BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 4054–4059
doi:10.1093/bioinformatics/bti660

Data and text mining

Data pre-processing in liquid chromatography–mass spectrometry-based proteomics

Xiang Zhang[1,*], John M. Asara[3], Jiri Adamec[1], Mourad Ouzzani[2] and Ahmed K. Elmagarmid[2]

[1]Bindley Bioscience Center and [2]Department of Computer Science, Purdue University, West Lafayette, IN, USA and [3]Beth Israel Deaconess Medical Center, Boston, MA, USA

Received on April 13, 2005; revised on August 1, 2005; accepted on September 1, 2005
Advance Access publication September 8, 2005

ABSTRACT

Motivation: In a liquid chromatography–mass spectrometry (LC–MS)-based expressional proteomics, multiple samples from different groups are analyzed in parallel. It is necessary to develop a data mining system to perform peak quantification, peak alignment and data quality assurance.

Results: We have developed an algorithm for spectrum deconvolution. A two-step alignment algorithm is proposed for recognizing peaks generated by the same peptide but detected in different samples. The quality of LC–MS data is evaluated using statistical tests and alignment quality tests.

Availability: *Xalign* software is available upon request from the author.

Contact: zhang100@purdue.edu

INTRODUCTION

Proteomics was initially envisioned as a technique to globally and simultaneously characterize all components in a proteome. In a typical liquid chromatography–mass spectrometry (LC–MS)-based protein expression profiling experiment, multiple samples collected from different patients are analyzed in parallel (Diamandis, 2004). Each sample is digested into peptides and subjected to multi-dimensional liquid chromatography for separation. Each peptide fraction is then analyzed on an LC–MS system. Ideally, the same molecules detected in the same LC–MS platform should have the same retention time, molecular weight and signal intensity. However, this is not the case due to experimental variations (Wang *et al*., 2003). It is very important to recognize peak variation in the same type of molecule, but from different samples, from millions of LC–MS peaks, and to compare them.

Spectral deconvolution, peak alignment and data quality assurance are common tasks in data pre-processing. Several methods have been developed to quantify peaks from LC–MS data (Li *et al*., 2003; MacCoss *et al*., 2003; Zhang *et al*., 2005). Peak alignment recognizes peaks of the same type of molecule occurring in different samples from peaks detected during the course of an experiment (Torgrip *et al*., 2003; Yu *et al*., 2004). This paper reports an LC–MS data pre-processing method for a bottom-up proteomics approach in which peaks from peptide profiles are analyzed. The objective of the work described here is to develop a method to (1) study the quality of the LC–MS results and to (2) align the LC–MS peaks for further statistical analysis.

*To whom correspondence should be addressed.

MATERIALS AND METHODS

The experimental method of this work is identical to the method described in Zhang *et al*. (2005). Briefly, serum albumin and human serum were individually digested with enzyme trypsin. The tryptic digest was then aliquoted into two groups. Each group was labeled with succinimidyl-($^1H_3$)-acetate and succinimidyl-($^2H_3$)-acetate, respectively. The light and heavy labeled peptide mixtures were then combined. Aliquots of 5 μl of the combined light and heavy labeled peptide mixtures were injected and acquired in positive ion mode by LC–MS using a Waters CapLC HPLC instrument and a Waters QTOF micro mass spectrometer. Microcapillary liquid chromatography was operated at 250 nl/min using a 360 μm o.d. × 75 μm i.d. microcapillary column from New Objective Inc. (Woburn, MA), self-packed to 10 cm in length with 10 μm C_{18} from YMC (Kyoto, Japan).

Methods and algorithms

The following sections present the algorithms for the proposed pre-processing LC–MS data. These algorithms include spectral deconvolution, data quality assurance and data alignment. These algorithms have been implemented in software *Xalign* using C++.

Spectral deconvolution Spectral deconvolution was performed using a modified algorithm reported by Zhang *et al*. (2005). The method uses chemical noise filtering, charge state fitting and de-isotoping to improve analysis of complex peptide samples. Spectral noise levels were initially determined based on peak density, and then adjusted using estimated peptide peak profile information. Any peak with intensity less than that of the adjusted noise level was filtered out. The rest of the peaks were further validated at the chromatographic level.

