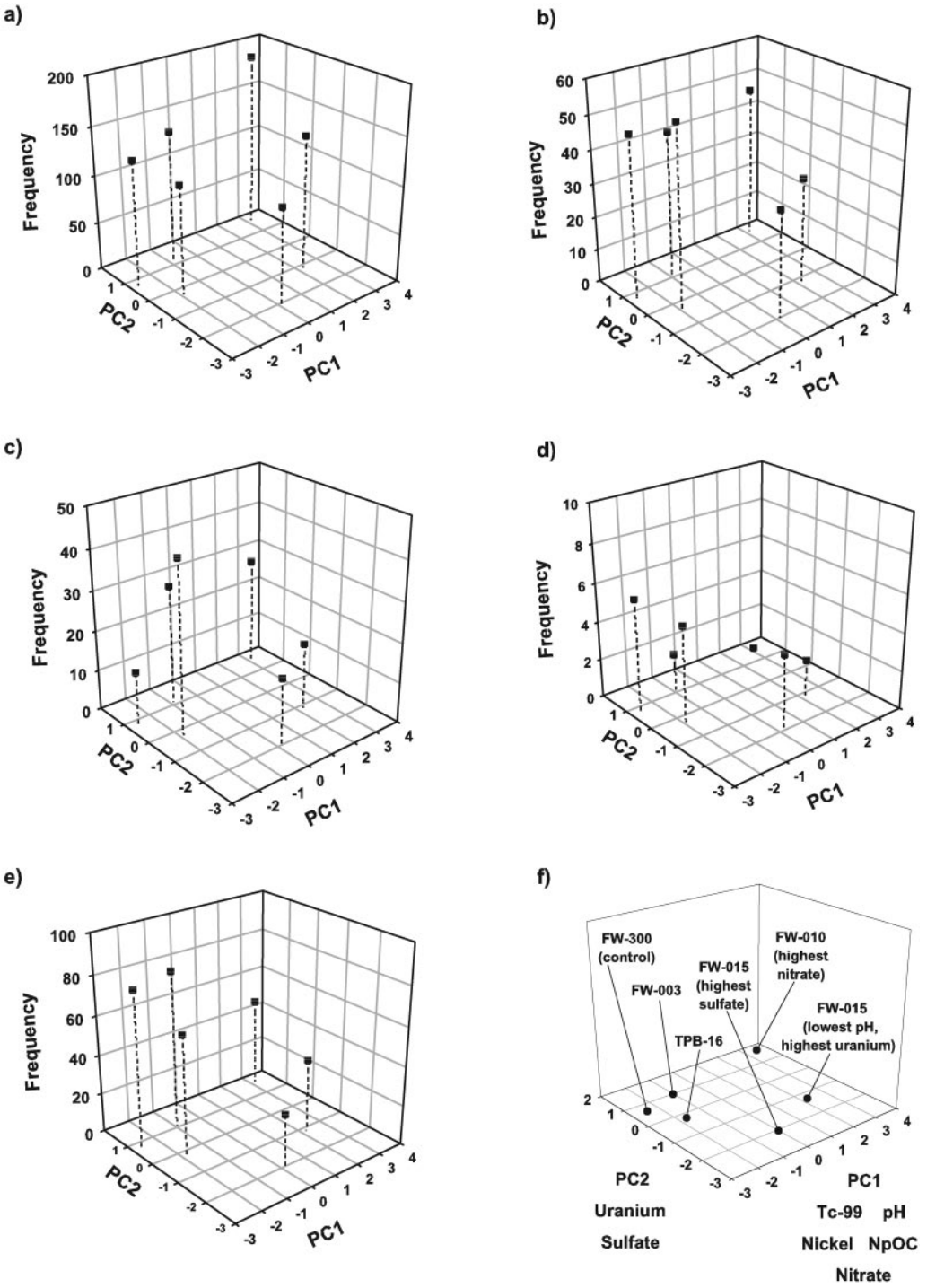

FIG. 7. Principal component analysis sample scores and *dsrAB*$_2$ frequencies for gene classes (a) A, (b) B, (c) C, (d) D, and (e) E. The samples are identified in panel f.

butions of gene clusters at the sites. This is equivalent to substituting the mean value of an input into the model for every sample and then measuring the increase in the prediction error based on only the other inputs.

