Using a Spike-In Experiment to Evaluate Analysis of LC-MS Data

Leepika Tuli, Tsung-Heng Tsai, Rency S. Varghese, Amrita Cheema, Habtom W. Ressom*

Lombardi Comprehensive Cancer Center, Georgetown University, Washington DC, USA

**hwr@georgetown.edu*

Abstract

Recent advances in liquid chromatography-mass spectrometry (LC-MS) technology have led to newer approaches for measuring changes in peptide/protein abundances. Label-free LC-MS methods have been used for extraction of quantitative information and for detection of differentially abundant peptides/proteins. However, difference detection by analysis of data derived from label-free LC-MS methods requires various preprocessing steps including alignment, baseline correction, filtering, noise reduction, peak detection, normalization, and peak quantification. Although several specialized tools have been developed to analyze LC-MS data, determining the most appropriate computational pipeline remains challenging partly due to lack of established gold standards. In this paper, we use a spike-in experiment to evaluate the performance of three software tools in accurately detecting changes in peptide abundances from LC-MS data obtained by a label-free LC-MS method. We observe that tools that incorporate peptide isotope cluster and multiple charge information lead to more difference detection with fewer false positives.

1. Introduction

Liquid chromatography-mass spectrometry (LC-MS) has been an indispensable tool for profiling peptide/protein expression, mainly because it is a sensitive technique that can detect minute differences in peptide/protein levels [1-4]. The changes in peptide/protein abundances are important for biomarker discovery research as they are generated in response to a perturbation, disease, morphogenesis, toxicity, or any other kind of cell stress, in a given system. A typical LC-MS proteomics run separates peptides/proteins on a chromatographic column to elute them later, for which mass spectra are subsequently acquired using a mass spectrometer. Spectral information is acquired at a specific scan rate (every few seconds). Each chromatographic time point records a series of mass signals. These mass signals (m/z values) at different chromatographic time points are then formatted as a matrix that describes the LC-MS data.

Difference detection between LC-MS runs is achieved by either using isotope-labeled internal standards such as iTRAQ, ICAT, O^{18} and N^{15} labels [5-7] or a label-free approach. Although labeling methods have been reliably used for obtaining both relative and absolute quantification, certain limitations exist including limited number of samples that can be analyzed, occurrence of artifacts from labeling, limited measurements, limited availability of isotope references, extensive procedural steps, and lack of tools to analyze data from different platforms [8-9]. As a result, a growing interest is seen in the direction of label-free methods for biomarker discovery studies.

Label-free methods offer an alternative strategy for measuring differential protein expression as they do not require mixing of samples and yield a wider dynamic range. These methods are more attractive as they are cost effective and involve simpler protocols.

Several label-free difference detection methods exist including spectral count [10], sequence coverage [11], peptide count [12], and precursor signal intensity [13-17]. The first three methods provide relative abundance information based on MS/MS fragmentation. However, these methods tend to discard low abundant ions that are not typically selected for MS/MS fragmentation. Alternatively, a direct comparison method, where all ions at survey scan (MS1) level are analyzed, is used to capture low abundant ions. However, this approach requires that preprocessing steps such as alignment, peak detection, and normalization are appropriately handled.

Various software tools exist for LC-MS data analysis including MZmine 2 [18], XCMS [19] and Progenesis LC-MS (NonLinear Dynamics, United Kingdom). With the increase in demand for label-free quantification tools, a suite of software packages have also been developed such as OpenMS, msInspect,

SpecArray, SuperHirn, Xalign, and MSView [20].

