Evaluation of Computational Platforms for LS-MS Based Label-Free Quantitative Proteomics: A Global View

Runxuan Zhang[1], Alun Barton[1], Julie Brittenden[2], Jeffrey T. -J. Huang[1,3]* and Daniel Crowther[1,3]*

[1]*Translational Medicine Research Collaboration, University of Dundee, James Arrott Drive, Ninewells Hospital Dundee, DD1 9SY, UK*
[2]*Department of Surgery, Polwarth Building University of Aberdeen, Aberdeen AB25 2ZD UK*
[3]*Pfizer Inc*

Abstract

Label-free shotgun proteomics is a promising semi-quantitative protein profiling method with capability of comparing a large number of samples in a single experiment. One of the key challenges in this proteomics approach is the high requirement of computational capability for tasks such as feature detection and LC-MS alignment due to the complexity of proteomics systems. Many software tools have been developed in recent years to aid these processes, yet it is often not clear to users whether these tools extract information from raw data correctly and comprehensively. In this paper, we described a comprehensive procedure to provide a fast and global view for performances of LC-MS label-free computational software. Two high quality mass spectrometry datasets with carefully controlled QC samples and spiked-in proteins were also provided as benchmark datasets for such evaluations.

Abbreviations: ADH: Alcohol dehydrogenase; CV: Coefficient of Variation; FWHM: Full Width at Half Maximum Intensity; PCA: Principle Component Analysis; QC: Quality Control; SD: Standard Deviation

Keywords: Bioinformatics; Label free; LC-MS; Alignment; Feature detection

Introduction

LC-MS based quantitative shotgun proteomics has gradually replaced traditional two-dimensional gel electrophoresis to become a method of choice for profiling protein composition in a given biological system. There are two major LC-MS based shotgun proteomics approaches commonly employed: stable isotope labeled method and label-free method. (For review on these methods, please see America and Cordewener, 2008; Gevaert et al., 2008; Goshe and Smith, 2003; Ong et al., 2003; Pan and Aebersold, 2007; Wang et al., 2008). With increased availability of high resolution and accuracy MS instruments, label-free shotgun quantitative proteomics has gained great popularity in recent years due to the capability of comparing a large number of samples without resource intensive and potentially biased labeling steps. Such a capability is particularly critical for clinical proteomics as inter-individual variation can be substantial and experiments with large sample sizes are required. In addition, the label free method has proven to provide a higher dynamic range for quantification and more analytical depth by saving all the machine cycles on fragmenting all forms of labeled peptides (Bantscheff et al., 2007; Mueller et al., 2008).

The principle of label-free quantitative proteomics is based on comparison of precursor ion intensities across all experiments after all features (defined as isotopic clusters) are aligned according to their LC retention time, m/z and charge states. Here the aligned features across multiple runs and experiments are defined as consensus features. There are two crucial requirements in order to make this approach successful. First, as the retention time, precursor mass/charge ratio and charge state are usually the only parameters by which a "feature" is defined, the requirements for reproducible LC retention time, high resolution and mass accuracy for mass spectrometers are critical to ensure that signals are well separated (Norbeck, et al., 2005) and the same peptides are compared across experiments. Recent advances in technology have made these requirements achievable.

For example, with instruments such as the Orbitrap and the new generation of Q-TOF, the resolution and mass accuracy can reach up to 100,000 (FWHM) and <1 ppm. Similarly LC retention time has becoming increasingly reproducible with novel technologies such as Agilent's HPLC-Chip (Vollmer and van de Goor, 2009). The other key requirement is good bioinformatics software that provides accurate computation since label-free proteomics experiments typically contain tens to hundreds of LC-MS runs, which yield tens of thousands of features in MS spectra and peptide information in MS/MS spectra, if available. In addition, the inherent biases and variations in MS data add another layer of complexity for computational tasks (Griffin et al., 2009). Therefore the computational capability required for label-free proteomics is very demanding. Computational platforms have been developed in recent years to aid these processes(Jaffe et al., 2006; Li et al., 2005; May et al., 2007; Mueller et al., 2007; Sturm et al., 2008). An assessment of software solutions for MS-based quantitative proteomics has been done by Mueller, et al (Mueller et al., 2008). However, general guidelines are missing for evaluating the software as a whole platform with respect to extracting information from raw data correctly and comprehensively.

