Microbial
Ecology

Application of Nonlinear Analysis Methods for Identifying Relationships Between Microbial Community Structure and Groundwater Geochemistry

Jack C. Schryver[1], Craig C. Brandt[2], Susan M. Pfiffner[3], Anthony V. Palumbo[2], Aaron D. Peacock[3], David C. White[3], James P. McKinley[4] and Philip E. Long[4]

(1) Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
(2) Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
(3) Center for Biomarker Analysis, The University of Tennessee, Knoxville, TN 37932, USA
(4) Environmental Technology Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received: 21 June 2004 / Accepted: 12 December 2004 / Online publication: 31 January 2006

Abstract

The relationship between groundwater geochemistry and microbial community structure can be complex and difficult to assess. We applied nonlinear and generalized linear data analysis methods to relate microbial biomarkers (phospholipids fatty acids, PLFA) to groundwater geochemical characteristics at the Shiprock uranium mill tailings disposal site that is primarily contaminated by uranium, sulfate, and nitrate. First, predictive models were constructed using feedforward artificial neural networks (NN) to predict PLFA classes from geochemistry. To reduce the danger of overfitting, parsimonious NN architectures were selected based on pruning of hidden nodes and elimination of redundant predictor (geochemical) variables. The resulting NN models greatly outperformed the generalized linear models. Sensitivity analysis indicated that tritium, which was indicative of riverine influences, and uranium were important in predicting the distributions of the PLFA classes. In contrast, nitrate concentration and inorganic carbon were least important, and total ionic strength was of intermediate importance. Second, nonlinear principal components (NPC) were extracted from the PLFA data using a variant of the feedforward NN. The NPC grouped the samples according to similar geochemistry. PLFA indicators of Gram-negative bacteria and eukaryotes were associated with the groups of wells with lower levels of contamination. The more contaminated samples contained microbial communities that were predominated by terminally branched saturates and branched monounsaturates that are indicative of metal reducers, actinomycetes, and Gram-positive bacteria. These results indicate that the microbial community at the site is coupled to the geochemistry and knowledge of the geochemistry allows prediction of the community composition.

Introduction

A major challenge to the successful implementation of bioremediation is to understand the structure of the indigenous microbial community and how this structure is affected by and affects the environment. As the majority of environmental microorganisms are unculturable [1], culture-independent approaches using biomolecular markers are now widely used for microbial community analysis [3, 10, 27, 28]. Often, the analysis of molecular biomarkers is based on a small set of samples generating a large number of measurements. Furthermore, field sampling of groundwater typically has uncertainty associated with the volume and geometry of the subsurface pores. The microbial cells actually sampled vary depending on pore-scale flow velocities and other details of the sampling process. The linkage between the pore water chemistry and the microbial biomarkers can therefore be modified or "smeared" during sampling. The potentially complex relationships among the large number of measurements obtained from a comparatively small number of samples makes conventional statistical analysis of such data difficult. Thus, new data analysis tools are needed to help understand these data.

Artificial neural networks (NNs) are one type of data analysis tool that can potentially accommodate the small

Correspondence to: Craig C. Brandt; E-mail: brandtcc@ornl.gov

DOI: 10.1007/s00248-004-0137-0 • Volume 51, 177–188 (2006) • © Springer Science+Business Media, Inc. 2006

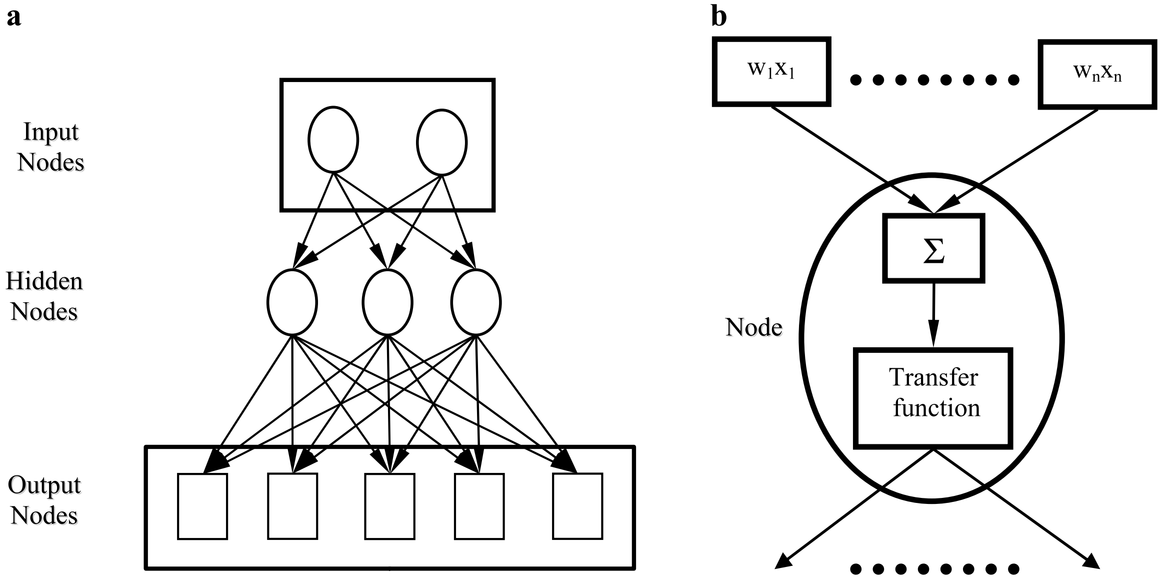

Figure 1. Schematic diagram of a hypothetical neural network. The input, hidden, and output nodes are shown in panel (a). The lines between layers represent the weights that connect the nodes. Each node consists of two parts as shown in panel (b). The first part sums the weighted inputs (w_ix_i), which is then passed to a nonlinear transfer function.

sample size and large number of measurements characteristic of biomolecular community studies [2, 6]. An NN consists of numerous processing elements called nodes. Figure 1 illustrates the architecture of a hypothetical NN consisting of two input nodes, three hidden nodes, and five output nodes. The nodes are interconnected by communication links, each with an associated weight. Each node has an internal state, called its activation level, which is a function of the inputs it receives. A node forwards its activation level via the communication links to the other nodes to which it is connected. The weights associated with the communication links represent the information used by the NN to solve a problem. An NN is trained by presenting it with a paired set of input and output patterns. The NN adjusts its weights to minimize the difference between the calculated and actual output pattern. NNs offer several advantages including (1) universal approximation for continuous functions, (2) distribution-free assumptions, (3) tolerance of missing or noisy data, and (4) applicability to feature identification and prediction problems. NNs can handle sparse, small sample size data sets and avoid the tendencies of large parameterized models to overfit data. Bishop [4], Haykin [11], and Jain *et al.* [12] provide basic introductions to NN theory and applications.