Although many tools are available for LC-MS data preprocessing, they either have been optimized for certain platforms or have their own computational requirements, or challenges on how the data should be processed. Each tool implements a different set of algorithms in its workflow characterized by its set of strengths and limitations. While some workflows perform well on data preprocessing steps, others focus primarily on statistical and machine learning methods for difference detection or visualization. The spike-in experiment presented in this paper allows us to evaluate such workflows. Fig. 1 illustrates the experimental design that involves two groups of samples: (i) serum samples with spike-in peptides, and (ii) serum samples alone. We generated LC-MS data using the Qstar Elite mass spectrometer equipped with nano-ionization source, connected to a nano-Acquity UPLC system. We use our LC-MS data to evaluate workflows implemented in three software tools (Progenesis LC-MS, XCMS, and MZmine 2). All of them provide modular workflows which enables us to assess each pipeline explicitly.

2. Methods

2.1. Serum

Blood was collected from five healthy individuals in a sterile tube and allowed to clot for 30 minutes. Serum (clear liquid) was collected in a sterile tube after centrifugation at 1000 rpm for 15 minutes. Serum was aliquot and stored at -80°C. Serum was allowed to thaw at room temperature prior to analysis.

2.2. Protein depletion and digestion

Abundant proteins such as albumin obscure separation, detection and quantification of low abundant proteins in serum. It is common to deplete serum of high abundant proteins such that low abundant proteins can be easily detected. For our study, we depleted 60 µl of human serum of IgG and Albumin using Aurum serum protein mini columns as described in BioRad protocol (http://www3.bio-rad.com/LifeScience/pdf/Bulletin_2823.pdf). Depleted serum was collected as the *unbound* eluate by centrifuging the column at 10,000g. Protein concentration was determined by Bradford assay. After acetone precipitation, 100 µg of depleted serum was set for trypsin digestion. Samples were reduced with 50 mM tris (2-carboxyethyl) phosphine (TCEP) at 60°C for 60 minutes. After cooling to room temperature, the sample was alkylated using 0.2 mM methyl methanethiosulfonate (MMTS) at 37°C for 30 minutes. Trypsin was added (protein:enzyme ratio of 50:1), followed by incubation at 37°C for 3 hrs. Another dose of trypsin was added after 3 hrs and the mixture incubated at 37°C overnight. The digested samples were cleaned and speed vac dried and reconstituted in mobile phase A solvent (0.1% formic acid in H_2O, 2% ACN) prior to injection.

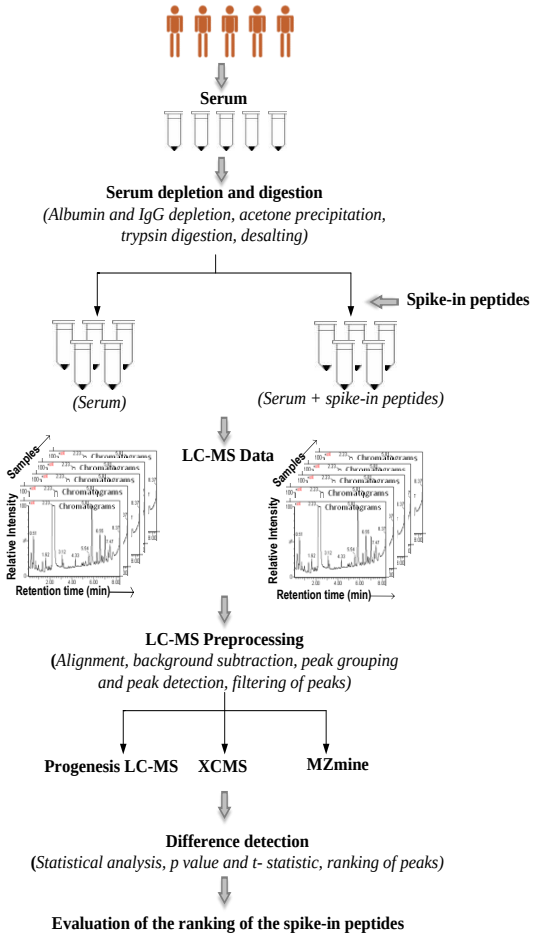

Fig. 1. Spike-in experiment to evaluate analysis of LC-MS data generated by a label-free LC-MS method.