In general, the computational analysis of label-free proteomics dataset consists of two major steps: feature detection and alignment. Feature detection is to extract peptide induced signals from ion

chromatography and represent each feature by mass charge ratios, charge, retention time and intensity values. It is an essential step and it serves as the basis for all the later steps. Feature detection usually consists of three steps: smoothing, baseline correction and peak detection (Yang et al., 2009). There have been many methods of different natures being proposed for these steps. Combinations of these methods have been used in MS data analysis software packages. A detailed list of these methods can be referred in (Yang et al., 2009). Alignment is a procedure to search correspondences of retention times across multiple LC-MS experiments such that each feature from the same molecule across different experiments can be grouped together (see Lange et al., 2008; Vandenbogaert et al., 2008for review).

Dozens of algorithms have been developed for feature detection and alignment (Reviews see. Vandenbogaert et al., 2008; Yang et al., 2009). Efforts have been made to evaluate the feature detection and alignment methods as individual steps (Lange et al., 2008; Yang et al., 2009), yet there has been no methods being proposed to evaluate computational platforms as a whole. For most of the MS analysis software packages, these two processes are highly interdependent and the effects of two processes are often convoluted together. It is very difficult to trace back problems to individual steps based on downstream analysis results. The methods employed in different steps are also very heterogeneous so that it is difficult to compare results from different software packages directly. Moreover, the general lacking of ground truth where characteristics such as the number of features in a data file are clearly defined makes the comparisons for these computational platforms even more difficult.

Thus a systematic evaluation method by downstream analysis is critical because the computational methods are heterogeneous and the quality of result highly depends on the performance of these software packages. In many cases, especially for commercial software, details of algorithms and intermediate results are not always available. Instead of unraveling the effects of individual steps, we adopt a statistical view on evaluating the performance of the whole computational platform by downstream analysis. To evaluate a software/algorithm for label free proteomics, two essential elements are required: (a) a comprehensive evaluation method; (b) high quality datasets. It is desirable to develop an evaluation method that reflects the performance of computational platforms in different aspects from downstream results without going into the details of the algorithms or doing additional specifically designed experiments.

In this study, we proposed a comprehensive evaluation method for LC-MS label-free computational platforms and two high resolution datasets with effective experimental designs and carefully controlled procedures were provided as benchmarks for such evaluation purposes.

Materials and Methods

Datasets

Label-free evaluation DATASET 1 was derived from a study comparing the protein composition between fresh frozen femoral and carotid plaques from atherosclerotic patients (gender and age matched). The details of experimental conditions can be found in the supplementary material. In brief, soluble proteins from human atherosclerotic plaque tissues were extracted followed by trypsin digestion. 400 ng peptides from each sample were subjected to nano-

LC-LTQ-Orbitrap. 25 samples in the femoral group and 29 samples in the carotid group were balanced randomized in the analysis. 11 QC samples (a pool of all samples) were also periodically performed to ensure the consistency of the instrument. Mass accuracy in the experiment was estimated < 5 ppm in MS1 and < 0.8 a.m.u. in MS2. The resolution for Orbitrap was set at 30,000. Principle component analysis of all samples (including QCs) in the soluble fractions indicated that QCs were tightly clustered together, suggesting that the system derived variation over the experimental period were well controlled (see Supplementary Materials).