The success of relating indicators of phylogenetic or functional groupings of bacteria (based on any technique) to environmental characteristics will depend on the validity of the assumption that microorganisms within a group can also be considered an ecotype that will respond in a similar manner to different geochemical conditions. This is the basis for using phylogenetic 16S rRNA gene probes to evaluate community changes within specific groups (37). Another approach is to group the clones of functional genes according to their distribution in the samples examined and relate those groupings to

the site characteristics. Thus, the groups being related to geochemistry or other characteristics can be distributionally related (they occur at a similar group of sites) but not related by sequence similarity. For example, Liu et al. (38) found relationships between nitrate and oxygen with community structure in denitrifying populations in marine sediments after grouping the clones of *nirS* and *nirK* by their distribution at the sampled sites. These types of clusters may not readily lend themselves to development and application of probes. However, approaches such as functional gene arrays (65) may be applicable in these situations as an alternative to cloning and sequencing.

Based on an examination of the literature and the results presented here, it may be uncommon for clone groups of specific functional genes to react in a similar manner to environmental characteristics, such as geochemistry. If this is true, probes for functional genes may not be as useful as those developed for 16S and other phylogenetic genes. General spatial distributions of phylogenetic groups have been observed (24) especially over wide ranges of environmental characteristics, such as the transition from fresh to salt water (51). Chang et al. (14) found by logistic regression that one cluster of *dsr* gene sequences (out of eight) was highly related to uranium at a mill tailings site. Also, there are examples where similar clones of functional genes have been found in a variety of dissimilar environments (4).

Conclusions. The inherent complexity and scale will often limit our ability to understand the relationships between phylogenetic or functional gene groups and environmental characteristics. To further our understanding we are attempting to capture patterns or correspondence between gene patterns and the geochemistry to infer relationships among them. If successful, we can begin to understand, at least in simple terms, what dominant variables show strong statistical influence or coincidence with specific populations. However, we will not always measure the critical environmental characteristics, or we may not measure them on an appropriate scale. Only when a characteristic we measure exerts a strong influence can we hope to make inferences between the geochemistry and gene distributions. Also, when we are successful it does not mean that other variables are not important but only that given the measurements made, the sites sampled, and the limits of our sampling methods we could not detect the influence of other variables.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Natural and Accelerated Bioremediation Research (NABIR) program, Biological and Environmental Research (BER), U.S. Department of Energy. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

1. **Affourtit, J., J. P. Zehr, and H. W. Paerl.** 2001. Distribution of nitrogen-fixing microorganisms along the Neuse River Estuary, North Carolina. Microb. Ecol. **41:**114–123.
2. **Almeida, J. S.** 2002. Predictive non-linear modeling of complex data by artificial neural networks. Curr. Opin. Biotechnol. **13:**72–76.
3. **Amari, S., N. Murata, K. R. Muller, M. Finke, and H. H. Yang.** 1997. Asymptotic statistical theory of overtraining and cross-validation. IEEE Trans. Neural Netw. **8:**985–996.
4. **Avrahami, S., R. Conrad, and G. Braker.** 2002. Effect of soil ammonium concentration on N_2O release and on the community structure of ammonia oxidizers and denitrifiers. Appl. Environ. Microbiol. **68:**5685–5692.
5. **Azimisadjadi, M. R., S. Ghaloum, and R. Zoughi.** 1993. Terrain classification in SAR images using principal components-analysis and neural networks. IEEE Trans. Geosci. Remote Sens. **31:**511–515.
6. **Bano, N., and J. T. Hollibaugh.** 2002. Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. Appl. Environ. Microbiol. **68:**505–518.
7. **Bishop, C. M.** 1995. Neural networks for pattern recognition. Clarendon Press, Oxford, United Kingdom.
8. **Blank, C. E., S. L. Cady, and N. R. Pace.** 2002. Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. Appl. Environ. Microbiol. **68:**5123–5135.
9. **Braker, G., J. Zhou, L. Wu, A. H. Devol, and J. M. Tiedje.** 2000. Nitrite reductase genes (*nirK* and *nirS*) as functional markers to investigate diversity of denitrifying bacteria in Pacific Northwest marine sediment communities. Appl. Environ. Microbiol. **66:**2096–2104.
10. **Brandt, C. C., J. C. Schryver, S. M. Pfiffner, A. V. Palumbo, and D. C. White.** 1999. Using artificial neural networks to assess changes in microbial communities, p. 1–6. *In* B. C. Alleman and A. Leeson (ed.), Bioremediation of metals and inorganic compounds. Proceedings of the Fifth International Symposium on In Situ and On-Site Bioremediation Symposium. Battelle Press, Columbus, Ohio.