In this article we describe the application of two types of NNs to the analysis of microbial biomarker data. The first is a predictive model in which an NN is used to estimate microbial phospholipid fatty acids (PLFA) from geochemistry measurements in groundwater samples. This type of NN can be viewed as a nonlinear regression estimator. The second is a type of nonlinear principal component analysis [25], in which an NN is used to map the PLFA data to itself for the purpose of data reduction. Multiple generalized linear regression (MGLR) and linear principal components analysis were also conducted to provide a benchmark for assessing the performance of the NN techniques. This research is part of a larger project to study the biotransformation processes occurring at the uranium mill tailings site at Shiprock, NM (http://www.pnl.gov/nabir-umtra/index.stm) [7].

Materials and Methods

Site Description. The U.S. Department of Energy is responsible for uranium mill tailings under the Uranium Mill Tailings Radiation Control Act (UMTRA) of 1978. The Shiprock UMTRA site is on Navajo tribal land in San Juan County, NM, located adjacent to and partly within the town of Shiprock. The Shiprock site has been exposed to elevated levels of uranium since the 1950s. A variety of other contaminants, particularly sulfate and nitrate, also occur at the site.

The site is on the south side of the San Juan River on an elevated terrace about 21 m above the river (Fig. 2). The uranium disposal cell is on unconsolidated alluvial terrace deposits underlain by Mancos Shale. Groundwater occurs at the contact between the terrace alluvium and the upper portion of the Mancos Shale, where it has been weathered. Uranium contamination occurs in the alluvium and upper Mancos Shale on the terrace and in the floodplain alluvium. The groundwater in the terrace alluvium and upper Mancos Shale beneath the site and in

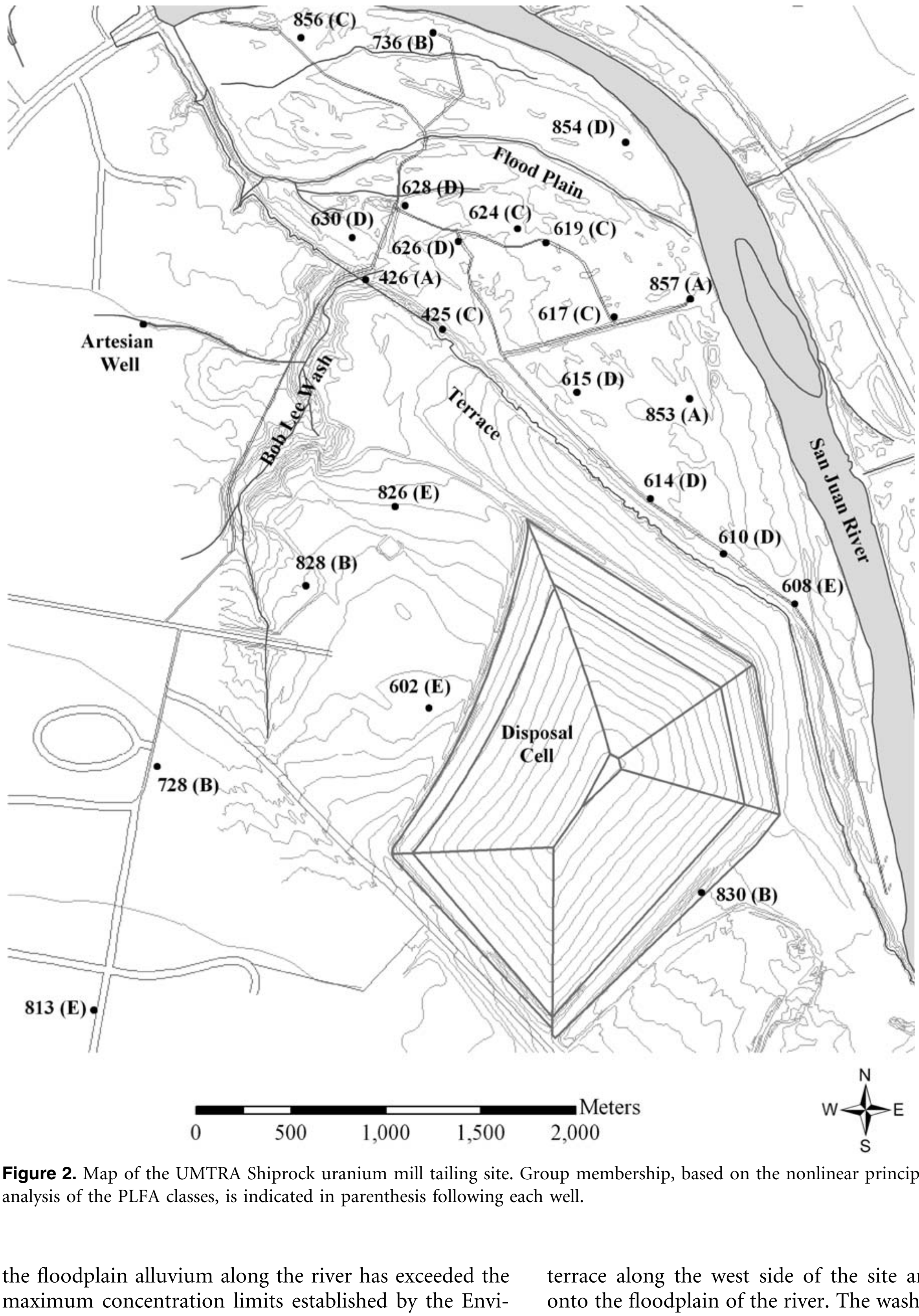

Figure 2. Map of the UMTRA Shiprock uranium mill tailing site. Group membership, based on the nonlinear principal components analysis of the PLFA classes, is indicated in parenthesis following each well.

the floodplain alluvium along the river has exceeded the maximum concentration limits established by the Environmental Protection Agency for nitrate and uranium. The volume of contaminated groundwater is estimated to be 610,000 m^3. Bob Lee Wash flows northward on the terrace along the west side of the site and flows down onto the floodplain of the river. The wash would contain flowing water ephemerally, but the lower 200 m of the wash receives a constant discharge of about 230 L min^{-1} from a potable-water artesian well. This water has created wetlands within Bob Lee Wash and at the mouth of the wash where it discharges onto the river's floodplain (proximal to wells 608, 610, 615, 617, 624, 626, 853, and 857; Fig. 2). Several drainage ditches in the floodplain contain water year-round [8].

Sample Collection. Prior to collection of the groundwater samples, all glassware was washed in a 10% (v/v) microcleaning solution (VWR Scientific, West Chester, PA, USA) and rinsed 10 times in tap water followed by 10 rinses in deionized water. Prior to use, the glassware was heated at 450°C for 4 h in a muffle furnace. The groundwater samples were collected in March 1999, using a downhole peristaltic or impeller pump. The wells were purged before sampling with a minimum of three well volumes. Between sampling events, the pump and associated tubing were decontaminated using a dilute detergent and rinsed with deionized water. Sample volumes, ranging from two to four liters each, were filtered through sterile (methanol rinsed) Anodisc™ filters (Whatman International Ltd., Maidstone, England, UK), 47 mm in diameter with a 0.2-μm pore size [22]. The filtration method was designed to ensure that all suspended particles including both sediment grains (with microorganisms attached) and individual microorganisms were retained for analysis. Filters were stored in sterilized glass petri dishes, preserved on dry ice, and shipped overnight for PLFA analysis.