Label-free DATASET 2 contains two samples: SSA001 and SSA002 which contains different levels of 4 proteins (yeast enolase, yeast alcohol dehydrogenase (ADH), bovine serum albumin, and rabbit phosphorylase B) in the background of tryptic digested serum proteins. The human serum sample (from a pool of mixed gender) was purchased from Sera Lab and was collected in plain tubes (BD Biosciences). The clotting was performed at room temperature for 30 minutes and the tube was centrifuged at 1,300g for 20 minutes at 4 °C before the storage at -80 °C. The two samples were analyzed with nano-LC-MS/MS in triplicates. The amount of proteins spiked-in for SSA001 were 50 pmol yeast enolase, 50 pmol yeast ADH, 50 pmol bovine serum albumin, 50 pmol rabbit phosphorylase B. The amount of proteins spiked-in for SSA002 were 100 pmol yeast enolase, 50 pmol yeast ADH, 400 pmol bovine serum albumin, 25 pmol rabbit phosphorylase B. The final volume was 500 μL. Mass accuracy in the experiment was estimated < 5 ppm in MS1 and < 0.8 a.m.u. in MS2. The resolution for the Oribtrap was set at 30,000.

Data processing

"Features", clusters of monoisotopic peaks with distinct m/z, retention time and charge state, in raw data were processed using Elucidator Version 3.3 (Rosetta[1]), and Progenesis Version 2.1 (Non-linear Dynamics). For Elucidator, the process started from the alignment of raw LC-MS images, followed by background removal and extraction and quantification of the peak regions. For Progenesis, features in all raw files were aligned against a pre-selected QC file based on the m/z, retention time and charge state of precursor ions. Intensity of each feature in each raw file was then calculated by the sum of area under curve of all monoisotopic peaks. To identify the spike-in proteins, database search on MS2 spectra were performed using Sequest through Biowork Version 3.3.1 SP1 against human IPI database version 3.3.8. The mass tolerance for precursor ions and fragment ions were 7.5 ppm and 0.8 a.m.u. The settings for static modifications were carboxyamidomethylation at cysteine residues and the dynamic modifications were oxidation at methionine residues. False discovery rates for protein identifications were controlled at 1.0%. The search results were output as xml format or out files, which were subsequently imported into Progenesis. Dataset containing all essential information (all basic characteristics of features, intensity of features in all samples and peptide identification) were then output to an excel format. Intensities of features were normalized to total ion current prior to further analysis. Finally the calculations of all the measurements proposed in this paper were carried out in R version 2.8.1.

Evaluation method

Given the consensus feature matrix, features are aligned across all the samples. All the following measurements are calculated based on the results by each of the software package.

A. Obtain the number of features detected per run for all 11 QC samples. Calculate the mean and the variance. A higher mean indicates higher detection sensitivity and a small variance indicates high consistency.

B. Determine the number of missing values for each consensus feature across all QC samples and group them in bins depending on the number of missing values per feature. A small number of features with missing values indicates a high consistency in feature detection and good performance in alignment.

C. Calculate the Coefficients of Variances (CV) for consensus features with no missing values across 11 QC samples. Low CV indicates the consistency in feature intensities.

D. Calculate the mean and variance of pairwise Pearson correlations of feature intensity between each pair of QC samples. A high correlation indicates consistency in feature intensities across the samples.

E. If two computational platforms are to be compared, match the consensus features from different software packages by their mass range as well as retention time. If there is an overlap between two aligned features in both mass range and elution time range, these are considered a matched feature pair. The unmatched features are considered as unique to each platform. A random sampling of 100 unmatched features is subject to manual validation. False discovery rate is then estimated based on manual validation results. Low false discovery rate indicates a high accuracy in feature detection.

F. Perform t-test for each consensus feature between comparative experiments, which were carotid and femoral plaque samples in this dataset, to test if there is any significance change of feature intensities between comparative samples. Obtain the number of differentially expressed features with a predefined p-value threshold. A high number of different expressed features indicates that the computational method preserves and elucidates differences between comparative samples effectively.

G. The quantification accuracy is evaluated on human blood serum dataset, which contains 4 spike-in proteins with predefined ratios. Calculate ratios between comparative experiments based on the average of the feature intensities of three replicates and then rescaled so that Yeast ADH should have a ratio of 1 Compare the calculated ratios with the predefined ratios. Small differences indicate higher accuracy in quantification.

