

# msInspect: An Informatics Tool for Integrated Analysis of LC-MS and LC-MS/MS Data Generated from Complex Protein Mixtures

Matthew Bellew<sup>1,2</sup>, Marc Coram<sup>1</sup>, Matthew Fitzgibbon<sup>1</sup>, Mark Igra<sup>1,2</sup>, Damon May<sup>1</sup>, Pei Wang<sup>1</sup>, Jimmy Eng<sup>1</sup>, Ruihua Fang<sup>1</sup>, Jeffrey Whiteaker<sup>1</sup>, Heidi Zhang<sup>1</sup>, Amanda Paulovich<sup>1</sup>, and Martin McIntosh<sup>1</sup>  
<sup>1</sup>Fred Hutchinson Cancer Research Center, Seattle WA <sup>2</sup>LabKey Software, Seattle WA

## OVERVIEW

Successful application of differential proteomics by liquid chromatography mass spectrometry (LC-MS) requires extraction of peptide features, estimation of peptide abundances, relative quantitation, alignment & normalization across multiple related runs, and identification of features. We have developed an open-source software application called **msInspect** that integrates novel algorithms for each step in this process.

msInspect has been designed to look for peptide signatures, rather than isolated peaks, in LC-MS data. This approach allows us to implement *feature-based* alignment algorithms and to apply techniques typically reserved for LC-MS/MS, such as relative quantitation with stable isotope labeling, directly to LC-MS data.

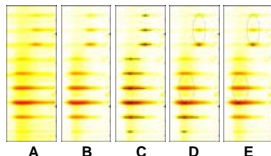
msInspect may be used from the command line for batch processing of very large data sets. It may also be invoked with a graphical interface that includes a number of visualization and analysis tools. New implementations of components, such as feature extraction, may be added to msInspect at run time, making it an excellent platform for the development and integration of new algorithms.

msInspect shares significant code components with CPAS [Rauch 2006], our data management and mining system which includes support for storing and analyzing LC-MS/MS sequencing results. We are in the process of integrating msInspect with CPAS, and we are applying both tools to large-scale biomarker discovery efforts.

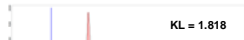
## PEPTIDE EXTRACTION

Recent advances in high mass-accuracy instrumentation have made it practical to identify the distinctive isotopic signatures of peptides in LC-MS data. An outline of our algorithm is:

- Spectra are loaded from an mzXML format file [Pedioli 2004].
- Local background is estimated and removed.
- Local maxima are located via wavelet decomposition, and those that persist over time are retained as peaks.
- Co-eluting peaks are grouped and compared to a model of the isotopic distribution associated with a peptide of a given mass. Groups of peaks that best match this model are added to our peptide list.
- "Stray" peaks that are not assigned to any group may be removed.



We use a Poisson model for the expected isotopic distribution of a peptide of a given mass. The Poisson rate is based on theoretical distributions computed from 539,957 peptides from a virtual tryptic digest of a recent release of the Human IPI database.



KL = 1.818

Successive panels show in blue the Poisson model constructed for different groups of peaks. In the top panel, the left-most peak is chosen as the monoisotopic peak. In the next panel the second peak is chosen, and finally the third is chosen.



KL = 0.062

For each combination, a metric (KL) is computed to describe how much the extracted peaks deviate from the model. A lower KL indicates a better match, and the group in the middle panel clearly provides the best fit.

In general, more than one charge state is possible and Poisson models for corresponding masses are considered.

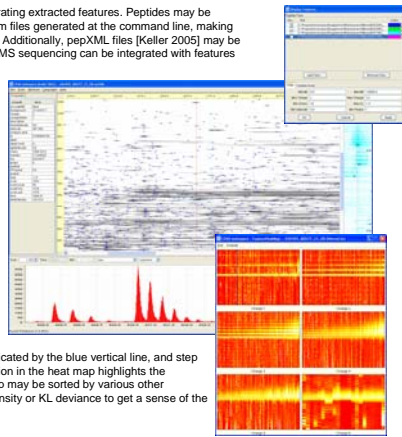
## VISUALIZATION AND CURATION

msInspect includes tools for visualizing, filtering, and curating extracted features. Peptides may be extracted directly in the graphical interface or loaded from files generated at the command line, making msInspect well suited for high-throughput environments. Additionally, pepXML files [Keller 2005] may be loaded so that peptides confidently identified by LC-MS/MS sequencing can be integrated with features extracted from LC-MS alone.

Features extracted from LC-MS may be filtered by intensity, charge, deviance from model isotopic distribution, number of isotopic peaks, number of scans, and several other measures. Exclusion of low quality features is critical for the success of relative quantitation and alignment of multiple runs.