Geochemical Analysis. Anions were determined using ion chromatography (Dionex Model DX-300; AS-4a column, chemical suppression, and conductivity detection) [16]. Samples were quantified against commercial standards that ranged from 0.1 to 100 mg L^{-1}. The concentration of uranium [U(VI)] was determined using a kinetic phosphorescence analyzer (Model KPA-11, Chemchek Instruments, Inc., Richland, WA, USA) with a detection limit of 0.3 μg U L^{-1} [16]. Quantitation was against NIST-traceable standards over a concentration range of 0.25–50 μg U L^{-1} in 11 steps. Samples were treated with a phosphorescent complexant and were run in batch using an autosampler. When necessary, samples were diluted and rerun so that raw results fell within the standard concentration range and yielded acceptable counting statistics. A set of standards was run at the beginning and end of each analytical sample set as an internal check on accuracy and precision. The ionic strength was calculated for each sample based upon the reported ion measurements. Tritium was analyzed by liquid scintillation counting after enrichment. Carbon was determined using a Dohrman DC-80 carbon analyzer (Rosemount Analytical, Santa Clara, CA, USA). Dissolved oxygen (DO) was measured using a flow cell during well purging. Stable (invariant) DO values typically occurred prior to completion of well purging; the minimum observed concentration was taken as the *in situ* value. The pH was measured by electrode against commercial standards.

Lipid Analysis. Glassware was washed in 10% (v/v) micro cleaner solution (VWR International, West Chester, PA, USA), rinsed 10 times in tap water, five times in deionized water, and then heated for 4 h in a muffle furnace at 450°C. The procedures for lipid extraction, separation, methylation, and analysis are described in [29]. Briefly, the lipids were extracted from the sample filters using the modified Bligh and Dyer procedure [5, 27], and the total lipid fraction was fractionated into glyco-, neutral-, and polar-lipids. All solvents were of GC grade (Fisher Scientific, Pittsburgh, PA, USA). The fatty acids in the polar fraction were converted into methyl esters using a mild alkaline methanolysis and recovered with hexane. The resulting PLFA were separated, quantified, and identified by gas chromatography-mass spectrometry. Fatty acids were identified by relative retention times, comparison with standards (Matreya, Inc., State Park, PA, USA), and mass spectrometry. The individual PLFA were reported as molar percent of the sample total. Absolute amounts of the PLFA were not used in the analysis because of possible sample biases due to differing extraction efficiencies.

Data Preparation. The original data set contained 23 samples with 182 different PLFA identified within this set of samples. To simplify interpretation, the PLFA data were combined into six classes based on the structural chemistry of the lipid moiety [13, 31, 32]. The normal saturates class (NSat) consists of straight-chain fatty acids that are generally 12–20 carbons in length. Members of the terminally branched saturates class (TBSat) have a methyl group attached to the penultimate and antepenultimate carbon atoms. The midchain branched saturates class (MBSat) has a methyl group attached to the chain at a location other than the penultimate or antepenultimate carbon atoms. The monounsaturates class (Mono) contains straight-chained fatty acids with one double bond, generally in the *cis* configuration, and cyclopropyl fatty acids. Members of the branched monounsaturates class (BMono) have one double bond and a moiety (e.g., methyl group) attached to the chain. The polyunsaturates class (Poly) contains straight-chained fatty acids with two or more double bonds. Of the 182 distinct PLFA, 4 were cyclics and 11 were unknowns, and these were excluded from the analyses because of their low abundances.

Prior to use in the predictive analyses, the geochemical and PLFA class data were rescaled to values between 0 and 1. Scaling of the geochemical data is used to prevent one input from dominating the other inputs [4]. Scaling of the PLFA data is required because the predictions of the NN model are constrained to values between -1 and 1. The scaling function we used was:

$$y_{ij} = x_{ij}^{b_i} \Big/ \left(x_{ij}^{b_i} + a_i\right)$$

where i is an index that identifies a geochemical analyte or PLFA class, j is the sample index, and a_i and b_i are parameters determined by the measurement being scaled. For this study, a_i and b_i were chosen such that the 18.6th quantile of measurement set i received a value of 0.1 and the 87.7th quantile was assigned a value of 0.9. This scaling allowed the NN model to extrapolate outside the range of the training data if necessary.

Data Analysis. We used MATLAB® version 5.2.1 (MathWorks, Inc., Natick, MA, USA) and the NETLAB toolkit [19] for the predictive NN, nonlinear principal component analyses and multiple generalized linear regression. We supplemented the NETLAB toolkit with our own modules for cross validation, pruning, and sensitivity analysis. Pearson product–moment correlations, linear principal components analysis, hierarchical cluster analysis, and analysis of variance (ANOVA) were done using SAS® version 8.1 (SAS Institute, Cary, NC, USA). Ward's method was used for the cluster analysis as it is space conserving and has been recommended for community analysis [15].

The initial architecture of the predictive NN consisted of eight input nodes representing the eight geochemical parameters, six output nodes representing the six PLFA classes, and twelve hidden nodes. The hyperbolic tangent (tanh) was used as the transfer function in the hidden layer and the logistic function was used in the output layer. The input, hidden, and output layers were fully connected, and the connection weights were adjusted by scaled conjugate gradient optimization [4]. A backward elimination pruning procedure was used to select the optimal number of input and hidden nodes in the NN [18]. Each input and hidden node was systematically removed from the NN and the mean squared error (MSE) of the reduced network was calculated using all of the data. The node producing the smallest increase in the MSE was deleted, and the new NN was retrained. This procedure was repeated until only a nominal network remained.

Training an NN model with many parameters can result in unwarranted complexity in the nonlinear mapping (overfitting). We used weight decay to avoid overfitting of the NN model during training [14]. The basic idea of weight decay is to modify the error function used during training to encourage smoother network mappings. A penalty term for excessive complexity is added to the error function. There is also a gain parameter that determines the importance of the error minimization. The gain parameter is selected to produce smooth mappings, but not to the extent that training error becomes unacceptably large. An important effect of weight decay is that it should reduce the verification error by producing networks that are more generalizable.

Linear models can be extended to encompass a more comprehensive family of models. Models that are linear in their inputs, but nonlinear in their outputs, are called generalized linear models. The logistic function is a commonly used example of such a model. In a multiple MGLR, several inputs are used to predict a nonlinear output. We performed MGLR by using the same NETLAB functions that were used to train artificial NN. The MGLR network architecture used a single node in the hidden layer with a linear transfer function and a logistic transfer function in the single output node. In the multivariate (multiple output) case, one model was required for each output.