Overlapping peptide signals in mass spectra were deconvoluted using a correlation with modeled peptide isotopic peak profiles. There are two major steps associated with deconvoluting peptide signals. One is ion charge state recognition; the other is correlation of experimentally measured isotopic peak clusters with theoretically predicted isotopic peak profiles. The initial charge assignment relies on the spacing of peaks in the mass-to-charge ratio (m/z) dimension. Peak intensities were used in a subsequent step to address the potential of overlapping signals from multiple peptides using isotopic peak profile information. The isotopic peak profiles for peptides were generated *in silico* from a protein database producing reference model distributions.

Peak alignment We have designed a gross-alignment algorithm to address systematic retention time shift. In the gross alignment, all possible significant peaks were first identified. A significant peak refers to a peak

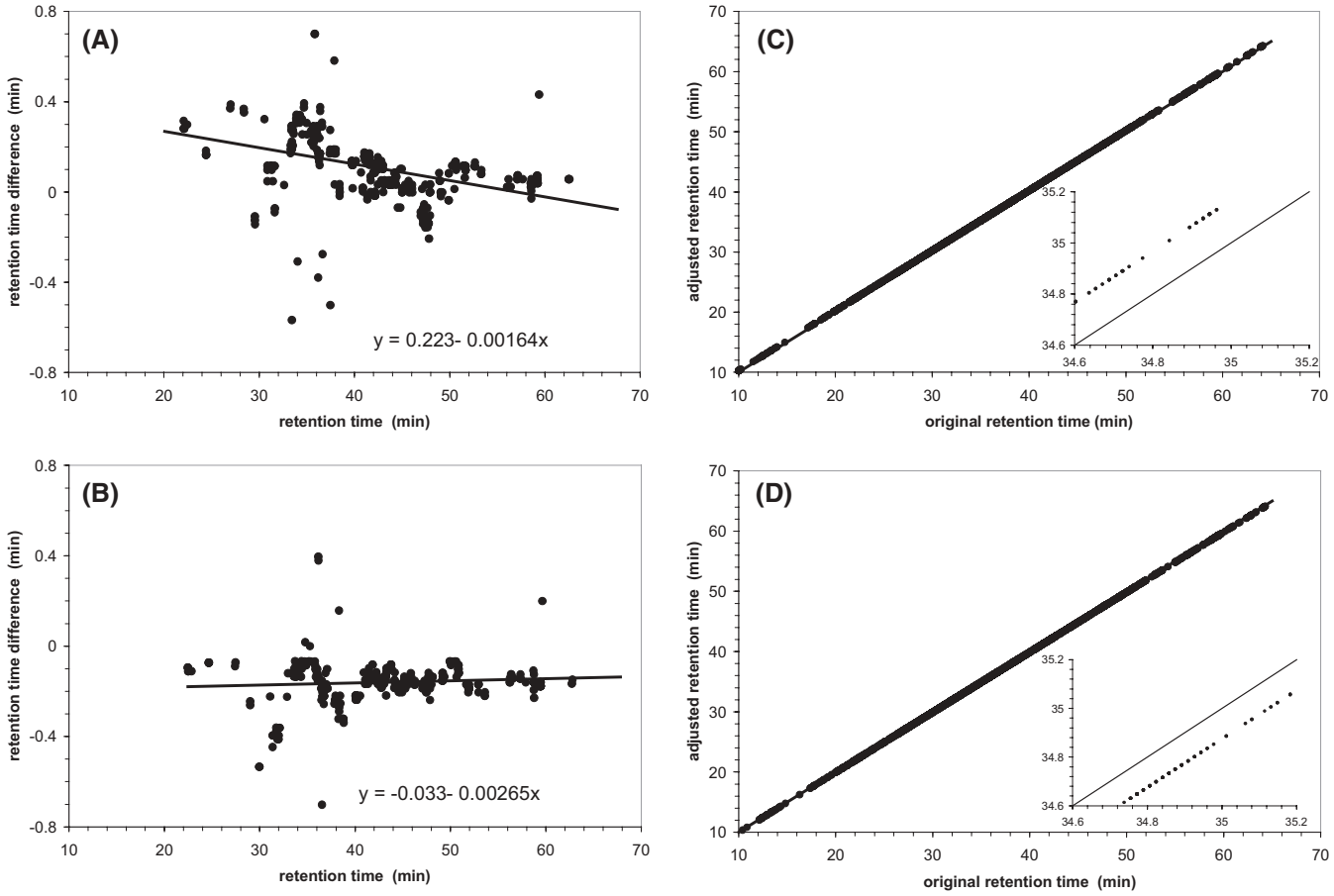

Fig. 1. Gross-alignment of LC–MS peak lists generated from the BSA experiments. Injection 9 was chosen as the median sample. (**A**) Retention time difference between all significant peaks in injection 1 and their corresponding peaks in the median sample. (**B**) Retention time difference between all significant peaks in injection 8 and their corresponding peaks in the median sample. (**C**) Retention time adjustment of injection 1 based on the retention time difference of all significant peaks in injection 1 and the median sample. (**D**) Retention time adjustment of injection 8 based on the retention time difference of significant peaks in injection 8 and the median sample.