11. **Brion, G. M., T. R. Neelakantan, and S. Lingireddy.** 2002. A neural-network-based classification scheme for sorting sources and ages of fecal contamination in water. Water Res. **36:**3765–3774.
12. **Brown, M. V., and J. P. Bowman.** 2001. A molecular phylogenetic survey of sea-ice microbial communities (SIMCO). FEMS Microbiol. Ecol. **35:**267–275.
13. **Cataltepe, Z., Y. S. Abu-Mostafa, and M. Magdon-Ismail.** 1999. No free lunch for early stopping. Neural Comput. **11:**995–1009.
14. **Chang, Y.-J., A. D. Peacock, P. E. Long, J. R. Stephen, J. P. McKinley, S. J. Macnaughton, A. Hussain, A. M. Saxton, and D. C. White.** 2001. Diversity and characterization of sulfate-reducing bacteria in groundwater at a uranium mill tailings site. Appl. Environ. Microbiol. **67:**3149–3160.
15. **Davis, J. W., J. M. Odom, K. A. Deweerd, D. A. Stahl, S. S. Fishbain, R. J. West, G. M. Klecka, and J. G. DeCarolis.** 2002. Natural attenuation of chlorinated solvents at Area 6, Dover Air Force Base: characterization of microbial community structure. J. Contam. Hydrol. **57:**41–59.
16. **Dhillon, A., A. Teske, J. Dillon, D. A. Stahl, and M. L. Sogin.** 2003. Molecular characterization of sulfate-reducing bacteria in the Guaymas Basin. Appl. Environ. Microbiol. **69:**2765–2772.
17. **Dollhopf, S. L., S. A. Hashsham, and J. M. Tiedje.** 2001. Interpreting 16S rDNA T-RFLP data: application of self-organizing maps and principal component analysis to describe community dynamics and convergence. Microb. Ecol. **42:**495–505.
18. **Drummond, S. T., K. A. Sudduth, A. Joshi, S. J. Birrell, and N. R. Kitchen.** 2003. Statistical and neural methods for site-specific yield prediction. Trans. ASAE **46:**5–14.
19. **Duda, R. O., P. E. Hart, and D. G. Stork.** 2001. Pattern classification. John Wiley and Sons, Inc., New York, N.Y.
20. **Finnoff, W., F. Hergert, and H. G. Zimmermann.** 1993. Improving model selection by nonconvergent methods. Neural Netw. **6:**771–783.
21. **Fishbain, S., J. G. Dillon, H. L. Gough, and D. A. Stahl.** 2003. Linkage of high rates of sulfate reduction in Yellowstone hot springs to unique sequence types in the dissimilatory sulfate respiration pathway. Appl. Environ. Microbiol. **69:**3663–3667.
22. **Fromin, N., J. Hamelin, S. Tarnawski, D. Roesti, K. Jourdain-Miserez, N. Forestier, S. Teyssier-Cuvelle, F. Gillet, M. Aragno, and P. Rossi.** 2002. Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns. Environ. Microbiol. **4:**634–643.
23. **Fujarewicz, K., M. Kimmel, J. Rzeszowska-Wolny, and A. Swierniak.** 2003. A note on classification of gene expression data using support vector machines. J. Biol. Syst. **11:**43–56.
24. **Gregory, L. G., P. L. Bond, D. J. Richardson, and S. Spiro.** 2003. Characterization of a nitrate-respiring bacterial community using the nitrate reductase gene (*narG*) as a functional marker. Microbiology **149:**229–237.
25. **Haykin, S. S.** 1999. Neural networks: a comprehensive foundation. Prentice-Hall, Englewood Cliffs, N.J.
26. **Hornik, K., M. Stinchcombe, and H. White.** 1989. Multilayer feedforward networks are universal approximators. Neural Netw. **2:**359–366.
27. **Jain, A. K., R. P. W. Duin, and J. C. Mao.** 2000. Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22:**4–37.
28. **Karkhoff-Schweizer, R. R., D. P. Huber, and G. Voordouw.** 1995. Conservation of the genes for dissimilatory sulfite reductase from *Desulfovibrio vulgaris* and *Archaeoglobus fulgidus* allows their detection by PCR. Appl. Environ. Microbiol. **61:**290–296.
29. **Kelly, K. M., and A. Y. Chistoserdov.** 2001. Phylogenetic analysis of the succession of bacterial communities in the Great South Bay (Long Island). FEMS Microbiol. Ecol. **35:**85–95.
