Using Artificial Neural Network Tools to Analyze Microbial Biomarker Data

C. C. Brandt[1](PI), J. C. Schryver[1], J. S. Almeida[2], S. M. Pfiffner[3], and A.V. Palumbo[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Medical University of South Carolina, Charleston, SC; [3]University of Tennessee, Knoxville, TN

New data analysis tools are needed to help understand biomolecular community characterization data. In our project, we have developed artificial neural network (ANN) tools for relating changes in microbial biomarkers to the concentration of heavy metals and radionuclides. ANNs are nonlinear pattern recognition methods that can learn from experience to improve their performance. Supervised learning is used for prediction and unsupervised learning provides dimension reduction. We have successfully applied these techniques to the analysis of membrane lipids and nucleic acid biomarker data from both laboratory and field studies. Although ANNs typically outperform linear data analysis techniques, the user must be aware of several issues to ensure that the analysis results are not misleading. This poster will provide guidance in the proper use of ANNs.

Overfitting is a potential problem when using ANNs to analyze data in which the ratio of the number of biomarkers to samples is large. The result of overfitting is that the ANN generalizes poorly to new data that were not used during training. We have implemented several techniques to help reduce the problem of overfitting. The new techniques include: (1) a sensitivity-based pruning algorithm to help identify efficient ANN models for nonlinear prediction, (2) cross validation, (3) weight decay, and (4) a global optimization (e.g., simulated annealing) technique to select parsimonious ANNs that generalize well to new data sets. We have also implemented an unsupervised learning technique called the input training technique that can be used to reduce the dimensionality of the input variables for visualization. This method is more efficient than a popular autoassociative ANN technique used for the same purpose.

Another challenge in using ANNs is the interpretation of the analysis results. ANNs are often treated as a black box, which is unsatisfactory for understanding relationships between sets of variables. For example, we often want to estimate the relative sensitivity importance of a set of explanatory variables for a given set of predicted variables. However, importance or sensitivity is often an ill-defined concept outside a specific problem context. Many measures are only locally defined or are biased toward large or small values. An unbiased local model-error-based importance index we are currently using is based on the amount of degradation in the accuracy of ANN predictions when a particular input is effectively removed from the model. Inputs that are “important” predictors should severely degrade ANN performance when their effects are removed from the model. We discuss several measures of importance/sensitivity and offer guidance on their use.