that is present in every sample and is the most intense peak in a certain m/z range ($m_i - \varepsilon_m$, $m_i + \varepsilon_m$) and retention time range ($t_i - \varepsilon_t$, $t_i + \varepsilon_t$), where m_i is m/z, t_i is retention time, ε_m is user-provided m/z variation and ε_t is user-provided retention time drift. During the course of finding significant peaks in each sample, an intensity weighted average peak is calculated using Equations (1) and (2).

$$M_j = \Sigma I_{i,j} M_{i,j} / \Sigma I_{i,j} \quad (1)$$

$$T_j = \Sigma I_{i,j} T_{i,j} / \Sigma I_{i,j} \quad (2)$$

where $I_{i,j}$, $M_{i,j}$ and $T_{i,j}$ are the peak intensity, m/z, and retention time of the significant peak j in sample i, respectively. M_j and T_j are the intensity weighted m/z and retention time of the averaged peak of the corresponding significant peak j. The intensity and retention time of each average peak were used to search for a significant peak in the subsequent sample. The procedure for significant peak selection is as follows.

(1) Selection of a sample with a minimum number of peaks as a reference sample, and sorting of all peaks in ascending order of retention time.

(2) Determination of the unprocessed peak with minimum retention time value ($t_{\min}$) in the reference sample. The most intense peak in retention time range ($t_{\min}$, $t_{\min} + \varepsilon_t$) is selected as a significant peak $S_{i,j}$. All peaks with retention time less than $t_{\min} + \varepsilon_t$ are recorded as processed peaks.

(3) Determination of all peaks in the next sample whose m/z and retention time fall into ($M_j - \varepsilon_m$, $M_j + \varepsilon_m$) and ($T_j - \varepsilon_t$, $T_j + \varepsilon_t$), respectively. Selection of the most intense peak as the significant peak in the current sample and recalculate T_j and M_j using Equations (1) and (2). If there is not a corresponding peak found in the current sample, all significant peaks related with peak $S_{i,j}$ will be removed and the program moves to Step 2.

(4) Repeat Step 3 until all samples are processed.

(5) Repeat Steps 2–4 until all peaks in the reference sample are processed.

After identifying all significant peaks, the retention time median μ_j is calculated to each significant peak j. An absolute value of retention time difference between $T_{i,j}$ and μ_j is calculated, where $T_{i,j}$ is the retention time of j-th significant peak of sample i. For each sample, the retention time difference of each significant peak is summed together using Equation (3), where D_i is the summed retention time difference of sample i. The median sample is then defined as a sample that has the minimum sum of retention time difference.

$$D_i = \sum_{j=1}^{j} |T_{i,j} - \mu_j| \quad (3)$$

The final step in the gross-alignment procedure is to align all peaks detected in each sample to the peaks in the median sample. During this

process, the retention time difference between the significant peaks of each sample and the corresponding significant peaks of the median sample will be used to minimize the overall retention time difference between each LC–MS experiment and the LC–MS evaluation of the median sample. It is very likely that some peaks selected as significant peaks are false-positives, which introduces a large retention time difference. It is necessary to filter these peaks out before adjusting the overall peak retention time. For this purpose, we sorted all significant peaks in each sample in ascending order of retention time and grouped data points into multiple groups. Each group contains a certain number of data points. From these, a median data point is found for each group. The retention time difference of each median data point is the median of retention time differences of all data points in that group. Consequently, we only retain the median data points while all others are ignored. Then, the retention times of the retained peaks in each sample are fitted with the retention time of the corresponding peaks in the median sample using a robust estimation. The concept of robust estimation in this instance refers to a statistical estimator that is insensitive to small departures from the idealized assumptions for which the estimator is optimized. *M*-estimates, using maximum likelihood by minimizing the mean absolute deviation, are applied in this work to find the idealized straight-line fit (Press *et al*., 2002). The parameters derived from the significant peak fitting are then applied to the remaining peaks to systematically shift their retention time toward the median sample.