2.3. LC-MS data acquisition

Two datasets were generated from five serum samples obtained from five healthy individuals. The first dataset was derived from the five serum samples mixed with known concentration of spike-in peptides. The second dataset was obtained from the five serum samples alone. In the first dataset, nine MassPrep peptides (designated as Peptides 1-9 in Table 1) were added prior to acquisition of LC-MS data (Fig. 1). The concentration of the MassPrep peptides (1 pmol/µl) was selected based on best detection of the lowest concentration of MassPrep peptides available on the Qstar instrument (data not shown). Both datasets were acquired by LC-MS/MS on the hybrid Q-TOF mass spectrometer using the same acquisition parameters.

Mass spectrometer was calibrated and nano-UPLC system equilibrated for optimal separation. Chromatographic separation was performed on a Waters Acquity BEH C_{18} column (75 µm X 150 mm, 1.7 µm particle size) equilibrated with 98% mobile phase A (0.1% formic acid in H_2O, 2% ACN) and 2% mobile phase B (0.1% formic acid in acetonitrile, 2% H_2O). The protein digest was preloaded on a nanoACQUITY UPLC Symmetry C18 trap column (180 µm X 20 mm) before separation. Both column and the autosampler were maintained at a temperature of 40°C and 4°C respectively.

A 180 minute linear gradient elution was performed as follows: 99% mobile phase A for the first 20 minutes, changing to 20% A by 200 minutes, 15 minutes with 1% A and finally back to 99% A for 25 minutes at 300nL per minute flow rate. The nano-Acquity UPLC system (Water, USA) was connected to a hybrid QStar-Elite MS equipped with a nano-electrospray ionization source (Waters Micromass, USA) for ESI-QTOF analysis of serum samples. The automated system ensured reproducible loading of samples using LC-autosampler. MS settings were as follows: Ionspray voltage (IS) 2300, Interface heater temperature (IHT) 160°C for the positive ionization mode, with a mass error tolerance set to100mDa. Analyst (v 2.0) was used for data acquisition and LC-MS operations.

A full mass scan (MS1) was obtained between ~9,000-12,000 resolution. The total scan cycle was 7 seconds long which included one MS scan from 350-2000 m/z and MS/MS scans for the five most intense product ions ranging from 50 to 2000 m/z using 30 seconds dynamic exclusion and rolling collision energy for fragmentation. The dynamic exclusion feature enables exclusion of ions that have been selected in the previous scan for MS/MS from being selected again for that period of time. Each sample was followed with four wash cycles to minimize sample carryover.

A 100 fmol *E.coli* beta-galactosidase digest was run at the beginning and end of a batch to monitor LC-MS profiles. It was observed that due to low intensity of spike-in peptides when mixed in serum, no MS/MS information would be available. Therefore, all samples were acquired using an inclusion list for spike-in peptides. For each set, an inclusion list was prepared based on one injection of Massprep peptides typically run at the beginning of a batch, acquired under information dependent acquisition (IDA) mode. Only peptides identified with a >10% confidence interval were retained on the list. This file was imported into the Inclusion tab of the IDA method in Analyst software. Both a time filter of ±1 minute and the predicted ion m/z value was entered in the inclusion list.

The data were searched against the International Protein Index (IPI, human v 3.6) database using Paragon Algorithm in ProteinPilot^{TM} software (v 3.0). Fixed modification was set to carboxymethyl (C) and variable modification to oxidation of methionine (M). Trypsin digestion with cysteine alkylation and ESI-Q-TOF instrument was chosen for the searched. An unused score of 1.32 (equivalent to 95% confidence) was used on Paragon for protein identification. For Progenesis LC-MS software, MASCOT search was done against a custom MassPrep database (MPDB), containing MassPrep peptides and human protein sequences with MassPrep sequence similarity. Both groups: with and without spike-in peptides were searched against MPDB. Compiling a custom MPDB was necessary as some MassPrep peptides were not present in human database and it reduced our database search time considerably.