A "heat map" visualization tool allows rapid exploration of a feature set to evaluate filtering parameters or to exclude specific, problematic features. In the image at right, a window around each identified feature has been extracted and mapped to a heat map color scale; red is low intensity and yellow high. Each panel contains features from one charge state. The y-axis is m/z, with the monoisotopic peak centered in each panel, and features are sorted left-to-right by increasing mass. The effect of mass accuracy is clearly visible.

The user may select any feature in the heat map, as indicated by the blue vertical line, and step forwards and backwards through the sort order. Navigation in the heat map highlights the corresponding feature in the main window. The heat map may be sorted by various other measures. It is particularly useful to sort features by intensity or KL deviance to get a sense of the appropriate thresholds for a given data set.



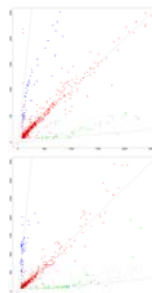
## RELATIVE QUANTITATION

Stable isotope labeling provides a means of performing relative quantitation. In a common realization, two samples are mixed after being labeled with a light reagent and a heavy reagent. LC-MS spectra from the mixture will contain paired features separated by a multiple of the difference between the heavy and light labels.

After extracting peptide features, msInspect can perform relative quantitation within a run by searching for label pairs. While some labeling methods attach a single label to each peptide, other methods label each occurrence of a specific amino acid. Because the peptide sequence is not necessarily available, msInspect searches a defined number of integer multiples of the mass difference in each direction from a given feature. The label weights, the maximum number of labels to consider, and the scan and mass tolerance for forming pairs are all configurable by the user.

An example from a cleavable ICAT experiment is shown at left. The upper feature is 9.008 Da heavier than the lower, and both reach maximum intensity at nearly the same time. We infer a pair of peptides, each with one labeled cysteine, and indicate this by connecting the pair with a line.

The scatter plots at right compare light intensities to corresponding heavy intensities for cleavable ICAT (top) and N-terminal labeling with deuterated succinic anhydride (bottom). Red points are from a 1:1 mix, blue from a 10:1 mix, and green from a 1:10 mix.

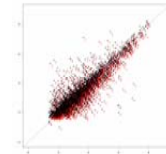


## ALIGNMENT AND NORMALIZATION

To support label-free quantitation, msInspect includes the ability to align multiple LC-MS runs and combine them into a "peptide array." The included algorithm operates on extracted peptide features rather than aligning total ion chromatograms or isolated peaks.

For a group of ten runs of unfractionated human serum, the upper plot at right illustrates the non-linear mapping constructed to transform each run onto the first, which is used as a reference. The horizontal axis indicates post-transformation scan number and the vertical axis indicates how much each scan was shifted to bring it into registration with the reference run.

The lower plot at right shows those features that aligned across all ten runs before (left) and after (right) registration. The vertical axis is mass, and the horizontal axis is scan number. Points are colored by charge and have size proportional to feature intensity.



Results of alignment can be saved as a "peptide array" with each row representing a peptide and each column the intensity observed in a run. Such data can be processed by a variety of tools generally applied to transcript microarray data.

Moment-based normalization methods, though often appropriate for microarray data, may introduce biases when applied to peptide arrays since low abundance peptides may be too faint to be detected by the instrument or by feature extraction algorithms. We have implemented a global normalization method [Wang 2006] that compensates for low-intensity missing data by considering the top ordered statistics when computing a re-scaling coefficient. The scatter plot at left compares the log intensities of one pair of runs from the above alignment before (red) and after (black) normalization.

## AVAILABILITY

msInspect is available from our website <http://proteomics.fhcr.org/> under an Apache 2.0 license. The software may be downloaded as an executable Java JAR file, or may be launched via Java Web Start directly from a web browser. The software is in regular use under Windows, GNU/Linux, and Mac OS X.

## ACKNOWLEDGEMENTS

This work was funded by National Cancer Institute contract 23XS144A.

## REFERENCES

- Rauch A, et al., Computational Proteomics Analysis System (CPAS): An Extensible, Open-source Analytical System for Evaluating and Publishing Proteomic Data and High throughput Biological Experiments. J Proteome Res 2006;5(1):112-121.
- Padioli PG, et al., A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004;22(11):1459-66.
- Keller A, et al., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Molecular Systems Biology, doi:10.1038/msb-4100024, Published online: 2 August 2005.
- Wang P, et al., Normalization regarding non-random missing values in high-throughput mass spectrometry data. Proceedings of the Pacific Symposium on Biocomputing 2006; 11:315-325.
- Bellew M, et al., A suite of algorithms for comprehensive analysis of complex protein mixtures using high-resolution LC-MS. In preparation.