After gross alignment, a micro alignment is used to identify peaks of the same molecule but in different LC–MS datasets. The procedure for the micro alignment is as follows:

(1) A sample that has the lowest number of unprocessed peaks is defined as the base sample. It is assumed that a peak could be detected in the rest of the samples if that peak can be detected in the base sample.

(2) Starting with the peak having the minimum m/z in the base sample, all peaks that overlap each other within a user-defined retention time and m/z window are selected. The most intense peak in the selected peak cluster is defined as a local base peak.

(3) All peaks in the second sample that overlap with the local base peak in m/z and retention time are selected.

(4) Discrete convolution is used to find the peak in the second sample that correlates with the current local base peak. A convolution is an integral that expresses the amount of overlap of one function *s* as it is shifted over another function *h*. Assuming both $s(t)$ and $h(t)$ are digital functions with a sampling interval of unity, the convolution operation is defined as

$$y(t) = s_k * h_k = \sum_{k=-\infty}^{+\infty} s_k h_{j-k} \quad j = 1, 2, 3, \ldots \tag{4}$$

The convolution is implemented as follows: all peaks in the base sample that overlap with the local base peak are defined as function *s* while the peak cluster selected from the second sample is defined as function *h*. The best match between *s* and *h* is defined as a match that gives the maximum value of $y(t)$. If there are multiple choices for the best match, the most symmetric match is defined as the best match. The peak in the second sample that matches the local base peak is considered as the corresponding peak in the second sample and is aligned to the base peak. It is also marked as a processed peak and removed from the second sample.

(5) The algorithm moves to the next sample. Processes 3 and 4 are repeated until all samples are examined using the current local base peak.

(6) The current local base peak is updated by repeating Step 2. The updated local base peak has a different m/z value. The Steps 3–5 are repeated.

(7) If all peaks in the base sample are processed, the algorithm moves to Step 1 to find the next base sample, and Steps 2–6 are repeated.

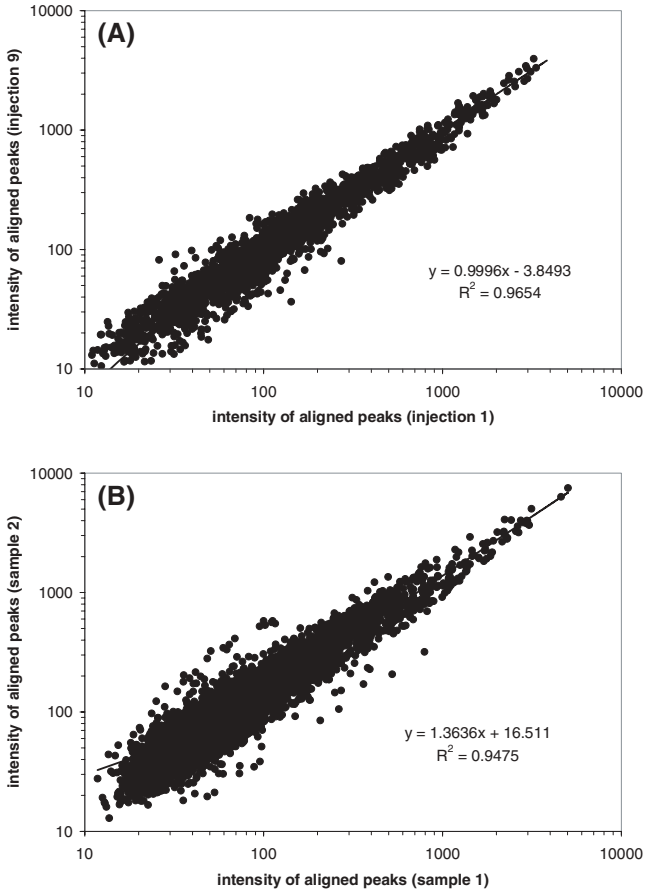

Fig. 2. Sample alignment results between two samples. (**A**) Alignment of peak lists generated from injections 1 and 9 in BSA experiments. (**B**) Alignment of peak lists generated from samples 1 and 2 in serum experiments.

Quality assurance For quality assurance purposes, the main factors of LC–MS data are peak retention time and m/z. Therefore, the two-dimensional Kolmgorov–Smirnov (K–S) test can be applied to study the peak distribution in retention time and m/z plane, where each given peak can be represented as (t_i, m_i). Each peak actually separates the plane into four quadrants $(t > t_i, m > m_i)$, $(t < t_i, m > m_i)$, $(t < t_i, m < m_i)$ and $(t > t_i, m < m_i)$. An integrated probability in each of these four natural quadrants around a given point can be calculated. The statistic *D* of K–S test is taken to be the maximum difference (ranging both over data points and over quadrants) of the corresponding integrated probabilities (Press *et al*., 2002).

