

# A Diverse Dataset to Assist in the Production of Improved Peptide and Protein Identification Software

John E. Klimek<sup>1</sup>; Laura J. Hohmann<sup>1</sup>; Jimmy K. Eng<sup>1</sup>; Jennifer Jackson<sup>1</sup>; Andrew B. Gemmill<sup>1</sup>; Daniel B. Martin<sup>1</sup>  
<sup>1</sup>Institute for Systems Biology, Seattle, WA

## Overview

- MS/MS spectra are often used in Proteomics to identify peptides and proteins.
- Programs such as Sequest and Mascot assign peptides to spectra.
- Most of the software used to determine probabilities for spectra based on Sequest/Mascot output have used data from a single source, a 3D ion trap, in their testing and validation.
- To address this we have created a public dataset of LC-MS/MS runs from various instruments for use as an aid in the evaluation and testing of new software.
- The standard peptide mixture used in the creation of this dataset is made from a trypsin digest of 18 commercially available proteins.
- Three replicate experiments were performed with each digest run ten times on a subset of eight instrument platforms.
- To date approximately 750,000 MS/MS spectra have been collected from 8 different mass spectrometers.
- This data is being analyzed for quality, and will be made available to the public shortly.
- This dataset will assist in software development for multiple platforms.

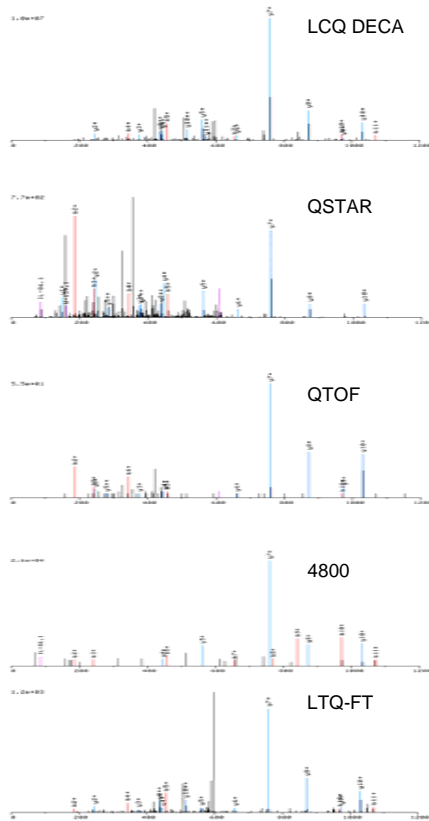
## Introduction/Background

- Mass Spectrometry based Proteomics uses software such as Sequest and Mascot to assign peptide ID's to MS/MS spectra
- Other software such as Peptide Prophet and Protein Prophet assign a probability to correct assignments
- Much of the software has been optimized for performance on a single instrument type, a 3D ion trap
- Intensity of MS/MS peaks, neutral loss, immonium ion production, etc varies greatly between and within different types of mass spectrometers
- Any next-generation software looking at such values will need to work well on various platforms
- We have created a public dataset of LC-MS/MS runs from various instruments, using a tryptic digest of a standard peptide mixture, for use as an aid in the evaluation and testing of new software

## Methods

- Standard peptide mixture made from 18 commercially available purified proteins
- 3 different batches used for experiment
- Starting concentration is 1 $\mu$ M of each protein
- Mixture reduced, alkylated, and digested with trypsin
- Peptides cleaned up for mass spectrometry using a waters MCX column
- Sample aliquoted to contain approx 500 fmol. of each protein per run
- Aliquots run 10 times on a subset of the following instruments
  - ThermoFinnigan LTQ
  - ThermoFinnigan LCQ DECA
  - Waters/Micromass QTOF Ultima
  - Applied Biosystems API QSTAR Pulsar
  - ThermoFinnigan LTQ-FT
  - Applied Biosystems 4800 TOF-TOF
  - Agilent Chip-XCT Ultra ion trap
  - Applied Biosystems 4700 TOF-TOF
- All samples run with a gradient of 10% to 35% ACN over the course of 60 minutes

## Sample MS/MS Spectra



This figure illustrates some of the differences in MS/MS spectra between different platforms. All spectra high confidence identifications are from MS/MS scans of the peptide ALGNTNPTNAEVK from the protein spj|P02602|MLE1\_RABIT MYOSIN LIGHT CHAIN 1