We also conducted a nonlinear principal components (NPC) analysis on the PLFA data. The NPC model contained two input, representing the hidden features, four hidden and six output nodes representing the six PLFA classes. The scores obtained from a two-component linear principal components analysis were used as the initial inputs in the NPC. The backpropagation algorithm in NETLAB [19] was modified to adjust the values of the input nodes in the NPC model during training. During training, the input nodes were adjusted to reduce the backpropagated error using the learning algorithm in a manner analogous to the adjustment of the weights in the predictive NN models. After training, the final values of the two input nodes represent the hidden features contained in the PLFA data [25].

Cross Validation Scheme. We employed a bootstrapped cross validation procedure to evaluate the performance of the MGLR and predictive NN models. The data set was randomly divided into two parts. The training part, consisting of 90% of the data, was used to train the NN and MGLR models. The remaining 10% of the data was used to verify the performance of the models. The procedure was repeated 101 times with replacement, and the training and verification (generalization) errors were reported for the best, worst, and median cases.

Sensitivity Analysis. A sensitivity analysis was performed on the median-performing trained NN to determine the relative importance of the geochemistry measurements in predicting the PLFA classes. In this study, we defined sensitivity as the increase in MSE resulting from the removal of a geochemical analyte as an input variable [18]. The MSE increase for each combination of geochemical input and PLFA output was estimated by substituting the mean value of the geochemistry variable in each sample. This approach

Table 1. **Pearson product–moment correlations of the geochemical and PLFA class data**

	NSat	*MBSat*	*TBSat*	*BMono*	*Mono*	*Poly*	*Inorg. C*	*Nitrate*	*Tritium*	*Uranium*	*Ionic strength*	*Chloride*	*Sulfate*	*DOC*
NSat	1.000													
MBSat	−0.133 (0.544)	1.000												
TBSat	0.354 (0.097)	0.465 (0.026)	1.000											
BMono	−0.378 (0.075)	0.425 (0.043)	0.071 (0.749)	1.000										
Mono	−0.410 (0.052)	−0.502 (0.015)	−0.640 (0.001)	−0.011 (0.959)	1.000									
Poly	−0.605 (0.002)	−0.301 (0.162)	−0.483 (0.020)	0.003 (0.989)	0.059 (0.789)	1.000								
Inorg. C	−0.169 (0.440)	0.685 (<0.001)	0.050 (0.822)	0.230 (0.290)	−0.415 (0.049)	−0.007 (0.975)	1.000							
Nitrate	−0.014 (0.950)	0.483 (0.020)	−0.050 (0.820)	0.219 (0.315)	−0.070 (0.750)	−0.192 (0.380)	0.212 (0.332)	1.000						
Tritium	−0.074 (0.742)	0.647 (0.001)	−0.038 (0.868)	0.275 (0.212)	−0.234 (0.295)	−0.128 (0.569)	0.564 (0.006)	0.785 (<0.001)	1.000					
Uranium	0.145 (0.510)	0.420 (0.046)	0.082 (0.710)	0.102 (0.643)	−0.287 (0.185)	−0.179 (0.413)	0.537 (0.008)	0.347 (0.104)	0.468 (0.028)	1.000				
Ionic strength	−0.004 (0.987)	0.618 (0.002)	0.038 (0.863)	0.196 (0.371)	−0.339 (0.114)	−0.113 (0.608)	0.741 (<0.001)	0.602 (0.002)	0.668 (0.001)	0.830 (<0.001)	1.000			
Chloride	0.017 (0.939)	0.546 (0.007)	0.058 (0.791)	0.121 (0.583)	−0.287 (0.184)	−0.161 (0.464)	0.757 (<0.001)	0.442 (0.035)	0.502 (0.017)	0.827 (<0.001)	0.941 (<0.001)	1.000		
Sulfate	−0.019 (0.932)	0.600 (0.002)	0.061 (0.783)	0.163 (0.458)	−0.368 (0.084)	−0.077 (0.726)	0.784 (<0.001)	0.474 (0.022)	0.701 (<0.001)	0.832 (<0.001)	0.978 (<0.001)	0.958 (<0.001)	1.000	
DOC	−0.182 (0.406)	0.579 (0.004)	−0.029 (0.896)	0.214 (0.328)	−0.289 (0.181)	0.011 (0.959)	0.907 (<0.001)	0.350 (0.101)	0.570 (0.006)	0.718 (<0.001)	0.866 (<0.001)	0.898 (<0.001)	0.888 (<0.001)	1.000

Significance probability of the correlation under the null hypothesis that the statistic is zero is given in parenthesis.

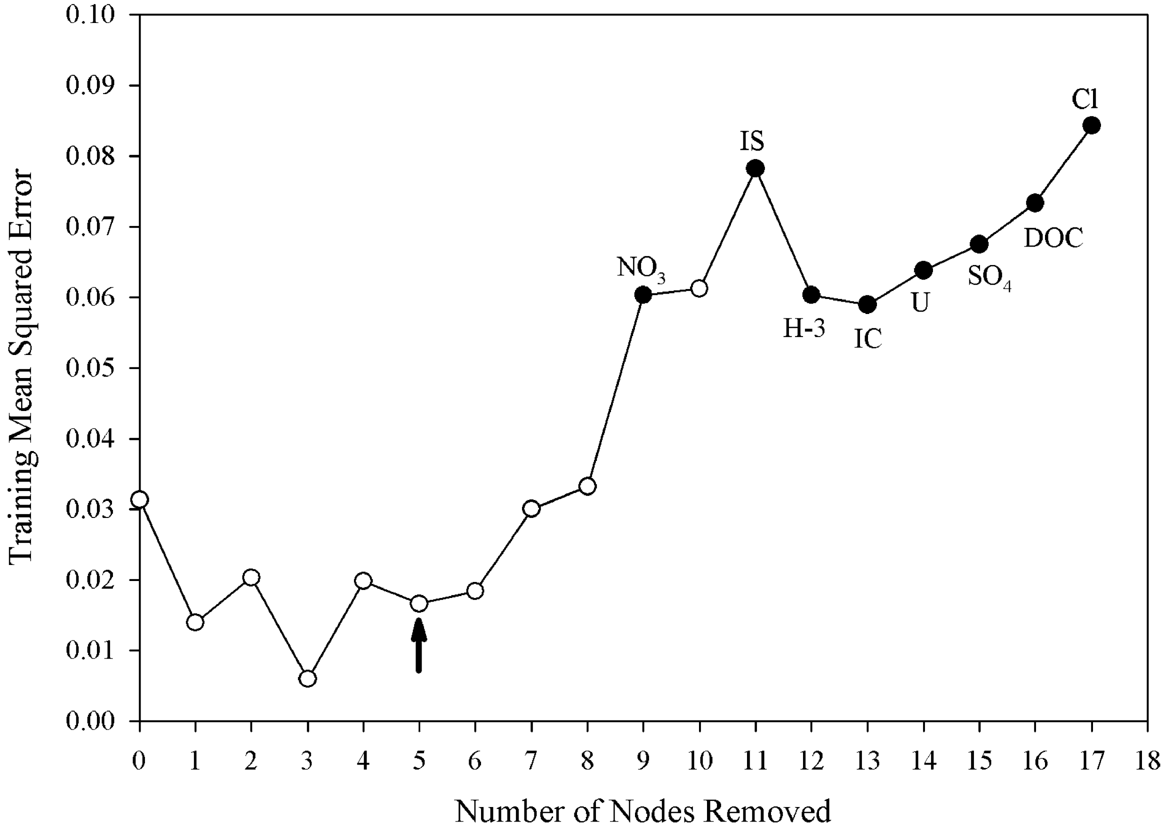

Figure 3. Change in the training error of the predictive NN model with removal of hidden and input nodes. The initial model contained eight input nodes, twelve hidden nodes, and six output nodes. *Solid circles* denote input variables and the *open circles* represent the hidden nodes. The *arrow* indicates the architecture selected.

decoupled the input from the output without affecting the calibration of the model weights and biases. Each MSE increase was divided by the total of the MSE increases for all of the geochemical inputs within each PLFA class. The resulting normalized indices provide a rank ordering of the geochemical sensitivities for each PLFA class.