The number of peaks detected in an LC–MS experiment can be affected by many factors, such as preparation of the sample, the amount of sample loaded, etc. Therefore, the number of peaks detected is also essential for evaluating experimental quality. During K–S test and peak number test, Sprent's equation (Sheskin, 2000) was used to find statistical outliers:

$$|X_i - M|/\mathrm{MAD} > \mathrm{Max} \tag{5}$$

where X_i is any score being evaluated with respect to whether it is an outlier, and *M* is the median of the scores in the sample. MAD is the median absolute deviation and Max is the critical value that the result to the left of the inequality must exceed in order to conclude that the value X_i is an outlier. The value Max is set as 5.0, which is extremely likely to identify scores that deviate from the mean by more than three standard deviations.

After peak alignment, the number of aligned peaks is studied to ensure the quality of the LC–MS data. The test was done by (1) selecting aligned peaks

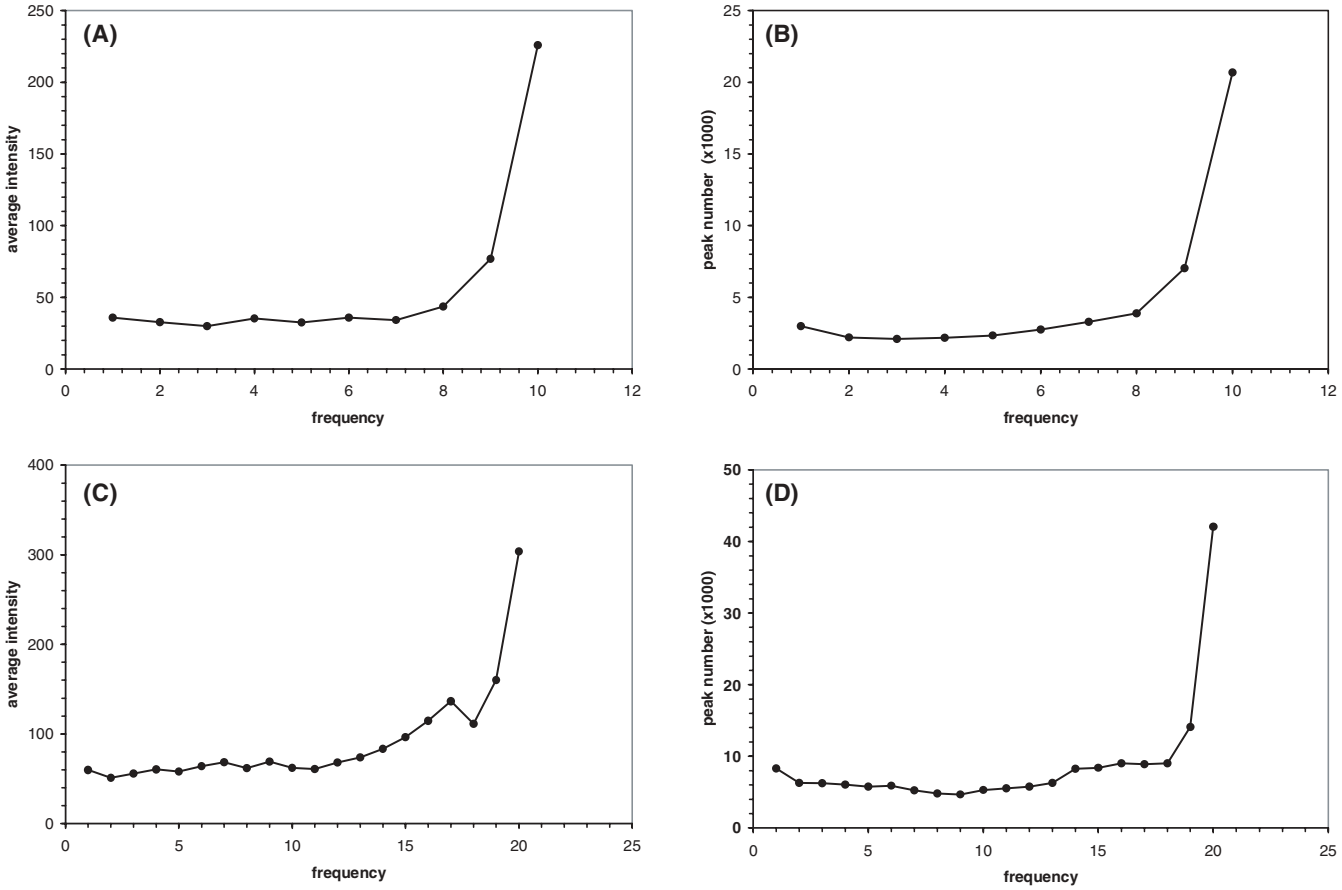

Fig. 3. Alignment statistics of the BSA and serum experiments. The *x*-axis refers to the number of injections/samples in which a peak had been detected, and also shows its alignment with corresponding peaks in the other injections/samples. (**A**) Average peak intensity distribution of BSA experiments. (**B**) Peak number distribution of the BSA experiments. (**C**) Average peak intensity distribution of serum experiments. (**D**) Peak number distribution of the serum experiments.

that were detected in >80% of samples, (2) counting the number of aligned peaks in each sample and (3) using Sprent's equation to detect samples containing a significantly low number of aligned peaks.

RESULTS AND DISCUSSION

All raw MS spectra were subjected for spectrum deconvolution (Zhang *et al*., 2005). Then, the peak lists generated from all BSA and serum experiments were aligned, respectively, assuming that the retention time drift was <0.5 min and m/z variation was <0.1 Da.

Alignment

