

# Data Analysis Pipeline for Ion Mobility Spectrometry-based Proteomics

Gordon A. Anderson, Anuj R. Shah, Erin S. Baker, Nikola Tolić, Ashoka D. Polpitiya, Anoop M. Mayampurath, Brian H. Clowers, Rui Zhao, Mikhail E. Belov, and Richard D. Smith  
Pacific Northwest National Laboratory, Richland, WA



## Overview

- Multiple 10-h LC-IMS-MS analyses performed on tryptic digests of *Shewanella oneidensis* MR-1, mouse plasma, and human plasma to characterize cross sections of peptides in a large dataset
- Overlapping peptide identifications from individual organism are used in a machine learning framework to predict IMS drift times
- High accuracy prediction of drift times for +2 charge states (95% accuracy with FWHM of 1 millisecond) and +3 charge states (90% accuracy with FWHM of 2 milliseconds)

## Introduction

Ion mobility spectrometry (IMS) coupled with liquid chromatography (LC) and mass spectrometry (MS) offers an additional dimension of separation that can be used to distinguish gas phase conformers, separate similar ions based , and reduce chemical noise (Figure 1).

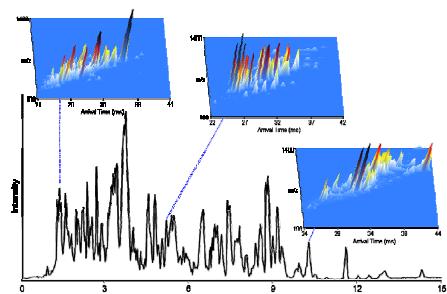
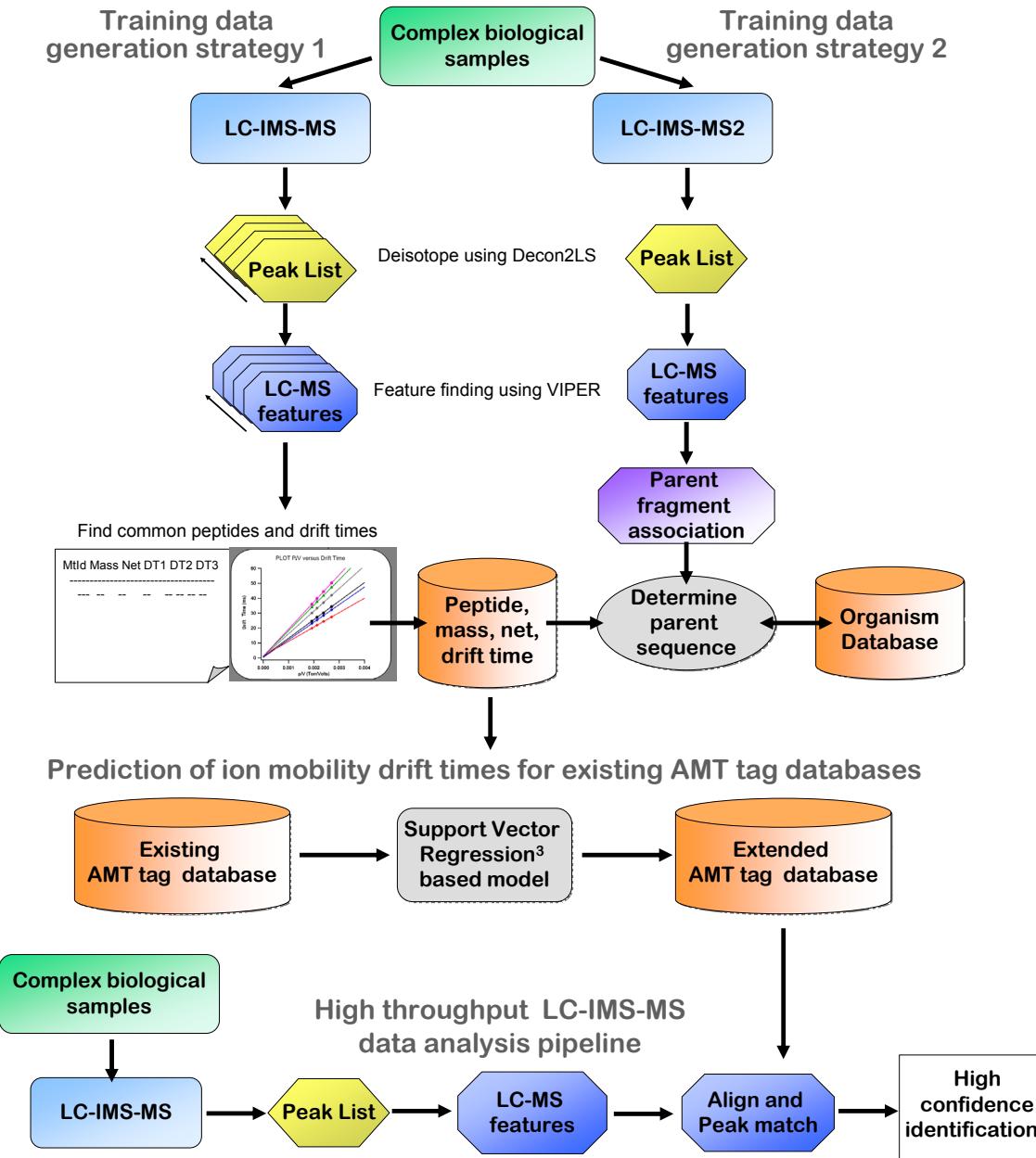


