B198 A Regression-Based Algorithm to Maximize the Confidence in Mass Spectral Identifications

Samantha A. Mehnert, West Virginia University, Morgantown, WV 26506; Brandon D. Lowe, St. Vincent College, Latrobe, PA 15650; Emily Ruiz, West Virginia University, Morgantown, WV 26505; J. Tyler Davidson, MS, West Virginia University, Morgantown, WV 26505; Glen P. Jackson, PhD, West Virginia University, Morgantown, WV 26506-6121*

Learning Overview: After attending this presentation, attendees will have learned the major sources of random and non-random variance in mass spectrometric analyses and the value of ion correlation analysis. Attendees will also know how to use the correlation that exists between ion abundances in replicate spectra to make compound identifications that are more confident and more accurate than existing algorithms.

Impact on the Forensic Science Community: This presentation will provide the forensic science community with a mathematical model for compound identification from mass spectrometric data that is more accurate and more precise than current static/discrete methods. The application of a more selective algorithm will decrease the incidence of false positives and further assist with the identification of unknown compounds in casework.

Hypothesis: It was hypothesized that a mathematical model that takes into account the covariance between ion abundances will provide better discrimination between true positives and true negatives than an algorithm that does not take into account the covariance between ion abundances, and for the model to work, several other hypotheses about the data must be also be true, one of which is the expectation that the residuals in the predictions should be normality distributed.

Methods/Results: Current mass spectrometric methods of substance identification use a "static" algorithm to determine the identity of a substance. The "static" approach assumes that there is one "best" or average exemplar of a substance in a library against which unknowns are compared. However, the variance in ion abundances for replicate spectra is around ±20% (95% Confidence Level [CL]), which can result in false positives in substance identifications. Instead, the algorithm uses a multivariate general regression model between ion abundances to make ion abundance predictions within a measured spectrum.

A mixture containing five drug standards was analyzed several times a day for approximately two weeks. From the approximate two dozen Gas Chromatography/Electron Ionization/Mass Spectrometry (GC/EI/MS) data files, hundreds of unaveraged mass spectra of each drug were extracted and split into a training set and an external validation set. The ion abundances were normalized to the base peak for each drug, and the 15 most abundant ions were selected iteratively to be the dependent variables within the general linear models. The model predicted ion abundances at each *m/z* value, and these predicted abundances were then compared to the measured spectra using Pearson Product-Moment Correlations (PPMCs) or Root Mean Squared Errors Of Predictions (RMSEP). The PPMCs for known positives and known negatives are then compared to a range of different threshold PPMC values to assess the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) rate at each threshold. These assessments were used to construct Receiver-Operator Characteristic (ROC) curves, which provided Areas Under the Curve (AUC) of 1, or errorless classification. A separate test was conducted to assess the number of spectra necessary to create an accurate model for a certain compound. Varying numbers of spectra were used to create general linear models, and the PPMCs and Root Mean Square Errors (RMSEs) were compared.

Whereas the AUC of ROC curves for the prediction of external validation spectra was 1 for each drug model, the PPMC threshold at which no mistakes were made varied slightly for each drug. For example, diacetylmorphine and fentanyl required PPMC thresholds between ~0.5–0.625 to distinguish TPs from TNs, but cocaine, ecgonine methyl ester, and 6-monoacetylmorphine required PPMC thresholds of ~0.88–0.95 to achieve the same errorless identifications. When changing the number of spectra used to make a Δ9-tetrahydrocannabinol model, a model constructed from as few as 30 spectra performed just as well as a model built with 200 spectra.

This project was supported by a grant awarded by the National Institute of Justice, Office of Justice Programs, United States Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the authors and do not necessarily reflect the views of the Department of Justice.

Dynamic Algorithm, Mass Spectrometry, Ion Correlations