There are two major steps during gross alignment: finding significant peaks and adjusting the overall retention time drift of all samples. The method used to find all significant peaks is biased to the reference sample, which contains the minimum number of peaks. A peak in the reference sample is considered the most intense peak, though it may not be the most intense peak in the other samples. However, the purpose of using the local intense peak is to increase the chance of finding peaks detected in all other samples. The peak intensity information will not be used to adjust overall retention time drift between LC–MS experiments. On the other hand, it is very typical that a few groups of samples are analyzed in comparative proteomics or metabolomics such as wild-type samples and disease samples. The profiles of the majority of the peaks in these two groups are similar. Only a few peaks (potential biomarkers) may have significant peak intensity differences.

Figure 1 shows a gross-alignment result of BSA experiments where the same BSA digest was injected onto the LC–MS system 10 times. Figure 1A depicts the retention time differences between each significant peak of injection 1 and its corresponding peaks in the median sample (injection 9). Figure 1B displays the same information between signification peaks of injection 8 and injection 9. The injection 1 experiment was performed ∼9 h before injection 9, while the injection 8 experiment was performed immediately before injection 9. It is apparent that there was a systematic retention time drift between two LC–MS experiments. In most cases, this type of drift is linear and can be corrected by a linear regression. It should be noted that some false-positive results could be introduced as significant peaks due to the method used to search these peaks. These false-positive results were greatly reduced by calculating the median value of every five data points. The median values were then used for robust estimation. The straight lines in Figure 1A and B are regression results from the robust estimation.

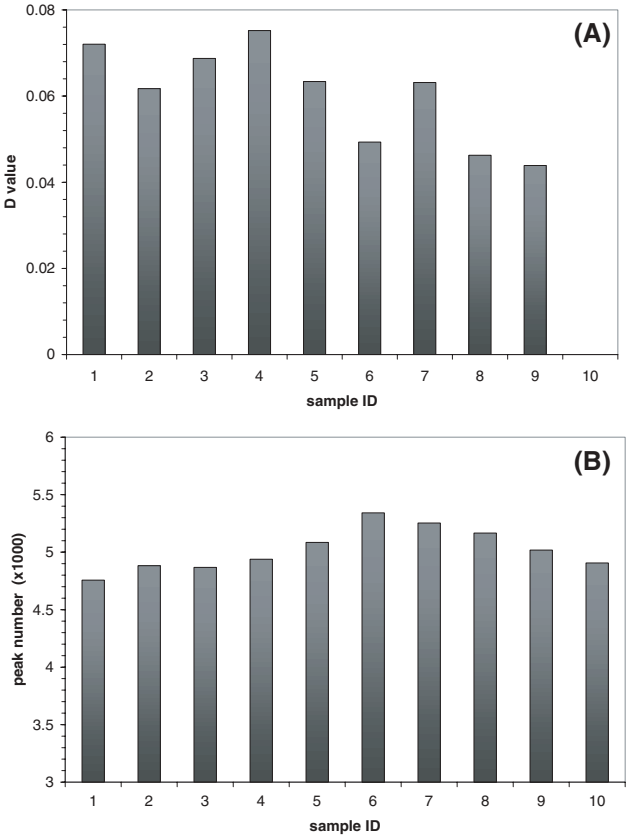

Fig. 4. LC–MS quality assurance of BSA experiments before peak alignment. (**A**) K–S test. (**B**) Peak number test.

