Chemometric Techniques for the Automated Batch Processing of GC-MS Data for Chemical Identification and Supervised Model Development

Bob Schweitzer[1], Spiros Manolakos[1], Kristi Miley[2], Mary-Ruth Joyce[2], Camila Trejo Paz[2], Cristina Davis[3], Mitchell McCartney[3], Eva Borras[3], and Ashish Chaudhary[1]

[1]Detect-ION | 6812 W Linebaugh Ave, Tampa, Florida

[2]Center for Global Health and Interdisciplinary Research, University of South Florida, Tampa, Florida

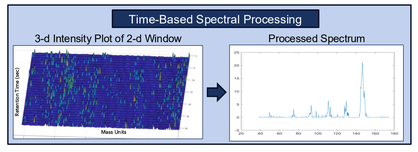
[3]Mechanical and Aerospace Engineering, University of California, Davis, California

INTRODUCTION

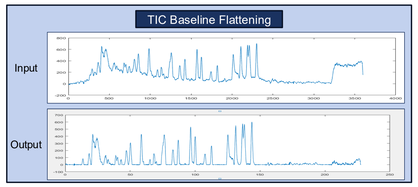
Analysis of volatile organic compounds (VOCs) in exhaled breath has emerged as a powerful tool to detect disease in early stages. Much of the research in this area, however, relies on extensive expert analysis of the collected data. There is a need for tools that support the automated analysis of large amounts of data and the development of models for disease state classification. Detect-ION has developed a set of MATLAB-based chemometric algorithms that provide autonomous processing of GC-MS data files for the identification of VOCs. This presentation demonstrates preliminary validation of these algorithms to data collected for exhaled breath analysis of subjects enrolled in a clinical trial to differentiate healthy individuals from those with specific respiratory infections.

Point-of-Care CLARION Breath Diagnostics (Top) powered by Detect-ION's chip-scale mass spectrometry (Right) enables real-time breath analysis for rapid detection of infectious diseases.

Acquired Data
- Convert sensor data to 2D raw data

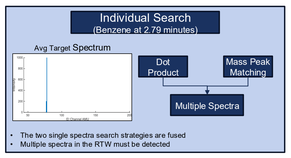


Processed Data
- Signal-to-noise enhancement
- Conversion to 1 AMU format

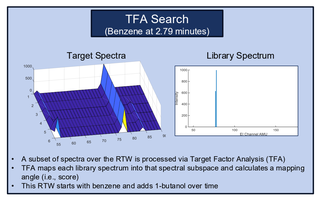


TIC
- Flatten the baseline
- Enumerate peaks in the TIC

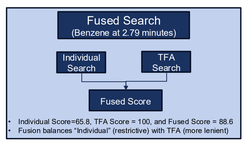
Calibration
- Perform RT calibration
- Perform quantitative calibration as needed



Spectral Library
- Develop via manual review of select GC-MS runs
- Perform library refinement over time



Chemical ID
- Loop over all VOCs in the library across RT windows
- Perform detection for each GC-MS spectrum
- Identify and output knowns and unknowns



Model Building
- Model input is the VOC ID pattern (spectra & RTs) for each data set
- Ground truth is a label associated with each pattern (ex, diseased vs healthy)
- Build a predictive model via PLS-DA (a supervised machine learning algorithm)

Clinical Data

We have established an institutional review board (IRB) and are currently enrolling asymptomatic and symptomatic human subjects in the first part (Cohort-1) of a breath collection campaign. Human subjects enrolled under this study provided breath specimens, nasal swabs, and sputum. Campaign-1 consists of 100 subjects, broken into 2 cohorts. Cohort-1 has finished enrollment of 49 subjects.

The BioFire® Pneumonia (PN) panel provided the ground truth (i.e., specific respiratory infections) for each subject. The BioFire® Pneumonia (PN) panel is based on the Polymerase Chain Reaction and is used to detect a comprehensive multiplexed array of bacteria and viruses associated with respiratory tract infections. The BioFire® Pneumonia (PN) panel tests for a total of 33 pathogens and of these, more than five subjects were found to test positive for Human rhinovirus/enterovirus, Haemophilus influenzae, and Staphylococcus aureus.

METHODS

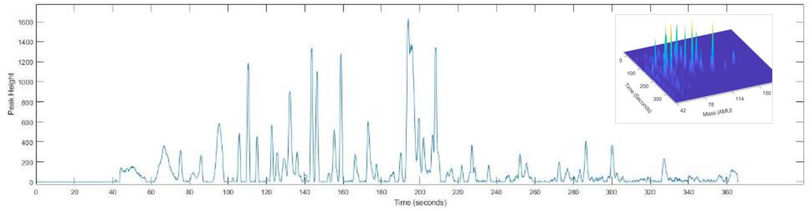
This work is funded by a three-year contract with the Defense Threat Reduction Agency (DTRA) and Defense Innovation Unit (DIU) to advance Detect-ION's Point-of-Care Breath Diagnostics prototype called "CLARION" and to apply that platform to the widespread screening and early detection of infections in warfighter populations via the identification and quantitation of key volatile organic compounds (VOCs) in breath. A manual review of a small number of data sets identifies the VOCs and their associated retention times (RTs) in each total ion chromatogram (TIC) and creates a corresponding mass spectral library. The completed library is used to batch process all data collected to create a feature vector of VOCs and associated RTs for each data set for use in model building.