Results

Instead of breaking down the effects of processing steps, we take a global view on the performance of the whole computational platform. A series of measurements were proposed to evaluate key qualities of label-free proteomics computational platform based on the end result output (Table 1). We started our analysis at the downstream results of a list of consensus features. A consensus feature is composed of aligned features from all the samples, represented by its m/z, retention time and intensity. We evaluated the software packages by looking through the statistics of these consensus features on detection sensitivities (A), consistencies (B, C, D) and accuracies (E), quantification accuracies (G), and statistical potential in detecting differences between comparative samples (F).

Two commercial software packages: Elucidator and Progenesis were used as a case study to illustrate the performance of the evaluation method, although this method can be applied generally to the comparison of any computational platforms. Only features with 2-4 charge will be considered in the following analysis as the charge state of most tryptic digested peptides are likely between +2 and +4.

A. **Number of features detected for each QC sample:** In general, it is expected that a high number of consensus features with few false detections would suggest better, more sensitive feature detection. The number of consensus features picked up by the platform per experiment can be an indicator for sensitivity of detection capability of the platform. Even with the same raw data files analyzed, there were significant differences in the number of features detected between the two software. Elucidator detected on average 23851 features, which was 43.17% more than Progenesis (Mean=16659.09). In addition, the data also showed a higher consistency in the number of features detected across all QC data using Elucidator (Elucidator, SD=2.9 vs. Progenesis, SD=658.2, Figure 1).

B. **Number of missing values of feature intensity across QC samples:** Missing values in a label-free proteomics dataset is very common. For repetitive runs such as the QC data provided in this study, we expected to have some missing values particularly in low signal features due to subtle changes in LC-MS system over time. However, as we used the same QC files for evaluating analytical platforms, a higher number of missing values could indicate either inconsistency of the feature detection or mis-alignment, or both. The number of consensus features with such a number of missing
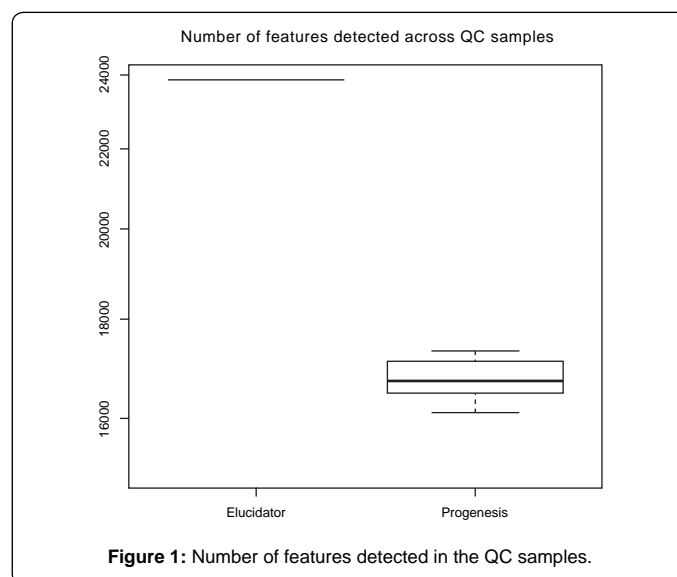
	Measurement/Test	Indications
A	Number of features detected for each QC sample	Detection sensitivity
B	Number of missing values of feature intensity across QC samples	Detection consistency
C	Coefficients of Variations of feature intensity for all QC samples	Intensity consistency & accuracy
D	Correlations of features intensity between QC samples	Intensity consistency
E	Manual curation of unmatched features from comparative platforms for all QC samples	Detection accuracy
F	T-test on features of comparative experiments (Carotid and Femoral)	Statistical capability
G	Quantification of spiked-in proteins on the human serum dataset	Quantification accuracy

Table 1: Measurements for computational platform evaluation.

Figure 1: Number of features detected in the QC samples.