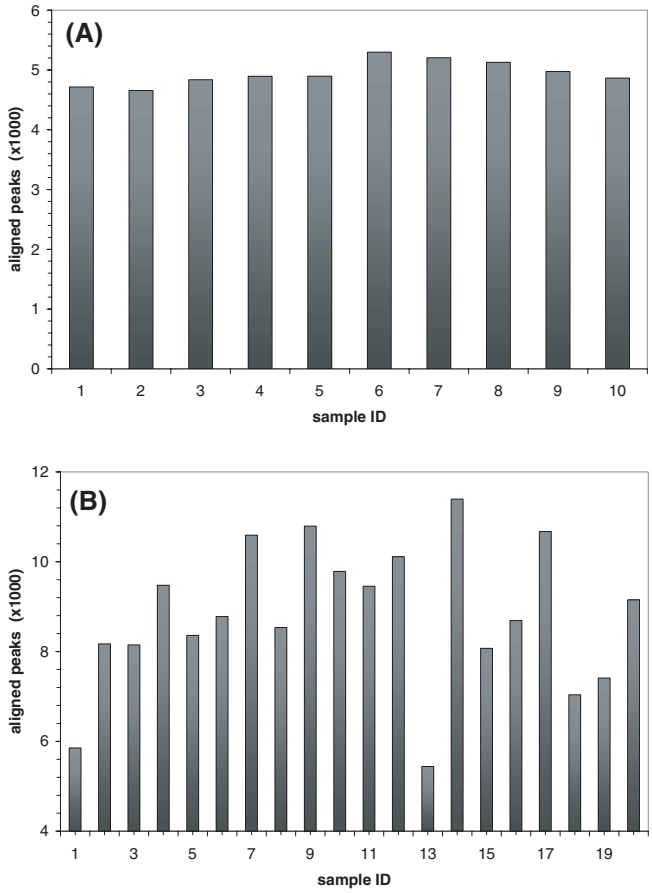

Fig. 5. LC–MS quality assurance after peak alignment. (**A**) BSA experiments. (**B**) serum experiments.

After robust estimation, the retention time of each peak in all samples was adjusted using the results of robust estimation. Figure 1C and D show the effectiveness of the gross alignment for injections 1 and 8, respectively. The straight line in each graph is a guideline showing retention time with zero retention time adjustment. The small figures in the bottom right show results of retention times between 34.6 and 35.2 min.

Following the gross alignment, micro alignment aligns all samples to the median sample. A total of 20 680 peaks were perfectly aligned from the BSA experiments, while 42 080 peaks were perfectly aligned from the serum experiments. A perfect alignment refers to a peak that was not only detected in each sample, but also aligned in each sample. Figure 2 displays alignment results of two randomly selected samples from BSA experiments (Fig. 2A) and serum experiments (Fig. 2B). Some peaks with larger intensity variations have been manually verified. The correlation coefficients indicate that there were some experimental variations in peak intensity, though the BSA experiments have less intensity variation than the serum experiments. The straight lines fitted by linear regression indicate that the overall peak intensities between two samples were slightly different. A peak intensity normalization method needs to be applied to make all samples comparable for further statistical analysis.

In order to evaluate the alignment results, the average peak intensity of each serial of aligned peaks was calculated, and the number of peaks in each aligned peak serial was counted. Figure 3A and B contains alignment results of the BSA experiments. The frequency refers to the number of injections from which a peak has been detected and aligned. Ideally, any peak detected in one injection should also be detected in the other nine injections because the samples were identical. However, due to experimental variation, this is not always the case. Although the majority of peaks were detected in all 10 injections, a fair number of peaks were detected in only one or two injections (Fig. 3A). It is likely that these are random peaks generated by the analysis system. Fortunately, these peaks can be differentiated from the true peptide peaks since most of these peaks are much less intense than the other peaks in the spectra (Fig. 3A). A similar observation can also be found in serum experiments (Fig. 3C and D).

The alignment method reported here uses user-estimated retention time drift and m/z variation as the base for alignment. This may cause a problem if the retention time drift and m/z variation provided by the user are less than the experimental variations. Therefore, some peaks with larger retention time drift or m/z variation will not be aligned. The software provides a simple mechanism to evaluate the alignment table in such a way as to take consider this. This was done as follows.