30. **Khan, S. T., Y. Horiba, M. Yamamoto, and A. Hiraishi.** 2002. Members of the family *Comamonadaceae* as primary poly(3-hydroxybutyrate-co-3-hydroxyvalerate)-degrading denitrifiers in activated sludge as revealed by a polyphasic approach. Appl. Environ. Microbiol. **68:**3206–3214.
31. **Kim, S., E. R. Dougherty, J. Barrera, Y. D. Chen, M. L. Bittner, and J. M.**

Trent. 2002. Strong feature sets from small samples. J. Comput. Biol. **9:**127–146.
32. **Kleijnen, J. P. C.** 1997. Sensitivity analysis and related analyses: a review of some statistical techniques. J. Stat. Comput. Simul. **57:**111–142.
33. **Klein, M., M. Friedrich, A. J. Roger, P. Hugenholtz, S. Fishbain, H. Abicht, L. L. Blackall, D. A. Stahl, and M. Wagner.** 2001. Multiple lateral transfers of dissimilatory sulfite reductase genes between major lineages of sulfate-reducing prokaryotes. J. Bacteriol. **183:**6028–6035.
34. **Kowalchuk, G. A., and J. R. Stephen.** 2001. Ammonia-oxidizing bacteria: a model for molecular microbial ecology. Annu. Rev. Microbiol. **55:**485–529.
35. **Krause, D. O., W. J. Smith, L. L. Conlan, J. M. Gough, M. A. Williamson, and C. S. McSweeney.** 2003. Diet influences the ecology of lactic acid bacteria and *Escherichia coli* along the digestive tract of cattle: neural networks and 16S rDNA. Microbiology **149:**57–65.
36. **Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei.** 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:**1244–1245.
37. **Layton, A. C., P. N. Karanth, C. A. Lajoie, A. J. Meyers, I. R. Gregory, R. D. Stapleton, D. E. Taylor, and G. S. Sayler.** 2000. Quantification of *Hyphomicrobium* populations in activated sludge from an industrial wastewater treatment system as determined by 16S rRNA analysis. Appl. Environ. Microbiol. **66:**1167–1174.
38. **Liu, X., S. M. Tiquia, G. Holguin, L. Wu, S. C. Nold, A. H. Devol, K. Luo, A. V. Palumbo, J. M. Tiedje, and J. Zhou.** 2003. Molecular diversity of denitrifying genes in continental margin sediments within the oxygen-deficient zone of the Pacific coast of Mexico. Appl. Environ. Microbiol. **69:**3549–3560.
39. **Lovell, C. R., C. E. Bagwell, M. Czákó, L. Márton, Y. M. Piceno, and D. B. Ringelberg.** 2001. Stability of a rhizosphere microbial community exposed to natural and manipulated environmental variability. FEMS Microbiol. Ecol. **38:**69–76.
40. **Mackay, D. J. C.** 1992. Bayesian interpolation. Neural Comput. **4:**415–447.
41. **Moody, J.** 1994. Prediction risk and architecture selection for neural networks, p. 147–165. *In* J. H. Friedman, V. Cherkassky, and H. Wechsler (ed.), From statistics to neural networks: theory and pattern recognition applications. Springer-Verlag, Berlin, Germany.
42. **Nabney, I. T.** 2001. NETLAB: algorithms for pattern recognition. Springer, London, United Kingdom.
43. **Nakagawa, T., S. Hanada, A. Maruyama, K. Marumo, T. Urabe, and M. Fukui.** 2002. Distribution and diversity of thermophilic sulfate-reducing bacteria within a Cu-Pb-Zn mine (Toyoha, Japan). FEMS Microbiol. Ecol. **41:**199–209.
44. **Noble, P. A., J. S. Almeida, and C. R. Lovell.** 2000. Application of neural computing methods for interpreting phospholipid fatty acid profiles of natural microbial communities. Appl. Environ. Microbiol. **66:**694–699.
45. **Pfiffner, S. M., C. C. Brandt, J. C. Schryver, A. V. Palumbo, and J. S. Almeida.** 1999. Using artificial neural networks to assess microbial communities, p. 205–212. *In* G. A. Uzochukwu and G. B. Reddy (ed.), Proceedings of the 1998 National Conference on Environmental Remediation Science and Technology. Battelle Press, Columbus, Ohio.
46. **Piceno, Y. M., P. A. Noble, and C. R. Lovell.** 1999. Spatial and temporal assessment of diazotroph assemblage composition in vegetated salt marsh sediments using denaturing gradient gel electrophoresis analysis. Microb. Ecol. **38:**157–167.