Number of missing values	Elucidator	Progenesis
0	23824	13175
1	19	1191
2	10	784
3	0	570
4	2	502
5	0	520
6	0	506
7	0	478
8	1	536
9	0	617
10	0	881
11	0	1734

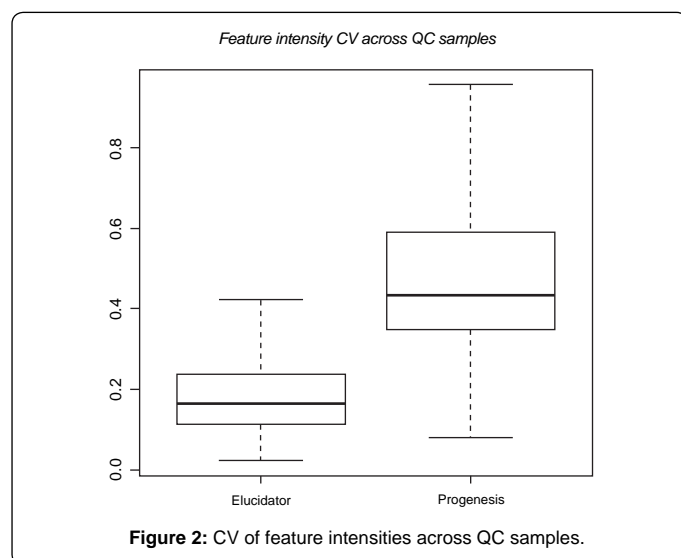**Table 2:** Number of missing values across QC sample.

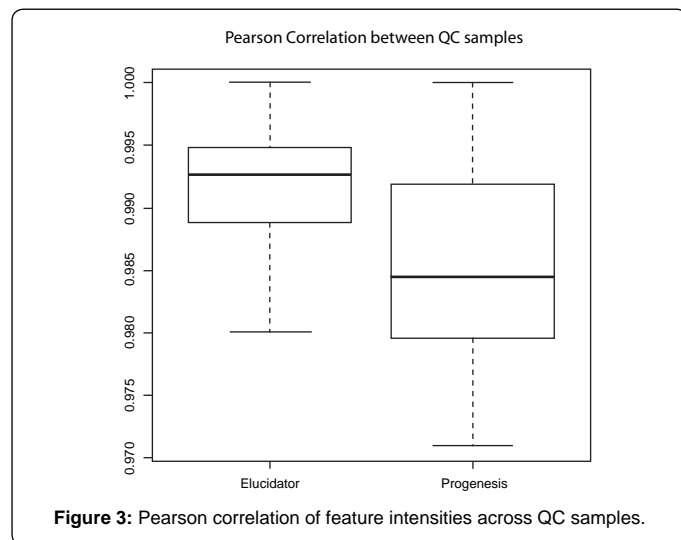**Figure 2:** CV of feature intensities across QC samples.

Figure 3: Pearson correlation of feature intensities across QC samples.

values across the 11 QC samples from both software were listed in Table 2. Elucidator showed a significantly smaller number of features with missing values compared to Progenesis, indicating that Elucidator detected feature signal consistently across all the QC samples and was able to align them together.

C. Coefficients of Variations of feature intensity for all QC samples: Being able to produce results with a small number of consensus features with missing values is necessary for a good computational

tool on label-free proteomics, but it is not sufficient because consensus features may be composed of bogus features, which could come from either random instrumental or chemical noise or peptides with similar retention time and m/z being clustered together by mistake.

Since our QC samples should be identical both in both composition and volume, correctly clustered consensus features which contain less bogus features are expected to have smaller variations in feature intensity across the QC samples. Coefficients of variation (CV) were used as a measure of the consistency of the consensus features. The CV, mean and standard deviation of all the consensus features for Elucidator and Progenesis are shown in (Figure 2). Lower CV with small deviation by Elucidator suggests high consistency in feature intensity as well as less bogus features or mis-alignment in the consensus matrix.

D. Correlations of feature intensities between QC samples: In addition to looking at the variations of consensus feature intensities, the quality of alignment can also be evaluated through the correlation of feature intensities between aligned experiments. The overall difference of feature intensities between aligned control experiments should be minimized. Pairwise Pearson correlations of feature intensities were calculated across QC samples in a one versus one fashion and yielded an 11×11 correlation matrix. The mean and standard deviation of the correlation values generated from the Elucidator and Progenesis software is shown in (Figure 3). The higher correlations (with smaller deviation) of the feature intensities aligned by Elucidator indicates the consistency of the software package, which agrees with the CV analysis results (Figure 2).