2.4. Analytical tools

We evaluated three software tools: Progenesis LC-MS (NonLinear), XCMS and MZmine 2. All tools perform the common steps including peak detection, peak grouping, alignment, and difference detection. In addition, Progenesis LC-MS allows peptide identification by analysis of tandem MS (MS/MS) data. Both Progenesis LC-MS and MZmine 2 provide solutions to identify charge states and isotopic patterns. We observed that the identification of these patterns greatly reduces ambiguous comparisons and improves the performance of the overall workflow.

Progenesis LC-MS generates an aggregate run containing the peptide ions from all analyzed runs. Semi-supervised alignment is achieved by choosing a reference run and aligning all runs to it based on landmark vectors (selected manually or automatically). The isotopic patterns are then identified. Irrelevant features can be further filtered out by selecting range of interest and charge state. For our analysis, since most of the peptides eluted before 150 minutes, the LC-MS map is filtered to retain features with only +2, +3 and +4 charges and within 150 minutes of gradient. Once filtered the data are normalized by calculating quantitative abundance ratio between the run being normalized and the reference run. Normalization is based on the assumption that most proteins do not change, and so the quantitative abundance ratio should be close to 1. This is followed by aggregating runs for difference detection and statistical analysis.

XCMS provides a series of LC-MS data preprocessing algorithms implemented in R language and is helpful to assess performance at each preprocessing step. The peak detection is performed on overlaid extracted ion chromatograms by a pattern matching approach, with kernel of Gaussian or second derivative of Gaussian. After peak detection, peaks that fall within a pre-defined range are grouped together

and better grouped peaks are identified as "well-behaved" peaks for alignment purpose. Alignment is preceded by applying medium filter and local regressor. Since XCMS was developed for metabolomics study, several issues related to proteomics study such as charge state selection and isotopic pattern identification are not addressed.

A similar workflow composed of peak detection, peak list deisotoping, alignment, peak row filtering and gap filling is carried out in MZmine 2. The peak detection is performed on each scan and only peaks which span a certain range are retained. The main difference between this workflow and XCMS is the deisotoping step which significantly eliminates irrelevant features.

3. Results and Discussion

The raw LC-MS data were exported as wiff files and converted to mzXML format using mzWiff v.4.2.1 before importing them into the software tools.

Analysis of the two groups of LC-MS datasets by Progenesis LC-MS detected 1,317 features. We used Progenesis LC-MS to identify features that are differentially abundant between the two groups of samples (serum samples mixed with spike-in peptides and serum samples alone). After applying correction for multiple testing, six features were found to be significantly different between the two groups (q-value <0.05). In addition to the workflow for difference detection, Progenesis LC-MS allows peptide identification by analysis of MS/MS data. We used the selected six features for peptide identification. Five features were identified as MassPrep peptides whereas one feature was identified as EVKSADR from kininogen protein (P08934). This was the only false positive detected by Progenesis LC-MS. Three other features that matched m/z and retention time values for the spike-in Peptides 1, 6 and 7 but were not detected in the first run of analysis. Manual addition to analysis resulted in correct identification of these peaks as RGDSPASSKP, DRVYIHPFHLLVYS, and WLTGPQLADLYHSLMK with a high MASCOT score. Figure 2 depicts the three-dimensional LC-MS maps zoomed into each of the eight features whose m/z and retention time values match with those of the MassPrep peptides. All features were correctly identified as MassPrep peptides with a high MASCOT score, except Peptides 8 and 9 for which no matching features were found.

For XCMS, the challenge was to correctly assign features across the multiple runs. By examining various peak profiles, we found the following parameters that worked best: fwhm = 30, snthresh = 40 for peak detection, and bw = 60, mzwid = 1.2 and bw = 40, mzwid = 0.5 for peak grouping in between retention time correction. With these parameters, 2,751 features were detected that were analyzed for difference detection. Out of the 2,751 features, 1,642 were significantly different between the two groups (q-value < 0.05). Although most of the 1,642 had a fold change < 2, a significant number of features (318) had a fold change > 2 and 106 features yielded a fold change > 2.