Results

Correlation Analysis. There was a wide range of community composition observed in the 23 samples and there were many significant correlations among PLFA classes and geochemical parameters (Table 1). PLFA classes were typically moderately intracorrelated (e.g., see upper left portion of Table 1) with the highest negative correlation between the terminally branched saturates (TBSat) and the monounsaturates (Mono) ($R = -0.640$). In contrast, the PLFA and geochemical variables (e.g., see lower left portion of Table 1) were generally not highly correlated, except for the midchain branched saturates (MBSat) which were positively correlated with the geochemical variables. Also, monounsaturated (Mono) PLFA were negatively correlated with a few of the geochemical variables. There were some strong correlations ($R > 0.850$) among the geochemical variables (e.g., see lower right region of Table 1). In particular, dissolved organic carbon (DOC) was strongly correlated with inorganic carbon, nitrate, and sulfate. Ionic strength, sulfate, and chloride were all strongly intercorrelated (Table 1).

Table 2. Performance of the neural network (NN) and multiple generalized linear regression (MGLR) predictive models

		Mean squared error		
Case	*Type of error*	*NN—No weight decay*	*NN—weight decay*	*MGLR*
Best	Training	0.1132	0.3907	0.4987
	Verification	0.1058	0.1866	0.2557
Median	Training	0.1509	0.3669	0.4494
	Verification	0.6025	0.6793	0.8194
Worst	Training	0.0839	0.2822	0.4041
	Verification	2.9912	2.1942	20.341

Performance is given for the best, median, and worst cases based on 101 iterations of the cross validation bootstrapping method. Smaller numbers indicate better performance.

Development of the Predictive NN. The initial network contained eight input (geochemical), twelve hidden, and six output (PLFA) nodes. We employed a pruning method to help select the number of hidden nodes to use in the predictive NN model. This method tracks the change in training error (mean of the squared differences between the predicted and observed outputs) as hidden and input (geochemical) nodes are systematically removed from the model. The results of the pruning study are shown in Fig. 3. The method eliminated eight hidden nodes and no input nodes before the training mean squared error increased dramatically. We made a conservative selection of seven hidden nodes (elimination of five hidden nodes rather than eight).

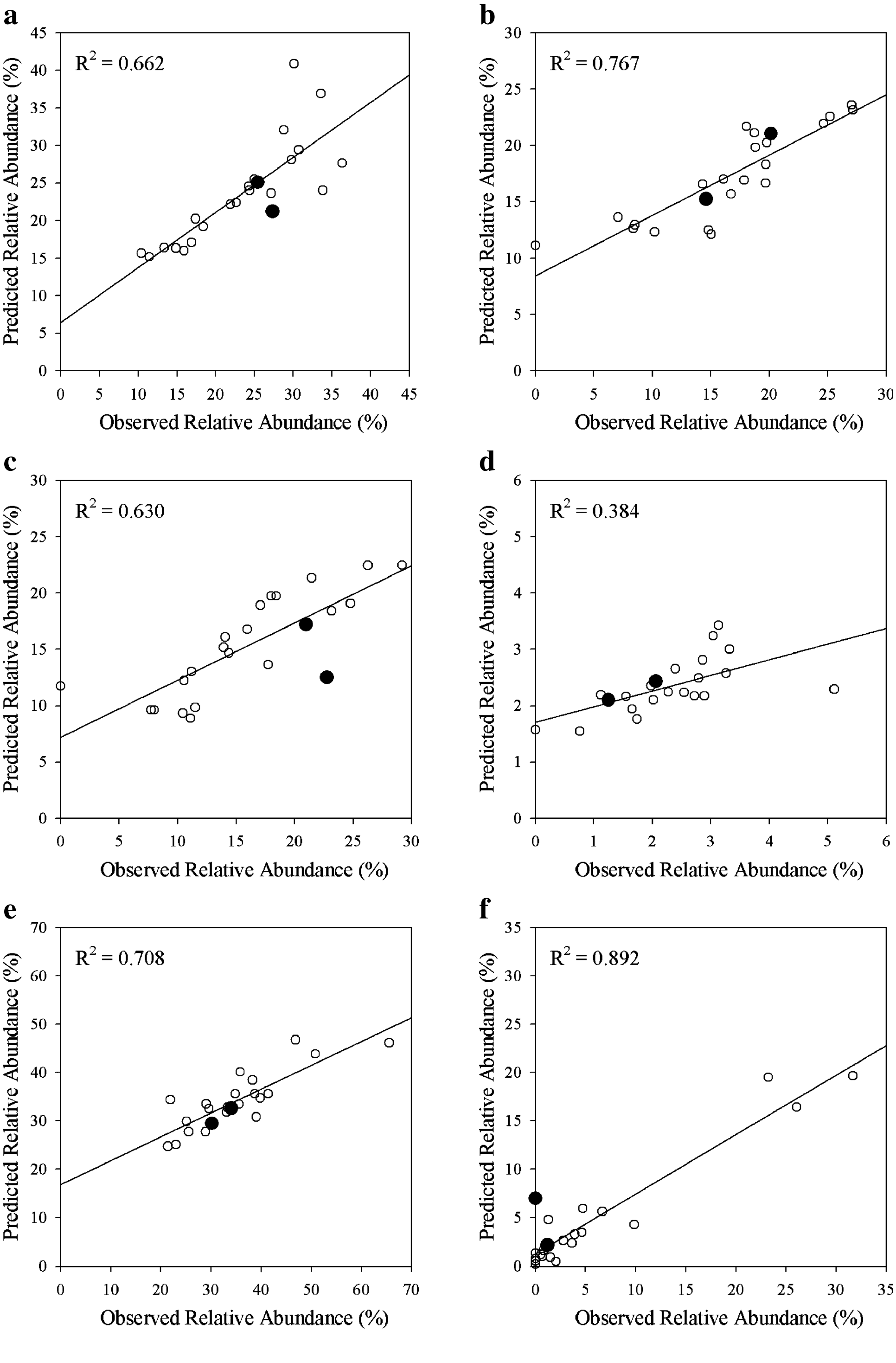

Figure 4. Observed–predicted scatter plots for the median-performing NN without weight decay for (a) NSat, (b) MBSat, (c) TBSat, (d) BMono, (e) Mono, and (f) Poly PLFA classes. The *solid line* represents the regression line. The training data are denoted by the *open circles* and the verification data are identified by *solid circles.*

