Clin Proteom (2009) 5:170–177
DOI 10.1007/s12014-009-9039-y

A Multi-feature Reproducibility Assessment of Mass Spectral Data in Clinical Proteomic Studies

Irene S. L. Zeng · Sharon R. Browning · Patrick Gladding · Mia Jüllig · Martin Middleditch · Ralph A. H. Stewart

Published online: 28 November 2009

Abstract
Background The use of mass spectrometry to investigate disease-associated proteins among thousands of candidates simultaneously creates challenges with the evaluation of operational and biological variation. Traditional statistical methods, which evaluate reproducibility of a single feature, are likely to provide an inadequate assessment of reproducibility. This paper proposes a systematic approach for the evaluation of the global reproducibility of multidimensional mass spectral data at the post-identification stage.
Methods The proposed systematic approach combines dimensional reduction and permutation to test and summarize the reproducibility. First, principal component analysis is applied to the mean quantities from identified features of paired replicated samples. An eigenvalue test is used to identify the number of significant principal components which reflect the underlying correlation pattern of the multiple features. Second, a simulation-based permutation test is applied to the derived paired principal components. Third, a modified form of Bland Altman or MA plot is produced to visualize agreement between the replicates. Last, a discordance index is used to summarize the agreement.
Results Application of this method to data from both a cardiac liquid chromatography tandem mass spectrometry experiment with iTRAQ labeling and simulation experiments derived from an ovarian cancer SELDI-MS experiment demonstrate that the proposed global reproducibility test is sensitive to the simulated systematic bias when the sample size is above 15. The two proposed test statistics (max t statistics and a sign score statistic) for the permutation tests are shown to be reliable.
Conclusion The methodology presented in this paper provides a systematic approach for the global measurement of reproducibility in clinical proteomic studies.

Keywords Multidimensional reproducibility assessment · ITRAQ · Cardiac proteomic · Principal component analysis · Eigenvalue testing · Permutation · Simulation

I. S. L. Zeng (✉) · S. R. Browning
Department of Statistics, University of Auckland, Auckland, New Zealand
e-mail: zeng@stat.auckland.ac.nz

I. S. L. Zeng · P. Gladding · R. A. H. Stewart
Green Lane Cardiovascular Service, Auckland City Hospital, Auckland, New Zealand

R. A. H. Stewart
Department of Medicine, University of Auckland, Auckland, New Zealand

M. Jüllig · M. Middleditch
School of Biological Sciences and Maurice Wilkins Centre for Biodiscovery, University of Auckland, Auckland, New Zealand

I. S. L. Zeng
Center for Clinical Research and Effective Practice, Auckland, New Zealand

Mass spectrometry and liquid chromatography are standard tools used to profile and quantify thousands of proteins simultaneously in clinical proteome research. To obtain reliable results, a high level of reproducibility in both protein identification and quantification is needed [1, 2]. Possible sources of variation may be technical or biological in origin. Technical sources of variation can alter the quantification of measured proteins due to small differences from sample preparation, chromatography, the condition of the ion source, and the

overall performance of the mass spectrometer. Biological sources of variation include differences between individuals within a population and physiological variation in individuals from one time to another. It is therefore important when evaluating clinical proteomic data to include an assessment of technical and clinical reproducibility.

Standard statistical methods used for evaluating reproducibility include the Bland Altman coefficient of reproducibility, the limit of agreement, the correlation coefficient, and linear regression. However, these assessments are generally limited to single measurements. In proteomic studies, reproducibility assessments are usually performed for a randomly selected sample of peaks or for candidate peaks of interest. The coefficient of variation, the correlation coefficient, or the limit of agreement is determined for one peptide or protein at a time. Few studies have evaluated reproducibility of mass spectral data at a multivariate level.

