G104 Bioinformatics Tools and R-Language Programming for the Classification of Soils

Julian L. Mendel, MSc, 10005 SW 141 Court, Miami, FL 33186; Natalie Damaso, 6151 W 22nd Lane, Hialeah, FL 33016; and DeEtta Mills, PhD, Florida International University, OE 167, Biological Sciences, 11200 SW 8th Street, Miami, FL 33199*

After attending this presentation, attendees will learn and better understand some of the principals of soil forensics. There are vast amounts of complex data that can be collected from soil samples and, by using bioinformatics tools and the R-language programming, it is possible to deconvolve these complexities and establish classification schemes for soil that can assist in its provenance.

This presentation will impact the forensic science community through the implementation of the methods and analyses described. Bioinformatics tools, once learned and understood, provide a simple and rapid means of analyzing large datasets, such as those obtained from soil samples. The potential of this study also lies in the creation of a searchable database for soil forensic identification.

Soil is a highly dynamic substrate consisting of many physical and chemical properties as well as a rich array of biological diversity. These soil properties make it a very useful and informative type of evidence for forensic investigators if all of this information can be analyzed. The ecological hypothesis states that soil type, which is characterized by the physical and chemical properties of the soil, is also highly correlated to the soil biota and microbial habitat.[1] Therefore, soil metagenomic profiling can provide a rapid tool for the discrimination and classification of soil. Previous studies have demonstrated the usefulness of physical and chemical properties of soil for accurate classification using elemental analysis but it was shown that 16SrRNA biotic profiling was more effective.[2] Another soil study applied bioinformatics tools such as support vector machines and K-nearest neighbors to distinguish soil bacterial community-pattern differences in soils from Idaho using the hypervariable 16SrRNA domains. These methods were able to predict the classification of soil (location and/or treatment) from the microbial profiles with high accuracy.[3]

Machine-learning tools are widely used for classification purposes and there are both supervised and unsupervised methods. These methods train on a known data set to recognize patterns; subsequently, unknown samples can be tested against the training set and classified as to their similarity. The current study used physical and chemical properties of ~1,270 soil samples from various geographic locations across Miami-Dade County, Florida. Biotic profiles were generated by Polymerase Chain Reaction (PCR) from the taxonomic soil groups- — bacteria, archaea, fungi, and plant — for each soil sample. The combined dataset of chemical/physical properties and microbial profile data were used to train the machine-learning algorithms that were programmed using R. In particular, random forest, decision trees, and neural networks were implemented and compared for accuracy of classification. When using abiotic and biotic data separately, the classification accuracy was not as high compared to when both were concatenated. This was true for both random forests and decision trees but, curiously, not of neural networks. Random forest analysis, when combining both biotic and abiotic data, was able to classify “test” samples to their known soil type with 100% accuracy;, decision tree analysis had a 98% accuracy rate. Neural networks were unable to classify soils accurately when both biotic and abiotic data were combined, with a low rate of 35%. Although, when tested separately, both abiotic and biotic data were able to classify soils with good accuracy, using neural networks. This approach demonstrated that the use of both chemical/physical properties as well as microbial community patterns can provide higher accuracy in the discrimination and classification of soils and provides a rapid method of analysis for forensic investigators to determine soil provenance. Future work will involve the implementation of K-nearest neighbor and support vector machine algorithms to add to the suite of learning tools for a more comprehensive comparison of their usefulness for soil classification.

References:

1. Girvan MS, Bullimore J, Pretty JN, Mark-Osborn A, Ball AS. Soil type is the primary determinant of the composition of the total and active bacterial communities in arable soils. Appl Environ Microbiol, 2003: 69:1800-1809.
2. Moreno LI, Mills DK, Entry J, Sautter RT, Mathee K. Microbial metagenome profiling using amplicon length heterogeneity polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. J Forensic Sci 2006: 51: 1315-1322.
3. Yang, C. Mills D., Mathee, K. Wang, Y. Jayachandran, K. Sikaroodi, M. Gillevet, P. Entry, J. Narasimhan, G. An ecoinformatics tool for microbial community studies: Supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA, J Microbiol Methods, 2006:65:49-62.

 ** Presenting Author*

Soil Forensics, Machine Learning, Microbial Profiling

 ** Presenting Author*