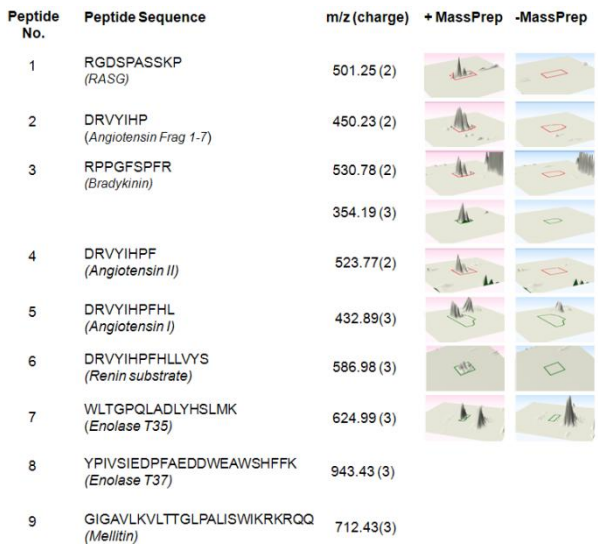

Fig. 2. Three dimensional maps of MassPrep peptides analyzed by Progenesis LC-MS. Comparison between serum with spike-in peptides (+MassPrep) and serum alone (-MassPrep) groups.

For MZmine 2, a span range of m/z = 0.05 Da, retention time = 1.5 minute, and presence across a minimum of five runs were selected for peak detection and filtering purposes. With these parameters, MZmine 2 detected 1,045 features were detected. Of these, 151 were significantly different between the two groups (q-value < 0.05). While 86 features had a fold change > 2, only 39 features yielded a fold change > 5.

Table 1 presents the ranks, p-values, q-values, and fold change for the spike-in peptides, obtained by comparing the LC-MS data from the two groups using the three software tools. Ranking of the peptides was done on the basis of q-values.

All workflows detected significant differences for Peptides 2-5, while Peptides 6-9 were consistently missing in all. Considering that we expect only the MassPrep peptides (Peptides 1-9) to be differentially abundant between the two groups of samples, Progenesis had only one false positive, while XCMS had the largest number of false positives in comparison to the other two tools. Since the feature-based alignment approach is applied in each of the three

software tools, it was straightforward to proceed with the alignment once peak detection and peak grouping were well done. However, feature-based alignment methods involve a trade-off between sensitivity (peak detection) and specificity (peak grouping) which is crucial to the alignment result. To get rid of noisy peaks, it is common to apply criteria regarding intensity and span range (available in the three software tools). For low-abundance peptides, more rigorous ways of considering peptide characteristics are desired.

Table 1. Rank, p-value, q-value, and fold change of spike-in peptides in comparing spike-in serum samples versus serum samples alone using Progenesis LC-MS, XCMS, and MZmine 2. Ranking is based on q-values. FC=fold change.

Progenesis LC-MS

Peptide No.	Peptide Sequence	m/z(charge)	Rank	p-value	q-value	FC
1	RGDSPASSKP	501.27 (2)				
2	DRVYIHP	450.25 (2)	3	4.0 E-09	8.2 E-07	21.1
3	RPPGFSPFR	354.21 (3)	2	3.7 E-10	1.1 E-07	Inf
		530.80 (2)	1	<1.1E-16	1.1 E-16	170.4
4	DRVYIHPF	523.78 (2)	5	9.3 E-07	1.1 E-04	340.5
5	DRVYIHPFHL	432.91 (3)	4	9.8 E-08	1.5 E-05	12.9
6	DRVYIHPFHLLVYS	587.00 (3)				
7	WLTGPQLADLYHSLMK	624.81 (3)				
8	YPIVSIEDPFAEDDWEAWSHFFK	943.29 (3)				
9	GIGAVLKVLTTGLPALISWIKRKRQQ	711.90 (3)				

