B138 Probabilistic Prediction of the Number of Contributors in DNA Mixtures Using a Machine Learning-Based Approach

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*

The goal of this presentation is to present the state-of-the-art approach to the prediction of the number of contributors in DNA mixtures, an area of significant need in the forensic science community.

The prediction of the number of contributors remains a significant gap in the available suite of forensic software tools. This is an approach that is believed could be seamlessly implemented into the standard workflow. This presentation will impact the forensic science community by informing attendees of possible methods to combat this issue. Attendees will be equipped with the information necessary to bring this information back to their labs to discuss the positives and negatives of this approach.

DNA mixture interpretation remains one of the most significant areas of need in the forensic community. The success of many current mixture interpretation approaches relies on the assumption that the number of contributors is correctly predicted. When this presumption is incorrect, it may lead to decreased likelihood ratios or incorrect exclusions or inclusions, particularly when faced with increasingly complex mixtures of three or more contributors. The use of machine learning to probabilistically estimate the number of contributors in a DNA mixture is proposed in this presentation. A machine learning algorithm is a statistical tool that can, after exposure to an initial set of data, be used to classify previously unseen data. Machine learning approaches excel when faced with complex problems involving implicit patterns (such as the problem of successfully classifying three- and four-person DNA mixtures). These algorithms are specifically capable of learning probabilistic models that can be subsequently used for prediction. Such a predictive model was created using a support vector machine, and the model's performance was compared against currently used methods for predicting the number of contributors.

The model had an accuracy of greater than 98% in identifying the number of contributors in a DNA mixture of up to four contributors. A comparison to approaches utilizing Maximum Allele Count (MAC) and Markov Chain Monte Carlo (MCMC) methods exhibited a greater than 6% improvement in classifying three-contributor samples and an improvement of more than 20% when assessing four-contributor samples. The Probabilistic Assessment for Contributor Estimation (PACE) also accomplishes classification of the number of contributors for mixtures of up to four contributors in less than one second using a standard laptop or desktop computer. The initial assessment of classification via machine learning relied on samples amplified using the AmpFℓSTR® Identifiler® Polymerase Chain Reaction (PCR) Amplification Kit, primarily due to the availability of large data sets. The functionality of the system has been broadened to address the new expanded-locus Combined DNA Index System (CODIS) kits, such as the PowerPlex® Fusion PCR amplification kit, with data analysis currently nearing completion.

Considering the high classification accuracy rates of PACE, the inherent complexity of standard methods to classify three or more contributors, and the lack of rapid alternatives, this approach provides a promising means of estimating the number of contributors and, subsequently, will lead to improved DNA mixture interpretation.

Number of Contributors, Mixture, Machine-Learning