A61 Multivariate Statistical Evaluation of Bacterial rRNA16S V4-V6 Sequencing to Identify Soil Evidence

James Hopkins, BA, and David R. Foran, PhD, Michigan State Univ, Forensic Science Program, 560 Baker Hall, East Lansing, MI 48824*

After attending this presentation, attendees will understand the forensic identification of soil samples based on next-generation sequencing and multivariate statistical analysis of 16S ribosomal RNA bacterial sequences.

This presentation will impact the forensic science community by introducing a novel molecular approach to soil identification. The advantages of assaying bacterial 16S rRNA loci compared to traditional physical and chemical evaluation of soil samples will be illustrated with an emphasis on the implementation of statistical methods for mathematically confident soil identifications.

Forensic soil analysis has historically been conducted through physical and chemical examination, including color determination, particle size distribution, and chemical component analysis. While these procedures have met with some success, there are a number of shortcomings, including lengthy preparation times, the amount of sample available for testing, subjectivity of results, and inconclusive data. Development of methods that reliably and statistically show soil origination is imperative.

Over the last eight years, forensic biologists at Michigan State University have been examining various molecular methods for characterizing and identifying soil samples based on their microbial populations. Early investigations using Terminal Restriction Fragment Length Polymorphism (T-RFLP) analysis showed that bacterial communities tend to differ among soils but often produce extremely complicated profiles, making reproducibility difficult. Further, these methods do not allow statistical analyses. Quantitative Polymerase Chain Reaction (qPCR) was then employed to assess if proportional comparisons of specific bacterial populations could differentiate soils. Pairwise comparisons of soil samples showed potential for differentiating them; however, not all soils could be completely separated.

With recent advances in technology, sequencing large quantities of DNA is both feasible and relatively inexpensive, making it a viable option for forensic utilization. In this study, soil samples from a wood lot, a marsh, and a yard were collected over time, and DNA was isolated. Using barcoded universal bacterial 16S rRNA primers, which allow for multiple samples to be sequenced and differentiated bioinformatically, the V4–V6 regions were amplified and sequenced using high-throughput pyrosequencing. Sequence files were processed using the software “mothur,” where barcodes and sequences too short or with ambiguous base calls were removed, and informative sequences aligned. A square phylip-formatted distance matrix was produced showing the pairwise distances between each sequence in the data set with a cutoff of distances greater than 0.30. Bray-Curtis and Sørensen indices were also calculated to exemplify the compositional dissimilarity between the samples. Additionally, the sequences were clustered into operational taxonomic units, which are bins composed of sequences 97% or more similar, based on the average neighbor method.

Nonmetric Multidimensional Scaling (NMDS) and Principal Component Analysis (PCA) were employed to evaluate the reproducibility of the analysis, and to examine which was more appropriate for the data set. NMDS is a numerical ordination technique that searches for a true best solution to the explicitly chosen axes. It has advantages over PCA in that it does not assume linear relationships within the data and is outside the restraint of the eigenvalue-eigenvector technique. However, NMDS is limited in that, being a numerical technique, it allows the possibility of different outcomes under the same conditions, and may not identify the true best solution based on computational limitations. PCA is an analytical ordination technique that identifies relationships based on variance. The first principal component, or axis, is plotted to include the greatest variance within the data set. A second principal component is then plotted perpendicular to the first along the axis of the second greatest variance. More principal components can be plotted; however, three is usually sufficient to elucidate relationships and more than three becomes difficult to visualize. PCA has advantages over NMDS in that the plot produced is easily understood and clearly illustrates relationships within the data set as well as reducing subjectivity in that axes are chosen based on the linear relationships in the data.

The ability to identify unknown soil samples via these techniques was then investigated. “Questioned” samples were picked blind from the original soil plots and processed as above. Based on both NMDS and PCA plots, the questioned samples were statistically included within the appropriate soil site. Likewise, samples 10 feet from the original collection sites were tested. The methodology proved useful for both assigning the origin of soil, and giving a percent confidence of association to that group. Plots designed by the software are easy to understand and explain in court. Furthermore, the addition of statistical confidence allows the technique to be more readily accepted as a reliable identification process, and helps meet *Daubert* considerations thus, this methodology has clear utility for forensic scientists.

Soil Identification, Multivariate Stats, DNA Pyrosequencing

Copyright 2013 by the AAFS. Unless stated otherwise, noncommercial *photocopying* of editorial published in this periodical is permitted by AAFS. Permission to reprint, publish, or otherwise reproduce such material in any form other than photocopying must be obtained by AAFS.