B99 A Hybrid Machine Learning Approach (MLA) for DNA Mixture Interpretation

Michael Marciano, MS, Forensic and National Security Sciences Institute, Syracuse University, 1-014 Center for Science and Technology, Syracuse, NY 13244-4100; and Jonathan Adelman, MS*, Syracuse University, CST 1-014, 100 College Place, Syracuse, NY 13244*

After attending this presentation, attendees will better understand machine learning approaches and the potential impact of applying machine learning to aid in the interpretation of DNA mixtures.

This presentation will impact the forensic science community by introducing the novel approach of incorporating machine learning as an analytical tool for DNA mixture deconvolution. In addition, its utility will extend past the current capillary electrophoresis-based data to utilize sequence data. This approach will produce highly informative, reliable intelligence or investigative leads from complex DNA mixtures that currently cannot be interpreted via other methods.

The challenge of DNA mixture interpretation is at the core of forensic genetic identification. Mastery of this interpretation can significantly impact the course of criminal investigations and/or quality of intelligence. In forensic settings, scientists responsible for mixture interpretations have relied on data from empirical validation studies, computation, and experience. Limitations are inherent in human-based analyses while other deficiencies are specifically related to computational capacity complexity and time constraints. There are expansive data sets that may be computationally leveraged to better address DNA mixture deconvolution; a classification approach capable of extracting maximum information from those data sets may be better able to interpret complex mixtures.

An MLA is ideally suited to such complex data sets and can indeed be used in classification problems. Hybrid systems capable of combining "human-like" subjective reasoning with the computational power of artificial intelligence techniques may be of special interest. These MLAs may provide higher-confidence and more rapid and expanded deconvolution capabilities. The power in such a system stems from: (1) the ability of the algorithm to learn from an initial data set and subsequently classify mixtures from previously unseen data; and, (2) the influence from the human analyst's experience-derived "rule sets." The approach utilizes the strengths of both computationally intensive algorithms and expert systems. Candidate features such as peak height and ratios were identified for extraction from mixture data sets, and the subsequent evaluation will be used as the feature vector used as input for the machine learning algorithms.

Several MLAs will be evaluated, with training data drawn exclusively from mixtures of known contributors and proportions. Once trained and validated, the chosen MLA will have the potential to rapidly (within minutes) deconvolute increasingly complex mixtures of at least three contributors. The architecture of the MLA permits mixture analyses using diverse data types including DNA fragment data, Polymerase Chain Reaction (PCR) amplification parameters, and a wide array of instrument parameters. The data-agnostic structure will allow increased flexibility in adapting to analyses of new data types, such as next generation DNA sequence data.

DNA Mixture, Machine Learning, Deconvolution

*Presenting Author