XCMS

Peptide No.	Peptide Sequence	m/z(charge)	Rank	p-value	q-value	FC
1	RGDSPASSKP	501.27 (2)				
2	DRVYIHP	450.25 (2)	1	4.2 E-06	0.0019	25.5
3	RPPGFSPFR	354.21 (3)	3	1.3 E-04	0.0205	42.7
		530.80 (2)	4	1.3 E-04	0.0205	41.8
4	DRVYIHPF	523.78 (2)	16	0.0013	0.0205	350.8
5	DRVYIHPFHL	432.91 (3)	44	0.0031	0.0205	4036.4
6	DRVYIHPFHLLVYS	587.00 (3)				
7	WLTGPQLADLYHSLMK	624.81 (3)				
8	YPIVSIEDPFAEDDWEAWSHFFK	943.29 (3)				
9	GIGAVLKVLTTGLPALISWIKRKRQQ	711.90 (3)				

MZmine 2

Peptide No.	Peptide Sequence	m/z(charge)	Rank	p-value	q-value	FC
1	RGDSPASSKP	501.27 (2)	679	0.3600	0.1700	0.58
2	DRVYIHP	450.25 (2)	1	1.9 E-05	0.0058	28.9
3	RPPGFSPFR	354.21 (3)	6	8.0 E-04	0.0370	65.8
		530.80 (2)	7	8.0 E-04	0.0370	40.7
4	DRVYIHPF	523.78 (2)	18	0.0027	0.0420	132.8
5	DRVYIHPFHL	432.91 (3)	19	0.0028	0.0420	729.6
6	DRVYIHPFHLLVYS	587.00 (3)				
7	WLTGPQLADLYHSLMK	624.81 (3)				
8	YPIVSIEDPFAEDDWEAWSHFFK	943.29 (3)				
9	GIGAVLKVLTTGLPALISWIKRKRQQ	711.90 (3)				

Based on MS/MS search results, we located regions of interest for LC-MS runs at MS1 level and examined the peak profiles of the MassPrep peptides. Intensities of differentially expressed peptides, such as Peptides 2, 3, 4, and 5 are more than ten times higher than Peptides 1, 6, and 7. We may include the low-abundance ones into analysis by decreasing the threshold for intensity or signal-to-noise ratio to a certain amount. However, a huge number of peaks (two times more) would also be detected and many of them around the low-abundance peaks. Since XCMS does not provide specific pipelines for further filtering the detected peaks, irrelevant peaks would inevitably be included. This consequently results in wrong estimation of peak information and ambiguous difference detection. Even with a conservative way for peak detection, the number of irrelevant features is still much higher in XCMS than Progenesis LC-MS and MZmine 2. The latter two consider isotopic peaks.

4. Conclusions

We present a spike-in experiment to evaluate the performance of three software tools in detecting differentially abundant peptides by a label-free LC-MS method. The performance of each tool is assessed by the ability to detect and pick spike-in peptides as differences between groups. Tools that accurately identify the spike-in peptides are likely to be useful in biomarker discovery research.

We observe that selection of the appropriate parameters in using each software tool is very important. Also, to assess the parameters for each pipeline explicitly, modular workflows are desired. From our study, it is evident that the identification of charge states and isotopic patterns greatly reduces irrelevant features and is crucial to improve the performance of a tool in difference detection.

All tools resulted in picking up only four out of nine spike-in peptides. Progenesis LC-MS led to the least number of false positives and the best ranking and statistical significance of the selected features. Although Progenesis LC-MS supports a complete pipeline for label-free proteomics and provides information on conflicting peptides identified for the same feature, certain parameters for peak grouping are not available for tuning. While XCMS and MZmine 2 provide the flexibility for software parameterization and extended analysis, they lack an identification module.