(1) All peaks (P_i) aligned in >90% of samples were found, and all the samples (S_1) that do not have P_i were recorded.

(2) All peaks (Q_i) aligned in <10% of samples were found, and all of the samples (S_2) that have Q_i were recorded.

(3) The Q_i list was searched to see whether there were some peaks (R_i) that have the same charge state and isotope label as the peaks in the P_i list.

(4) The retention time difference and m/z difference between each peak in the R_i list and the corresponding peaks in the P_i list were calculated. If both the retention time difference and m/z difference are less than two times the user provided value, a message is sent to the user to verify the estimated retention time drift and m/z variation.

Quality assurance

Data quality assurance was performed before and after peak alignment. Figure 4A depicts K–S test results of the BSA experiments. Injection 10 (sample ID 10) was chosen as a reference sample because the peak distribution of injection 10 has the greatest similarity to the remaining 9 samples. The peak distribution of the other 9 samples was then compared with the peak distribution of injection 10. Although there was not any injection detected as a significant outlier in the K–S test (Fig. 4A) and peak number test (Fig. 4B), experimental variations do exist. Generally, <10% of peak number variation can be observed (Fig. 4B). This number may be reduced by filtering peaks based on the signal-to-noise ratio. However, it is not favorable to remove peaks in order to increase the dynamic range of the analytical system.

In order to check the data quality in terms of alignment, the number of aligned peaks in each sample was counted. Of these, each aligned peak must be detected in >80% of the total number of samples. Figure 5 shows the results from the BSA experiments (Fig. 5A) and the serum experiments (Fig. 5B). These figures illustrate that serum samples have a larger variation in aligned peaks. This is primarily due to the increased sample complexity.

In most cases, the quality assurance measures the technical variations caused by the analytical system. Therefore, the evaluation of quality assurance is only used to flag the quality of a dataset and not to determine whether a sample gave rise to a questionable dataset unworthy of further statistical analysis. One reason for this utilization is that all quality control methods reported here focus on peak distribution, retention time drift and m/z variation. As long as the data-mining algorithm such as peak alignment can correctly adjust these variations, the sample dataset may still have some values. The other reason is that biological variation can be larger than technical variation. It is common for scientists to use a ratio of 2.0 as a cutoff value, together with a probability value, to determine whether a change is meaningful. This cutoff value is much higher than the intensity variation caused by the analytical system. The peak intensity variation arising from this analysis system is typically <30% (without normalization) in the case of BSA multiple injection experiments. Therefore, it is commonly accepted that a sample with large technical variation can still provide useful biological information. It is our practice that the decision to remove a sample from the experimental dataset will be made during the statistical analysis.

CONCLUSIONS

We presented a method of pre-processing LC–MS-based proteomics data. A spectral deconvolution method processes LC–MS spectra to the level of a peptide ion. The two-step alignment algorithm provides reliable alignment results. The gross alignment adjusts the overall retention time drift between samples, while the micro alignment focuses on the local complexity and aligns peaks together. Data quality assurance is performed at several different steps. We provide a K–S test, a peak number test and an alignment quality test. The combination of these tests provides a reliable quality assurance system for LC–MS experimental data. This method can also be used to analyze metabolomics data.

ACKNOWLEDGEMENT

This research was partially supported by grant NIH 0GI93.

Conflict of interest: none declared.

REFERENCES

Diamandis,E.P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell. Proteomics*, **3**, 367–378.

Li,X. *et al.* (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.*, **75**, 6648–6657.

MacCoss,M.J. *et al.* (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.*, **75**, 6912–6921.

Press,W.H., Teukolsky,S.A., Vettering,W.T. and Flannery,B.P. (2002) *Numerical Recipes in C++, The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge, UK, pp. 650–654.

Sheskin,D.J. (2000) *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL, pp. 271–272.

Torgrip,R.J.O. *et al.* (2003) Peak alignment using reduced set mapping. *J. Chemometrics*, **17**, 573–582.

Wang,W. *et al.* (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.*, **75**, 4818–4826.

Yu,W. and Zhao,H. (2004) Aligning spectral peaks in mass spectrometry data with a robust point matching approach. In *52nd ASMS Conference on Mass Spectrometry and Allied Topics,* Nashville, TN, May 23–27.

Zhang,X. *et al.* (2005) An automated method for the analysis of stable isotope labeling data in proteomics. *J. Am. Soc. Mass Spectrom.*, **16**, 1181–1191.