Some proteomic studies have borrowed statistical methods from those developed for genomic studies because of similarities in the properties of the data. In microarray reproducibility studies [3–5], correlation coefficients and autocorrelations have been used to assess the association between replicates of microarray data. The percentage of overlapping genes is used to assess the proportion of common identification between replications [5]. McShane et al. [6] introduced two global measures of reproducibility in the high-dimensional space of microarray data. They employed a robustness index (*R* index) and a discrepancy index (*D* index) to assess the reproducibility of components of interest formed by cluster analysis in replicates. The *R* index estimates the proportion of pair specimens in replicates that form the same cluster as the original data. The *D* index estimates the number of discrepancies between the original clusters and the best-matched cluster from the replicates.

Statistical methods applied to assess the reproducibility of mass spectral data have shown similarity to those used in microarray studies. In an early study [7], inter-laboratory reproducibility was assessed by four measures: (1) coefficient of variation, (2) resolution, (3) signal to noise ratio, and (4) normalized intensity for three chosen diagnostic peaks. They also assessed the classification agreement across laboratories by applying boosted logistic regression and boosted decision trees. The preprocessing of the data was standardized by a robotic system. The *m*/*z* values of peaks were controlled to within ±0.2%. The coefficient of variation (CV) for the intensity of the three peaks used in the assessment was 15–36%. Four out of the six labs obtained perfect agreement in the classification of patients and controls. The study was well designed with standardization and blind controls.

A study by Pelikan and Bigbee [8] introduced methods to assess the multivariate reproducibility of proteomic studies. This study simulated the sequential features of clinical proteomic data from multiple time intervals (sessions). The authors assessed the reproducibility of signal, discriminative features, and multivariate classification models between replicates from different sessions. They suggested a signal difference score to assess the reproducibility of profile signals. This signal difference score measures the average Euclidean distance, d_E, between all pairs of spectra, with smaller values indicating more similarity. Both the real signal (peak) and the noise were included in the measurement of similarity between spectra. Another reproducibility score used was the differential expression score, which assesses the reproducibility of discriminative features. The score quantifies the difference observed in a single profile feature between the case and control groups. It is similar to the Fisher-like score $\left|\frac{\mu^+-\mu^-}{\sigma^+-\sigma^-}\right|$, where μ and σ represent the mean and variance of the sample, while + and − represent patients and controls, respectively.

Chong et al. [9] conducted a reproducibility study of liquid chromatography tandem mass spectrometry (LC-MS/MS) iTRAQ data. In this study, the authors used three different model organisms as well as a double database search strategy, which aimed to minimize the false positive rate. They also employed multiple LC-MS/MS analyses to achieve better reproducibility. The CV was the only measure used to quantify precision. The iTRAQ quantification was highly reproducible with an average CV of 0.09 (range 0.04–0.14).

Of these proteomic reproducibility studies, only Pelikan's group introduced a global measure to assess reproducibility in mass spectral signal data. They tried to minimize the information loss by using the whole range of spectrum, but at the cost of increased noise. It is therefore difficult to distinguish poor reproducibility (real changes in the quantities of peaks) from noise. This paper proposes a permutation method to assess the global reproducibility of multiple features (proteins or peaks) in the dimension-reduced principal component space simultaneously and a discordance index based on cluster analysis methodology to summarize the bias between replicated samples.

Methods

Types of Data and Preprocessing of the Data

The format of feature quantification from different types of MS experiments can be the actual or relative intensity such as the area of peaks or other derived quantities. Most of the peak identification algorithms include baseline subtraction and normalization for preprocessing raw MS data. Normalization reduces the variation among identified proteins.

Global Reproducibility Testing

A global permutation reproducibility test based on all identified features (proteins or peaks) is proposed. This reproducibility assessment tests the hypothesis that there is no significant difference in the paired quantities of multiple features. First, the averages of all paired quantities are projected into p dimensional principal component (PC) space where p equals the number of features minus 1. Second, an eigenvalue test is used to verify how many of these p PC dimensions explain significant amounts of variance of the quantification data. The resultant m significant PC dimensions form the PC space for further analysis. Two global test statistics, the maximum t statistics and the sign score statistics, are proposed for a global permutation test in the principal component space. The empirical distributions of these two test statistics are simulated for comparison with the observed sample test statistics. This post hoc assessment can identify systematic bias between paired quantifications. Each step is described in more detail below.