Our analysis also highlighted the problem of optimal intensity threshold used for complex mixtures of peptide. We observed that by reducing the intensity threshold, three spike-in peaks that were earlier not detected could now be included in the analysis. However, this alone will increase the number of extracted features that affects peak grouping.

The LC-MS data from our spike-in experiment can be used as a benchmark for developing and optimizing LC-MS data preprocessing algorithms and to evaluate workflows implemented in existing software tools.

5. Acknowledgements

This work was supported in part by the National Science Foundation Grant IIS-0812246 and the NIH Research Project Grant Program RO1GM086746. The LC-MS data presented in the manuscript were generated through support from the Proteomics and Metabolomics Shared Resource at Lombardi Comprehensive Cancer Center.

6. References

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198-207, Mar 13 2003.

[2] E. P. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Mol Cell Proteomics*, vol. 3, pp. 367-378, Apr 2004.

[3] S. D. Patterson and R. H. Aebersold, "Proteomics: the first decade and beyond," *Nat Genet*, vol. 33 Suppl, pp. 311-323, Mar 2003.

[4] M. Tyers and M. Mann, "From genomics to proteomics," *Nature*, vol. 422, pp. 193-197, Mar 13 2003.

[5] S. P. Gygi, *et al.*, "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags," *Nat Biotech*, vol. 17, pp. 994-999, 1999.

[6] D. K. Han, *et al.*, "Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry," *Nat Biotech*, vol. 19, pp. 946-951, 2001.

[7] Y. Shiio and R. Aebersold, "Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry," *Nat. Protocols*, vol. 1, pp. 139-145, 2006.

[8] J. Lill, "Proteomic tools for quantitation by mass spectrometry," *Mass Spectrom Rev*, vol. 22, pp. 182-94, May-Jun 2003.

[9] D. R. Goodlett and E. C. Yi, "Stable isotopic labeling and mass spectrometry as a means to determine differences in protein expression," *TrAC Trends in Analytical Chemistry*, vol. 22, pp. 282-290, 2003.

[10] W. M. Old, *et al.*, "Comparison of label-free methods for quantifying human proteins by shotgun proteomics," *Mol Cell Proteomics*, vol. 4, pp. 1487-502, Oct 2005.

[11] L. Florens, *et al.*, "A proteomic view of the Plasmodium falciparum life cycle," *Nature*, vol. 419, pp. 520-6, Oct 3 2002.

[12] J. Gao, *et al.*, "Changes in the protein expression of yeast as a function of carbon source," *J. Proteome Res.* , vol. 2, pp. 643-9, 2003.

[13] D. Radulovic, *et al.*, "Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry," *Mol Cell Proteomics*, vol. 3, pp. 984-97, Oct 2004.

[14] A. Prakash, *et al.*, "Signal maps for mass spectrometry-based comparative proteomics," *Mol Cell Proteomics*, vol. 5, pp. 423-32, Mar 2006.

[15] W. Wang, *et al.*, "Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards," *Anal Chem*, vol. 75, pp. 4818-26, Sep 15 2003.

[16] K. M. Pierce, *et al.*, "A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data," *Analytical Chemistry*, vol. 77, pp. 7735-7743, 2005.

[17] P. Kearney and P. Thibault, "Bioinformatics meets proteomics--bridging the gap between mass spectrometry data analysis and cell biology," *J Bioinform Comput Biol*, vol. 1, pp. 183-200, Apr 2003.

[18] T. Pluskal, *et al.*, "MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data," *BMC Bioinformatics*, vol. 11, pp. 395, 2010.

[19] C. A. Smith, *et al.*, "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, pp. 779-787, 2006.

[20] P. M. Palagi, *et al.*, "Proteome informatics I: Bioinformatics tools for processing experimental data," *Proteomics*, vol. 6, pp. 5435-5444, 2006.