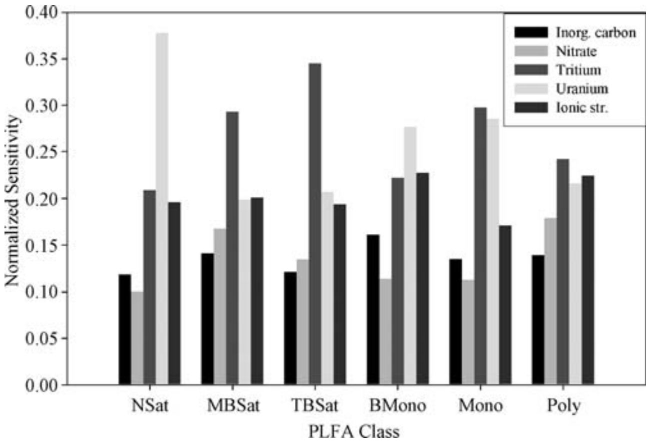

Figure 5. Sensitivities of geochemical variables in predicting PLFA classes with the median-performing NN model. A larger sensitivity value indicates that the PLFA abundance is more sensitive to small changes in the concentration of the corresponding geochemical variable.

To further simplify the NN model, we used correlation analysis to identify redundant inputs (Table 1). Dissolved organic carbon (DOC) was found to be almost completely redundant (highly correlated) with inorganic carbon ($R = 0.907$); therefore, DOC was eliminated from the predictive model. Sulfate, chloride, and ionic strength were also highly correlated (Table 1), and we chose to retain only ionic strength for analysis. The final network architecture contained five geochemical inputs (nitrate, ionic strength, inorganic carbon, uranium, and tritium), seven hidden nodes, and six PLFA outputs (NSat, TBSat, MBSat, Mono, BMono, and Poly).

Predictive Model Analysis. Model performance was assessed by calculating the mean squared error (MSE) of the predicted *versus* observed results. The training MSE was calculated for the 90% sample subset from each iteration of the cross validation scheme. The remaining 10% of the data from each bootstrap sample was used to estimate the verification error. The NN model (NN—No Weight Decay in Table 2) always yielded a lower MSE than the MGLR model. The consistently better results obtained with the NN model compared with MGLR models indicate that there were likely nonlinear influences of the geochemistry on the microbial community composition.

Although a higher verification error than training error is expected, a large difference between these errors may indicate an overfitting problem. Weight decay has been proposed as a method to decrease the verification error of NN models [14]. The introduction of weight decay (Table 2, NN—Weight Decay) did not lower the verification MSE of the NN models, indicating that overfitting may not be a problem with the NN architecture for this data set. Figure 4 shows plots of the observed *versus* predicted PLFA class abundances for the median performing NN model without weight decay. The fit between the observed and predicted values, as measured by linear regression, was high for all PLFA classes except BMono. It is interesting to note that this class did not show any significant correlations with the geochemistry variables (Table 1).

The sensitivity analysis of the predictive NN model indicated that, in general, tritium and uranium were the most important geochemical factors in predicting PLFA classes (Fig. 5). In this analysis, a large increase in the MSE after substituting the mean value of a geochemical variable was associated with high sensitivity. The TBSat and MBSat classes, which are indicative of Gram-positive bacteria, actinomcyetes, and sulfate reducers [30, 32], were most responsive to tritium. Also most responsive to tritium were the Mono class, a marker for Gram-negative bacteria, and the Poly class, an indicator of eukaryotes. The NSat class, ubiquitous in prokaryotes and eukaryotes, was most responsive to uranium as was the BMono class, which is indicative of metal reducers. Ionic strength was the next most important contributor. The least sensitive geochemical variables were nitrate and inorganic carbon, both of which were the least sensitive predictor for three of the six PLFA classes.

Principal Components Analysis. The nonlinear principal component (NPC) analysis of the PLFA class data outperformed the linear principal components analysis. The variance explained was 70% for two linear components and 91% for two nonlinear components. We used a hierarchical cluster analysis of the NPC

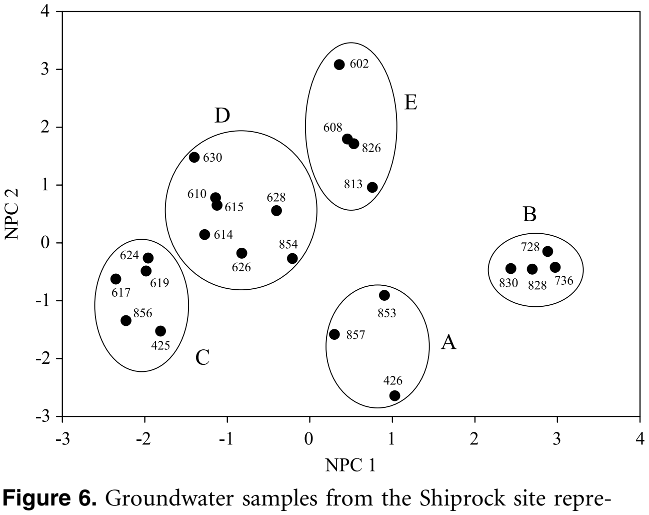

Figure 6. Groundwater samples from the Shiprock site represented as a plot of coordinates from the nonlinear principal component analysis of the PLFA data. Groups, based on a hierarchical cluster analysis of the NPC coordinates, are outlined and labeled A, B, C, D, and E.

Table 3. Means and standard errors (SE) of PLFA and geochemical measurements for groups identified in the nonlinear principal component analysis of the PLFA class data

	Group				
	A (n = 3)	*B (n = 4)*	*C (n = 5)*	*D (n = 7)*	*E (n = 4)*
Measurement	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
PLFA					
NSat (%)	29.48 (3.47)	12.49 (1.00)	29.92 (1.03)	26.16 (1.67)	17.15 (0.51)
MBSat (%)	7.37 (4.34)	10.46 (1.50)	17.05 (1.10)	18.46 (0.60)	26.01 (0.64)
TBSat (%)	7.35 (3.68)	9.31 (0.86)	24.57 (1.40)	18.47 (1.33)	14.31 (1.35)
BMono (%)	1.52 (0.84)	2.21 (0.13)	1.28 (0.17)	2.91 (0.40)	3.09 (0.10)
Mono (%)	45.83 (3.23)	43.61 (7.35)	26.70 (2.29)	32.36 (2.05)	32.39 (4.01)
Poly (%)	2.72 (2.02)	21.42 (5.84)	0.48 (0.22)	1.50 (0.53)	4.93 (1.79)
Geochemistry					
Inorganic carbon (mg L^{-1})	72.2 (5.5)	79.0 (32.1)	139.4 (33.2)	127.7 (28.0)	356.8 (63.5)
Nitrate (mmol L^{-1})	0.99 (0.99)	3.61 (2.15)	4.87 (2.13)	30.27 (11.27)	42.23 (28.03)
Tritium (mg L^{-1})	9.4 (1.9)	9.3 (2.5)	13.9 (3.2)	22.0 (10.6)	51.7 (8.4)
Uranium (mmol L^{-1})	1.07 (0.12)	0.95 (0.35)	2.89 (0.99)	5.71 (2.14)	5.39 (2.29)
Ionic strength	252 (45)	336 (99)	540 (132)	897 (281)	1219 (87)

coordinates to identify groups of stations. The station groups are shown in a plot of the sample scores for the NPC (Fig. 6), and the mean and standard error of the PLFA and geochemistry measurements for the groups are presented in Table 3.