Step I Principal component analysis and limit of agreement in the first principal component subspace

To begin, common features (proteins or peaks) from all individual spectra are identified for principal component analysis (PCA). A high proportion of common features identified from each run indicates good reproducibility in the identification process. Once the quantification format of data for analysis is determined, PCA is applied to the average quantities [10] of paired samples from all common features to create the orthogonal principal unit projection vectors for p PC dimensions. The resultant PCA unit projection vectors are used to project the individual runs separately back onto the PC space.

In principal component analysis, the first principal component explains the highest percentage of the variance from the data and has the highest eigenvalue. An assessment of the agreement in the first principle component provides an initial estimate of overall agreement. This can be visualized by the first principal component (FPC) plot modified from the Bland Altman plot [10]. The first principal component plot also has features similar to those of the MA plot in a microarray study. The SAS macro % FPC used for creating the FPC plot is available from the corresponding author on request.

Step II Eigenvalue testing

The proteomic profile of each sample contains proteins that are correlated and may belong to the same functional group. Principal component analysis projects these correlated data into independent PC dimensions to identify groups of proteins. While the collected data are a sample from the population of interest, the PC space formed by the principal component vectors may vary from sample to sample. In principal component analysis, a positive eigenvalue of the principal component vector reflects how much variance is explained by this component. The first principal component vector with the largest eigenvalue explains the largest percentage of the data variance, while a small positive eigenvalue could result from random noise. The eigenvalue test provides evidence of how many of the observed positive eigenvalues from the sample are not due to chance.

The eigenvalues from principal component analyses are random variables with their own distribution [11]. In a matrix with $n < p$, where n is the number of observations and p is the number of dimensions, the number of positive eigenvalues is $n - 1$. Based on random matrix theory, eigenvalue testing evaluates how many statistically significant components exist. Onatski [12] proved that the asymptotic distribution of test statistics $\mathrm{MAX}_{K0 < i < K\max}(\lambda_i - \lambda_{i+1})/(\lambda_{i+1} - \lambda_{i+2})$, where λ_i is the ith largest eigenvalue of the sample covariance matrix, equals the distribution of $\mathrm{MAX}_{0 < i < k\max - k0}(\mu_i - \mu_{i+1})/(\mu_{i+1} - \mu_{i+2})$, where $\mu_1,\ldots u_{k\max - k0}$ have the joint $(k_{\max}–k_0)$-dimensional Tracy–Widom distribution. Using this test statistic, eigenvalues are assessed from the largest to the smallest until a non-significant positive eigenvalue is identified. The m identified significant eigenvectors, λ_1 to λ_{m}, are used to derive the $n \times m$ dimension principal components for permutation. An alternative subjective approach to determine the number of significant dimension is to use scree plots [13].

Step III Permutation to test global reproducibility

The permutation method has been widely used to simulate the empirical distributions of test statistics for comparing quantities between two groups [14–16]. In the context of a proteomic reproducibility study, we propose the permutation method to test whether there are significant differences in the paired multiple-feature quantities in m significant PC dimensions. Permutation provides the empirical distributions of the global test statistics. The observed global test statistics are compared with these empirical distributions to derive the permutation p values.

We propose both a parametric and a non-parametric test statistic. The parametric statistic is the maximum t statistics ($\mathrm{Max}_{\ 1 < i < m}\ T_i$) of the m paired PC differences. Set $T_i = (\mu_i - 0)/\mathrm{std}_i$, where μ_i is the mean difference and std_i is the standard deviation of the difference in the ith PC. The non-parametric statistic is a two-dimensional sign score, $\log(P_+/P_-)$, where P_+ is the total number of positive differences in m PCs of n samples, that is, $P_+ = \sum_{j=1}^{n}\sum_{i=1}^{m} g_{ij}$, where $g_{ij}=1$ when the difference between the two replicates is positive and 0 otherwise, and P_- is the total

number of negative differences in m PCs of n samples, $P_{-} = \sum_{j=1}^{n}\sum_{i=1}^{m} f_{ij}$, where $f_{ij}=1$ when the difference between the two replicates is negative and 0 otherwise.