Model Details

Data was collected for a novel Point-of-Care Breath Diagnostics prototype ("CLARION") developed by Detect-ION and for an Agilent commercial benchtop GC-MS. The amount of data available for PLSDA model building for both sensors is listed in the table below, where columns 3-8 are BioFire-specific.

	Total	Pos	Neg	H. Infl	Rhino	Staph	Any of 3
Agilent	49	29	20	8	9	17	26
CLARION	37	23	14	7	8	12	20

The figure below is an example GC-MS trace for exhaled human breath on the CLARION system. The inlay is a 3D plot of the chromatogram.



Typical chromatography for exhaled human breath on the CLARION system, utilizing the LTM GC prototype with a column trap as the secondary collector/injector. *Inlay:* 3D plot of chromatogram

A mass spectral library of 176 VOCs was developed via a manual review of multiple GC-MS files from the benchmark system using the NIST High Resolution Mass Spectrometry (HRMS) library. This library supported the development of a corresponding mass spectral library of 126 VOCs for the CLARION system.

The automated batch processing algorithm was run for both the Agilent and CLARION data for all data files. The resulting set of detected VOCs represents a fingerprint pattern consisting of calibrated retention times and VOC identities for each peak in each GC-MS file. These patterns along with ground truth are provided as input to a Partial Least-Squares Discriminant Analysis (PLSDA) model building tool, which is employed using leave-one-out cross-validation. A key statistic is the root mean square error of cross-validation (RMSECV).

A new descriptor (i.e., VOC) removal technique was developed and applied to the CLARION H. Influenza model. A set of 10 descriptors was removed from an initial pool of 48 descriptors using an iterative approach in which one descriptor per iteration is identified as being the most detrimental to the model building.

A given iteration involves the development of 48 models, for example, in which each model leaves out one of the descriptors. The most detrimental descriptor is the one that results in the best model as judged by RMSECV across that set of 48 models. This is because its addition results in the highest amount of degradation of the model (i.e., the model would be much better without that descriptor).

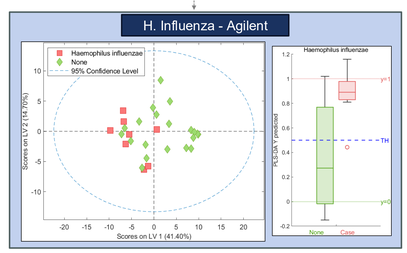
For each iteration, the previously accumulated set of descriptors is removed, and the process is repeated to find the next most detrimental descriptor to remove. Leave-one-out cross-validation is used in all model building. Detrimental descriptors represent VOCs that are confounders to the model building.

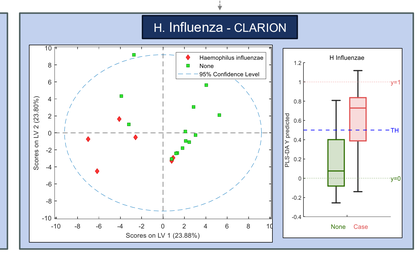
RESULTS

Models were built for each of the three primary Bio-Fire pathogens along with a fourth model for all Bio-Fire pos vs Bio-Fire neg samples for the Agilent data and the CLARION data. The following table provides the resulting statistics for these models. Figures are also provided for the Agilent and CLARION H. Influenza models.

Model	Agilent			CLARION		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
All controls vs positive cases	0.50	0.57	0.58	0.35	0.36	0.35
Human rhinovirus/enterovirus	0.67	0.53	0.67	0.50	0.71	0.64
Haemophilus influenzae	0.67	0.53	0.64	0.43	0.64	0.57
Staphylococcus aureus	0.83	0.71	0.79	0.58	0.57	0.58
Haemophilus influenzae*				0.71	0.86	0.81

*Uses the new descriptor removal technique





Left: Detection performance using HRMS-based breath analysis. Right: Detection performance using PoC CLARION Breath Diagnostics

DISCUSSIONS

- The Agilent (HRMS) models give somewhat better results on average than the Point-of-Care CLARION models. This is expected since the Agilent HRMS data is higher sensitivity GC-MS data, the models have more VOCs from breath to consider, and the optimization routine had more flexibility of being selective about removing more descriptors.
- The pathogen-specific models give better results than the "All neg vs pos controls" model. This is a very common occurrence in model-building as the general model is higher variance in VOC profiles than the pathogen-specific models.
- The new descriptor removal technique as applied to the CLARION data provides better performance for the one case in which it was applied than the Agilent model. This initial finding is promising, as it points to the potential of an inexpensive sensor to generate similar quality results as a commercial benchtop GC-MS.

UPCOMING TASKS

- Redo the CLARION analysis after making improvements to the library via more library development efforts. Inclusion of several low intensity VOCs that were excluded (due to identification challenges) may impact the overall discrimination model.
- Perform more experiments to validate the initial finding with the new descriptor removal technique. This may be interesting to evaluate both for high intensity VOC signals as well as low intensity.
- Repeat the processing and analysis with larger clinical data, as expected with the completion of ongoing Cohort-2 (Size 50; Summer 2025) and the planned Cohort-3 (Size 250; Summer/Fall 2025).

ACKNOWLEDGEMENT

This work is supported by the Defense Threat Reduction Agency (DTRA) and Defense Innovation Unit (DIU) EXHALE program via contract HQ00342390001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of DTRA or DIU. The authors declare no competing financial interest.

PoC: bob.schweitzer@detect-ion.com, ashish.chaudhary@detect-ion.com