Gram-negative bacteria were higher in the two groups of stations (A and B) with the lowest contamination levels. Also, the Poly class, indicative of eukaryotes, was higher in the group of wells that had the lowest contaminant levels (group B). The TBSat class was higher in the more contaminated groups (C, D, and E). Increases in the TBSat class have been documented at petroleum and metal-contaminated sites [3, 10, 22, 24]. At such sites, the TBSats are indicative of Gram-positive bacteria and at some sites these PLFA are attributed to anaerobic Gram-negative bacteria. Kostka (pers. comm.) demonstrated that Gram-positive dissimilatory iron-reducing bacteria were found at this UMTRA site. The highest levels of MBSat, indicative of actinomycetes and sulfate-reducing bacteria (SRB), were observed in the highly contaminated group E stations. This would coincide with the detection of SRB using molecular methods by Chang *et al.* [7]. The increased abundances of TBSat and MBSat in the more contaminated samples indicate that the site is predominated by anaerobic processes such as metal and sulfate reduction, and involve both Gram-negative and Gram-positive bacteria. The BMono class, indicating metal reducers and SRB, made up significantly greater percentages of the total PLFA in the groups of stations with the highest (D and E) and moderate (B) levels of contamination. It appears that the microbial community in the more contaminated areas is dominated by Gram-positive and some anaerobic Gram-negative organisms with the capacity for metal and sulfate reduction. The community appears to shift to more Gram-negative and eukaryotic influences as the contamination decreases.

Analysis of variance revealed significant differences between the station groups for some of the geochemical measurements (Table 4). The overall tests of group differences were highly significant for inorganic carbon (R^2 = 0.65, F = 8.27, p = 0.001) and tritium (R^2 = 0.52, F = 4.97, p = 0.007) and marginally significant for ionic strength (R^2 = 0.40, F = 3.02, p = 0.045). We performed a Tukey's studentized range test on the group means for inorganic carbon, tritium, and ionic strength to identify significant pairwise group differences. For inorganic carbon, group E had a significantly (p = 0.05) higher mean than any of the other groups. Group E also had a mean tritium that was significantly (p = 0.05) higher than groups A, B, and C. The mean tritium for group D was not significantly different from any of the other groups. No significant differences in the group means were found for ionic strength.

Group assignments are shown on the site map (Fig. 2). Group A was relatively uncontaminated

Table 4. Overall analysis of variance results of geochemistry measurements for groups identified in the nonlinear principal components analysis of the PLFA classes

Measurement	R^2	*F*	*Significance*
Inorganic carbon	0.65	8.27	0.001
Nitrate	0.28	1.80	0.173
Tritium	0.52	4.97	0.007
Uranium	0.25	1.53	0.235
Ionic strength	0.40	3.02	0.045

The significance of the overall test (significance) is the probability that the observed differences between the one or more group means is due to chance alone.

(Table 3) and these wells were primarily located at the fringe of the site along the river (right center of Fig. 2) except for well 426, which was near the other border of the site (left center of Fig. 2). Wells in group B are located on the fringe of the site (lower left, lower right, and top of Fig. 2). This group was characterized by low values (Table 3) for all the geochemical variables and also represents relatively uncontaminated wells. The most contaminated samples (Table 3) appear at the top center of Fig. 6 (group E) and were primarily located at fringe of the cap toward the center of the site (Fig. 2). The next most contaminated group (group D) primarily consists of a transect through the center of the site (Fig. 2). This was paralleled by a group of wells (group C) on either side that is slightly closer to the river (Fig. 2) near the center with the same orientation and lower contamination (Table 3).

Discussion

There appears to be a complex set of relationships between the PLFA and geochemical measurements collected in this study. These relationships are difficult to interpret from examination of the individual sample lipid profiles or the pairwise Pearson product–moment correlations. The predictive NN models explained more than 60% of the variability for five of the six lipid classes (Fig. 4). The predictive NNs estimated membership in the six PLFA classes better than the MGLR without evidence of overfitting. This comparative result suggests that significant nonlinear relationships exist among microbial biomarkers and geochemistry.

The sensitivity analysis indicated that tritium and uranium were the most important geochemical factors in predicting biomarker abundance with the NN model. Simple correlation analysis did not reveal the importance of these factors. Previous linear analysis of clone libraries taken at the site indicated that uranium and sulfate were important in predicting the abundance of one of the clusters of sulfate-reducing bacteria [7]. Predictions of the other clusters were not very successful. The indications of nonlinearity that we observed from the PLFA data may indicate that other clone library classes might have been better predicted with nonlinear techniques.

Although the nature of the relationships between geochemistry and PLFA is not identified by this method, it is possible to use the NN results as a guide to more detailed geochemical interpretation. For example, tritium serves as a tracer for river water and to a lesser extent the contaminant effluent from the tailings [17]. The microbial community associated with tritium may therefore reflect the impact of river water or the effect of mixing of river water with the tailings pile effluent. The relationship between the NSat and uranium suggests that there may be a diverse microbial community associated with the uranium-contaminated effluent.

Neural networks are sometimes criticized as a "black box" method of prediction that yields little insight into relationships among input and output variables. Several methods have been proposed to help understand the sensitivity of the model outputs to the various input variables (e.g., [20, 26]). Saltelli *et al.* [23] review several measures of sensitivity. One popular index is calculated by measuring the effects on the output variables (e.g., PLFA class) produced by small perturbations of an input (e.g., geochemical) variable. A global measure can be obtained by taking a mean of the absolute values of these individual sensitivity values. However, this measure is biased because it yields greater sensitivities for larger output values. Another index can be generated by multiplying the previous index by the input/output ratio. This normalizes the index but results in a bias for smaller output values. Another commonly used index is calculated as the partial derivative of the output with respect to an individual input variable at a given value. Other indices are also frequently used [23]. Unfortunately, the variety of sensitivity indices often means that the definition and interpretation of importance is limited to the context of a specific problem.

The sensitivity index we employed in this article is a one-factor-at-a-time method based on the model error associated with the scaled outputs instead of the output values themselves. Conceptually, this index is more compatible with the trained ANN model, because it measures the extent to which variation in the inputs affects the accuracy of the prediction. Although it is a local index, it has the advantage of deemphasizing the contribution to importance in those regions where the output is not well predicted by the inputs.