47. **Prechelt, L.** 1998. Automatic early stopping using cross validation: quantifying the criteria. Neural Netw. **11:**761–767.
48. **Priemé, A., G. Braker, and J. M. Tiedje.** 2002. Diversity of nitrite reductase (*nirK* and *nirS*) gene fragments in forested upland and wetland soils. Appl. Environ. Microbiol. **68:**1893–1900.
49. **Qiu, X., L. Wu, H. Huang, P. E. McDonel, A. V. Palumbo, J. M. Tiedje, and J. Zhou.** 2001. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl. Environ. Microbiol. **67:**880–887.
50. **Ramadan, Z., X. H. Song, P. K. Hopke, M. J. Johnson, and K. M. Scow.** 2001. Variable selection in classification of environmental soil samples for partial least square and neural network models. Anal. Chim. Acta **446:**233–244.
51. **Rappé, M. S., K. Vergin, and S. J. Giovannoni.** 2000. Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems. FEMS Microbiol. Ecol. **33:**219–232.
52. **Saltelli, A., S. Tarantola, and F. Campolongo.** 2000. Sensitivity analysis as an ingredient of modeling. Stat. Sci. **15:**377–395.
53. **Sarle, W.** 1995. Stopped training and other remedies for overfitting, p. 352–360. *In* Computing science and statistics: proceedings of the 27th Symposium on the Interface. Carnegie Mellon University, Pittsburgh, Pa.
54. **Schenker, B., and M. Agarwal.** 1996. Cross-validated structure selection for neural networks. Comput. Chem. Eng. **20:**175–186.
55. **Senko, J. M., J. D. Istok, J. M. Suflita, and L. R. Krumholz.** 2002. In-situ evidence for uranium immobilization and remobilization. Environ. Sci. Technol. **36:**1491–1496.
56. **Sjöberg, J., and L. Ljung.** 1995. Overtraining, regularization and searching for a minimum, with application to neural networks. Int. J. Control **62:**1391–1407.
57. **Stepanauskas, R., M. A. Moran, B. A. Bergamaschi, and J. T. Hollibaugh.** 2003. Covariance of bacterioplankton composition and environmental variables in a temperate delta system. Aquat. Microb. Ecol. **31:**85–98.
58. **Stone, M.** 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B **36:**111–147.
59. **Takai, K., T. Komatsu, F. Inagaki, and K. Horikoshi.** 2001. Distribution of archaea in a black smoker chimney structure. Appl. Environ. Microbiol. **67:**3618–3629.
60. **Tetko, I. V., D. J. Livingstone, and A. I. Luik.** 1995. Neural-network studies. 1. Comparison of overfitting and overtraining. J. Chem. Inf. Comp. Sci. **35:**826–833.
61. **Thomsen, T. R., K. Finster, and N. B. Ramsing.** 2001. Biogeochemical and molecular signatures of anaerobic methane oxidation in a marine sediment. Appl. Environ. Microbiol. **67:**1646–1656.
62. **Ventura, S., M. Silva, D. Pérez-Bendito, and C. Hervás.** 1997. Computational neural networks in conjunction with principal component analysis for resolving highly nonlinear kinetics. J. Chem. Inf. Comp. Sci. **37:**287–291.
63. **Vetriani, C., H. W. Jannasch, B. J. MacGregor, D. A. Stahl, and A.-L. Reysenbach.** 1999. Population structure and phylogenetic characterization of marine benthic archaea in deep-sea sediments. Appl. Environ. Microbiol. **65:**4375–4384.
64. **White, H.** 1990. Connectionist nonparametric regression-multilayer feedforward networks can learn arbitrary mappings. Neural Netw. **3:**535–549.
65. **Wu, L., D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje, and J. Zhou.** 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl. Environ. Microbiol. **67:**5780–5790.
66. **Yan, T., M. W. Fields, L. Wu, Y. Zu, J. M. Tiedje, and J. Zhou.** 2003. Molecular diversity and characterization of nitrite reductase gene fragments (*nirK* and *nirS*) from nitrate- and uranium-contaminated groundwater. Environ. Microbiol. **5:**13–24.
67. **Zhou, J., M. E. Davey, J. B. Figueras, E. Rivkina, D. Gilichinsky, and J. M. Tiedje.** 1997. Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. Microbiology **143:**3913–3919.
68. **Zhou, J., B. Xia, D. S. Treves, L. Y. Wu, T. L. Marsh, R. V. O'Neill, A. V. Palumbo, and J. M. Tiedje.** 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl. Environ. Microbiol. **68:**326–334.