E. Manual curation of unmatched features: Since the consensus features found by two platforms are less likely to be false positives, we focused on the features which are unique to each platform for manual validation. We matched the consensus features from different software packages by their mass range as well as retention time. In this experiment, if there was an overlap between two aligned features in both mass range and elution time range (with a +/-2.5 mins tolerance), they were considered a matched feature pair. The unmatched features were considered as unique to each platform. A random sampling of 100 unmatched features from each of the software was subject to manual validation. False discovery rate was then estimated based on manual validation results. For progenesis, 76 out of 100 features were validated as correct detections, while 99 features picked by Elucidator were validated as correct, which yield a higher peak detection accuracy of 99%. A two-sided test of proportion was carried out and the null hypothesis of Progenesis has the same feature detection accuracy as Elucidator was rejected with p value 2.5E-06.

F. T-test on features of comparative experiments: In addition to achieving high consistency within QC samples, good software packages are also expected to preserve the differences between comparative experiments because features with significant expression level changes are often the most relevant biologically, which are of great interest for such experiments. Good computational methods should preserve such differences as much as possible while keeping the data among the replicates highly consistent. Thus in this dataset, t-tests were performed on the consensus feature intensities between femoral and carotid samples. Lists of differentially expressed features were generated and there were 7876 and 3595 features (p<0.01) differentially

Protein	Sample 1 (pmol)	Sample 2 (pmol)	Expected ratios	Elucidator	Progenesis
Yeast Enolase	50	100	2	2.02	1.65
Yeast ADH	50	50	1	1	1
Bovine Serum Albumin	50	400	8	7.7	2.94
Rabbit Phosphorylase B	50	25	0.5	0.51	0.42

Table 3: Quantification results for Spike-in Proteins.

expressed from Elucidator and Progenesis individually. The result indicates that Elucidator was not only able to produce consistent results across the QC samples, but also to extract higher number of differences between comparative samples.

H. Quantification of spiked-in proteins on the human serum dataset: Finally the evaluation of quantification accuracy is based on the human blood serum data set, which contains 4 spiked-in proteins with pre-defined ratios. The ratios were calculated based on the average of the feature intensities of three replicates, and then linearly scaled so that Yeast ADH should have ratio of 1. The quantification results by Elucidator and Progenesis are shown in Table 3. The smaller error between the expected ratio and the calculated ratio indicated Elucidator was more accurate in quantification as well.

Discussion

Results interpretation

The evaluation method we used is based on the downstream analysis results, which can be influenced by any of the upstream procedures. Rather than dissecting the effects of individual steps, here we set out to interpret the evaluation result from a global view, which is more of interest when a decision on the choice of software needs to be made. Such a method is advantageous because often the computational methods are heterogeneous and the quality of result highly depends on the performance of these software packages. In many cases, details in algorithms and intermediate results are not always available. Thus a statistical evaluation the performance of the whole computational platform by downstream analysis is adopted. A table of measurements serving for these purposes and their indications are shown in Table 1.

When evaluating the performances of computational platforms, all the related measurements should be considered together. For example, Elucidator detected not only high number of features with few missing values but also with small intensity CVs across all the QC samples. It implies that higher number of features with less missing values by Elucidator was not attained by sacrificing the accuracy of feature detection and alignment. While the measurements on QC samples evaluate platform consistency, the t-test of comparative experiments helps to confirm the superior performances were not based on over-normalization of the datasets. Meanwhile, the manual validation results were also used to estimate the false discovery rates of identified features in order to rectify the detection of true signals.

Benchmark datasets

The availability of high quality benchmark datasets is one of critical parts for computational platform evaluations. Previous efforts have been made to set up common benchmark sets for such purposes (Lange et al., 2008). In this study, we have provided two high quality datasets generated by nano-LC-LTQ-Orbitrap where mass accuracy and retention time were well-controlled and the experiment was designed with great care as indicated by PCA (Supplementary Materials). The raw data are also included in the Supplementary Materials.