## Proteins Used

AMYLASE ALPHA (A4551)  
ACTIN GAMMA (A3653)  
CYTOCHROME C (C2037)  
CATALASE (C40)  
GLYCOGEN PHOSPHORYLASE (P6635)  
ALKALINE PHOSPHATASE PRECURSOR (79377)  
ALPHA-LACTALBUMIN PRECURSOR (L6010)  
BETA-GALACTOSIDASE (G5635)  
CARBONIC ANHYDRASE II (C2522)  
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE (G2267)  
MANNOSE-6-PHOSPHATE ISOMERASE (P2621)  
OVALBUMIN (A2512)  
MYOGLOBIN (M0630)  
MYOSIN LIGHT CHAIN 1 (M9891)  
REGULATORY LIGHT CHAIN 2 (M9891)  
BETA CASEIN PRECURSOR (C6905)  
BETA-LACTOGLOBULIN PRECURSOR (L0130)  
SERUM ALBUMIN PRECURSOR (A3059)  
SEROTRANSFERRIN PRECURSOR (T0178)

These are the proteins used in the construction of the database all runs were searched against. All proteins were purchased from Sigma-Aldrich. Numbers in parenthesis are Sigma product numbers.

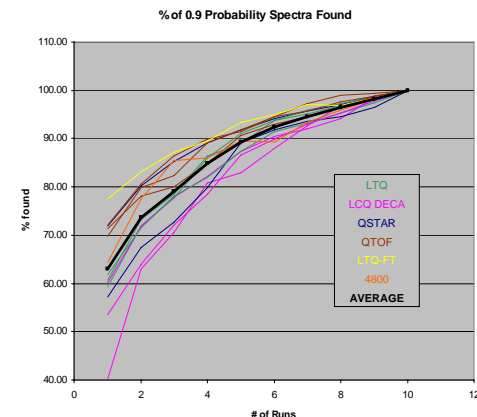
## Data Summary

All values averaged over 10 runs

Instrument	Spectra	Unique Peptides 0.9 Prob.	Std Dev. of Unique Peptides	Range of Unique Peptides
LTQ mix 1	8658.0	627.9	32.9	565 to 686
LCQ DECA mix 1 (30min)*	3122.9	282.9	40.4	218 to 326
QSTAR mix 1	3084.0	593.0	25.2	524 to 611
LTQ-FT mix 1	16098.0	483.9	16.4	449 to 500
LTQ mix 2	8329.3	874.2	48.0	813 to 965
LCQ DECA mix 2	6069.6	670.2	14.5	653 to 707
QSTAR mix 2	2678.0	432.0	29.9	394 to 489
4800 mix 2	2600.8	601.5	115.6	397 to 701
QTOF mix 2	2844.5	484.8	22.1	452 to 522
LTQ mix 3	6298.8	539.3	11.4	519 to 571
LCQ DECA mix 3	4349.0	334.6	19.2	307 to 378
QTOF mix 3	2383.1	210.3	13.2	180 to 234

Above are the results for complete datasets. Incomplete datasets exist for the Applied Biosystems API QSTAR Pulsar mix 3, Agilent Chip-XCT Ultra ion trap mix 2, and Applied Biosystems 4700 TOF-TOF mix 3. These datasets will be added to the public database upon completion. \*Mix 1 was done on the LCQ DECA as a pilot study, with a gradient from 10% -35% ACN over 30 minutes, instead of the standard 60 minutes. The sample was exhausted before a 60 minute gradient series could be run.

## Runs to Saturation



This graph illustrates the overlap between each run of the standard mixture on an instrument. For all platforms, over 80% of unique peptides found over the course of the 10 runs are found in the first 5 runs. This degree of overlap allows for multiple subgroups to be made on the same instrument's data and compared; for example a training and validation group.

## Summary

- We have collected 750,000 MS/MS spectra from the 10 replicates of the three sample mixes run on a subset of 8 different mass spectrometers.
- This dataset has been designed as a resource to facilitate the development of new peptide and protein identification software.
- It will serve as a resource for software developers for applications across multiple platforms.
- The data is being analyzed for quality and will be made available to the public shortly.

## References/Contact Info.

Keller, A. et al. **Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search**, Anal. Chem. 2002, 74, 5383-5392  
Nesvizhskii, A. et al. **A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry**, Anal. Chem., 75, 4646-4658  
Keller, A. et al. **Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis**, OMICS 2002, 6, 207-212

If you have are interested in communicating with the authors about this presentation please contact Dan Martin, dmartin@systemsbiology.org; or John Klimek, jklimek@systemsbiology.org.