We found artificial neural networks to be a robust and powerful class of data analysis tools for uncovering complex relationships in the data. These tools can be used to reduce the dimensionality of complex microbial ecological data or to predict the microbial community structure from environmental data. Although additional samples would improve the quality of the predictive models, the techniques described in this article appear to be capable of extracting meaningful information from measurement-rich analyses conducted on limited sets of samples. Furthermore, these and other related techniques can be applied to other types of biomarker data such as the abundance of clone libraries [21] and terminal restriction fragment length polymorphisms [9].

Acknowledgments

This research was funded by the U.S. Department of Energy, Office of Science, Natural and Accelerated Bioremediation Research Program (NABIR), and the

UMTRA Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

References

1. Amann, RI, Ludwig, RW, Schleifer, K-H (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol Rev 59: 143–169
2. Almeida, JS (2002) Predictive non-linear modeling of complex data by artificial neural networks. Curr Opin Biotechnol 13: 72–76
3. Bååth, E, Diaz-Ravina, M, Frostegård, A, Campbell, CD (1998) Effect of metal-rich sludge amendments on the soil microbial community. Appl Environ Microbiol 64: 238–245
4. Bishop, CM (1996) Neural Networks for Pattern Recognition. Clarendon Press, Oxford
5. Bligh, EG, Dyer, WJ (1959) A rapid method of total lipid extraction and purification. Can J Biochem Physiol 37: 911–917
6. Brandt, CC, Schryver, JC, Pfiffner, SM, Palumbo, AV, Macnaughton, S (1999) Using artificial neural networks to assess changes in microbial communities. In: Leeson, A, Alleman, BC (Eds.) Bioremediation of metals and inorganic compounds. Battelle Press, Columbus, pp 1–6
7. Chang, Y-J, Peacock, AD, Long, PE, Stephen, JR, McKinley, JP, Macnaughton, SJ, Hussain, AKMA, Saxton, AM, White, DC (2001) Diversity and characterization of sulfate-reducing bacteria in groundwater at the uranium mill tailings site. Appl Environ Microbiol 67: 3149–3160
8. Department of Energy (DOE) (2000) Final site observational work plan for the Shiprock, New Mexico, UMTRA project site. GJO-2000-169-TAR, MAC-GWSHP 1:1. Rev 2. U.S. Department of Energy, Grand Junction, CO
9. Dollhopf, SL, Hashsham, SA, Tiedje, JM (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
10. Frostegård, A, Tunlid, A, Bååth, E (1996) Changes in microbial community structure during long-term incubation in two soils experimentally contaminated with metals. Soil Biol Biochem 28: 55–63
11. Haykin, SS (1999) Neural Networks: A Comprehensive Foundation. Prentice-Hall, New Jersey
12. Jain, AK, Duin, RPW, Mao, JC (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22: 4–37
13. Kaneda, T (1991) Iso-fatty and antieiso-fatty acids in bacteria—-biosynthesis, function, and taxonomic significance. Microbiol Rev 55: 288–302
14. MacKay, DJC (1992) Bayesian interpolation. Neural Comput 4: 415–447
15. McCune, B, Grace, JB (2002) Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon
16. McKinley, JP, Stevens, TO, Fredrickson, JK, Zachara, JM, Colwell, KB, Wagnon, KB, Smith, SC, Rawson, SA, Bjornstad, BN (1997) Biogeochemistry of anaerobic lacustrine and paleosol sediments within an aerobic unconfined aquifer. Geomicrobiol J 14: 23–29
17. McKinley, JP, Long, PE, Elias, DA, Krumholz, LR (2001) Chemical evidence for uranium bioreduction at Shiprock, New Mexico. EOS Transactions, American Geophysical Union, 82(47) Fall Meeting Supplement Abstract B42B-0133
18. Moody, J (1992) Prediction risk and architecture selection for neural networks. In: Cherkassky, V, Friedman, JH, Wechsler, H (Eds.) From Statistics to Neural Networks: Theory and Pattern Recognition Applications, NATO ASI Series F. Springer-Verlag, Berlin
19. Nabney, IT (2002) NETLAB: Algorithms for Pattern Recognition. Springer-Verlag, London
20. Olden, JD, Jackson, DA (2002) Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model 154: 135–160
21. Palumbo, AV, Schryver, JC, Fields, MW, Bagwell, CE, Zhou, J, Yan, T, Liu, X, Brandt, CC (2004) Coupling of functional gene diversity and geochemical data from environmental samples. Appl Environ Microbiol 70: 6525–6534
22. Pfiffner, SM, Palumbo, AV, Gibson, T, Ringelberg, DB, McCarthy, JF (1997) Relating groundwater and sediment chemistry to microbial characterization at a BTEX-contaminated site. Appl Biochem Biotechnol 63–65: 775–788
23. Saltelli, A, Tarantola, S, Campolongo, F (2000) Sensitivity analysis as an ingredient of modeling. Stat Sci 15: 377–395
24. Stephen, JR, Chang, Y-J, Gan, YD, Peacock, A, Pfiffner, SM, Barcelona, MJ, White, DC, Macaughton, SJ (1998) Microbial characterization of a JP-4 fuel-contaminated site using a combined lipid biomarker/polymerase chain reaction-denaturing gradient gel electrophoresis (PCR-DGGE)-based approach. Environ Microbiol 1: 231–241
25. Tan, S, Mavrovouniotis, ML (1996) Reducing data dimensionality through optimizing neural-network inputs. AIChE J 41: 1471–1480
26. Wang, W, Jones, P, Partridge, D (2001) A comparative study of feature-salience ranking techniques. Neural Comput 13: 1603–1623
27. White, DC, Davis, WM, Nickels, JS, King, JD, Bobbie, RJ (1979) Determination of the sedimentary microbial biomass by extractable lipid phosphate. Oecologia 40: 51–62
28. White, DC, Stair, JO, Ringelberg, DB (1996) Quantitative comparisons of *in situ* microbial biodiversity by signature biomarker analysis. J Ind Microbiol 17: 185–196
29. White, DC, Ringelberg, DB (1998) Signature lipid biomarker analysis. In: Burlage, RS, Atlas, R, Stahl, D, Geesey, G, Sayler, G (Eds.) Techniques in Microbial Ecology. Oxford University Press, New York
30. White, DC, Pinkart, HC, Ringelberg, DB (1997) Biomass measurements: biochemical approaches. In: Hurst, CJ, Knudsen, GR, McInerney, MJ, Stetzenbach, LD, Walter, MV (Eds.) Manual of Environmental Microbiology. ASM Press, Washington, DC
31. Wilkinson, SG (1988) Gram-negative bacteria. In: Ratledge, C, Wilkinson, SG (Eds.) Microbial Lipids. Academic Press, London
32. Zelles, L (1999) Fatty acid patterns of phospholipids and lipopolysaccharides in the characterization of microbial communities in soil: a review. Biol Fertil Soils 29: 111–129