Ground truth definition

The evaluation of label-free proteomics computational software has always been a difficult task because of the general lack of ground truth on which this evaluation can be made. Two commonly used alternative methods are manual curation or using MS/MS identifications to evaluate associated features. However, it should be noted that: firstly MS/MS identification is not always available for label-free experiments. Secondly, MS/MS identifications show bias towards features with high intensity, while features derived from low intensities peaks are more likely to problematic for the software packages. Thirdly, assignment of MS/MS identifications to features is a potential biased process due to the low sampling rate and inconsistency of fragmentation precursor selection. Thus the manual curation was employed in this work. It enables a detailed look at the differences of comparative computational platforms and further identifying the origin of differences.

It should be noted that all the measurements that are compared between software packages are parameter dependent. The choice of parameters always influences the sensitivity and accuracy at the same time- higher sensitivity often leads to lower accuracy and vice versa. The selection of optimal parameter is to find the desired trade off between sensitivity and accuracy. The measurements proposed in the paper are also applicable to the optimization of parameters.

In summary, we have developed a simple method for evaluating computational platform and provided datasets for these evaluations. This method adopts a statistical view on evaluating the performance of the whole computational platform by downstream analysis. The proposed evaluation method can be applied for the comparison of computational platforms, for optimization of parameter settings and for understanding the performance of alignment in a given label-free dataset.

Conflict of interests

The authors declare that they have no conflict of interests.

References

1. America AH, Cordewener JH (2008) Comparative LC-MS: a landscape of peaks and valleys. Proteomics 8: 731-749.

2. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389: 1017-1031.

3. Gevaert K, Impens F, Ghesquiere B, Van Damme P, Lambrechts A, et al. (2008) Stable isotopic labeling in proteomics. Proteomics 8: 4873-4885.

4. Goshe MB, Smith RD (2003) Stable isotope-coded proteomic mass spectrometry. Curr Opin Biotechnol 14: 101-109.

5. Griffin NM, Yu J, Long F, Oh P, Shore S, et al. (2009) Label-free, normalized

quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol : 83-89.

6. Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, et al. (2006) PEPPeR, a platform for experimental proteomic pattern recognition. Mol Cell Proteomics 5: 1927-1941.

7. Lange E, Tautenhahn R, Neumann S, Gropl C (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics 9: 375.

8. Li XJ, Yi EC, Kemp CJ, Zhang H, Aebersold R (2005) A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. Mol Cell Proteomics 4: 1328-1340.

9. May D, Fitzgibbon M, Liu Y, Holzman T, Eng J, et al. (2007) A platform for accurate mass and time analyses of mass spectrometry data. J Proteome Res 6: 2685-2694.

10. Mueller LN, Brusniak MY, Mani DR, Aebersold R (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 7: 51-61.

11. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, et al. (2007) SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics 7: 3470-3480.

12. Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, et al. (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. J Am Soc Mass Spectrom 16: 1239-1249.

13. Ong SE, Foster LJ, Mann M (2003) Mass spectrometric-based approaches in quantitative proteomics. Methods 29: 124-130.

14. Pan S, Aebersold R (2007) Quantitative proteomics by stable isotope labeling and mass spectrometry. Methods Mol Biol 367: 209-218.

15. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, et al. (2008) OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics 9: 163.

16. Vandenbogaert M, Li-Thiao-Te S, Kaltenbach HM, Zhang R, Aittokallio T, et al. (2008) Alignment of LC-MS images, with applications to biomarker discovery and protein identification. Proteomics. 8: 650-672.

17. Vollmer M, van de Goor T (2009) HPLC-Chip/MS technology in proteomic profiling. Methods Mol Biol 544: 3-15.

18. Wang M, You J, Bemis KG, Tegeler TJ, Brown DP (2008) Label-free mass spectrometry-based protein quantification technologies in proteomic analysis. Brief Funct Genomic Proteomic 7: 329-339.

19. Yang C, He Z, Yu W (2009) Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics 10: 4.