Figure 1: 15-min LC-MS base peak chromatogram with corresponding IMS-MS spectra at 3 different elution times

## Methods



## Results

### Improved false discovery rates

Table 1: Profiles of four different algorithms for characterizing drift times. Tolerances for mass, normalized elution time and drift time are 9 ppm, 0.01, and 2 msec, respectively

Drift time profile	Average	Weighted Average	Max Peak	Conformer Detection
Extended AMT Database size <sup>1</sup>	1864	2776	3662	3699 (3000 unique)
LC-MS identifications <sup>2</sup>	1429	2211	2913	2287
False identifications <sup>3</sup>	113	174	208	98
False discovery rate <sup>4</sup>	7.9%	7.9%	7.0%	4.3%
LC-MS-DT Identifications <sup>5</sup>	1119	1884	2597	2076
False identifications	62	100	102	59
False discovery rate	5.5%	5.3%	3.9%	2.8%
% Reduction in FDR	29.9%	32.6%	44.0%	33.8%

<sup>1</sup>The Extended Accurate Mass and Time (xAMT) tag database is built by selecting highly confident overlapping peptides identifications from four experiments at different pressure and voltage values for *Shewanella oneidensis* MR-1.

<sup>2</sup>Unique identifications as obtained by the AMT tag approach using only LC and MS dimensions

<sup>3</sup>False identifications are calculated by shifting the masses of peptides within the mass tag database by 11-Daltons

<sup>4</sup>False discovery rate were calculated by shifting the masses of peptides by 11-Da

<sup>5</sup>Unique identifications as obtained by the AMT tag approach with the additional dimension of IMS drift times

### Drift time prediction

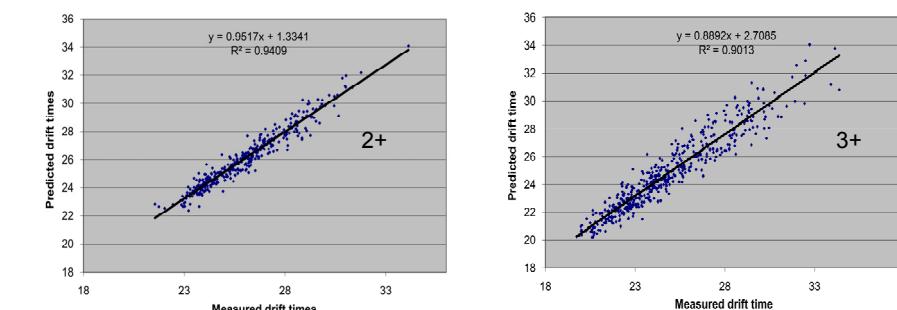


Figure 2: 5 fold cross validation results for prediction of drift times for charge states +2 and +3

- A Support Vector Regression<sup>3</sup> (SVR)-based model was trained and tested on the xAMT tag database.
- 132 features based on properties calculated directly from peptide sequence were used to predict drift times.
- The prediction method illustrated 95% accuracy for predicting 2+ charge state peptides and 90% accuracy for 3+ charge state (Figure 2).
- The final goal is to predict drift times for existing AMT tag databases to resolve ambiguities in peak matching and improve FDRs (Table 1).

## Conclusions

- A support vector regression-based model was developed to predict the drift times directly from peptide sequences. Our model exhibits very high accuracy on the xAMT database and will be tested in the future on larger AMT tag<sup>1</sup> databases.
- Preliminary results indicate the potential of predicting the drift times for existing AMT tag databases and using the additional dimension of drift times to reduce FDRs.
- A pipeline of processing steps is outlined for high throughput analysis of IMS coupled proteomics datasets.

## Acknowledgements

This work was funded by The Next Generation Proteomics Measurement Platform Initiative at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. Samples were analyzed using capabilities developed under the support of the NIH National Center for Research Resources (RR18522) and the U.S. Department of Energy Biological and Environmental Research (DOE/BER).

Significant portions of the work were performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

## References

- Zimmer JS, Monroe ME, Qian WJ, Smith RD. Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev* 2006; 25:450-482.
- Paša-Tolić L, Masselon C, Barry RC, Shen Y, Smith RD. Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques* 2004; 37:621-624, 626-33, 636.
- Vapnik VN. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, 1995.

**CONTACT:** Gordon A. Anderson  
Biological Sciences Division, K8-98  
Pacific Northwest National Laboratory  
P.O. Box 999, Richland, WA 99352  
E-mail: gordon.anderson@pnl.gov