Let $Z_{m,n}$ represent the data matrix of paired differences of m PCs by n samples. In each Monte Carlo permutation [14], the sign of each element of the $Z_{m,n}$ matrix is independently switched with probability 0.5. Equivalently, for each i and j ($1 \le i \le m$, $1 \le j \le n$), the original and replication values are independently permuted. One thousand Monte Carlo permutations provide the empirical distributions of the two proposed global test statistics [14]. The permutation p value is the proportion of permutations in which the absolute observed test statistics is greater than the absolute permutation test statistic.

Summary Statistics for Agreement with Multiple Features

In addition to detecting bias in the reproducibility, a global index of reproducibility is found by applying cluster analysis to the data, fixing the number of clusters at the sample size. Ideally, each sample should cluster with its replicate. The discordance index measures the proportion of samples that fail to cluster with their replicates.

Results

Two different types of quantification data (SELDI-MS and LC-MS/MS with iTRAQ labeling) were used to demonstrate the proposed method. In the SELDI-MS experiment, common peaks were identified with the PROcess algorithm (Li, Xiaochun http://bioconductor.org/packages/2.4/bioc/html/PROcess.html) where the local maxima of intensities in each identified peak region were used as the analyzed quantity. In the LC-MS/MS labeling experiment, peptides identified by ProteinPilot™ with "used" indicator = 1 were filtered by confidence score and aligned across different runs. For the purpose of this reproducibility analysis, the weighted average of reporter ion peak areas was calculated for peptides that have more than one observation in a single protein summary. The resultant peptide areas are summed for each protein that they belonged to. Within each run, median normalization was applied to the summed areas across labels on the natural log scale. After preprocessing, a relative protein quantity was derived for each sample. This preprocessing reduces the variance across different proteins and corrects for labeling effects.

Case Study

Coronary plasma blood samples of eight ischemic patients before and after an angioplastic procedure were collected from the Green Lane Cardiovascular Service of Auckland City Hospital and analyzed by LC-MS/MS with iTRAQ labeling at the Centre for Genomics and Proteomics, University of Auckland. Prior to the LC-MS/MS analysis, a depletion process was used to exclude the twelve most highly abundant proteins. For the purposes of this replication study, it is hypothesized that there are no changes in the proteomic expression before and after the angioplasty procedure, so the post-procedure samples are treated as the replications of the baseline sample to demonstrate the reproducibility assessment. Peptide profiles from four different runs of ProteinPilot™ were aligned, and the areas under the peaks were log-transformed and normalized by the median within each run. Two hundred common peptides from the four different runs were used to construct the relative intensity of proteins for the reproducibility assessment. Principal component analysis was performed on the quantities of the 28 proteins found in all four runs.

The eigenvalue test and scree plot (results not shown) indicated that the first two eigenvalues was significant, and the corresponding eigenvectors explained 95% of the total variance, with 82% explained by the first principal component. The FPC plot (Fig. 1a) shows a significant difference in the relative protein quantities between the post- and pre-angioplasty samples; the PC of post-procedure samples tends to be lower than the pre-angioplasty samples overall. This trend is consistent with the pattern in the second plot where the difference in relative quantity between the pre- and post-procedure for all proteins is plotted. Details of the post-angioplasty expression change will be reported separately.

Simulation Experiments

A simulation experiment was used to investigate the sensitivity of the proposed method. Different types of noise, with different distributions and parameters, were added to the relative peak quantities of 30 ovarian cancer patients to simulate different replicates from MS experiment. The mass spectral intensity data are a random sample from a large ovarian proteomic experiment, downloaded from the proteomic databank of Center for Cancer Research (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). Sixty-one common peaks were identified from these 30 subsamples using the PROcess bioconductor package; preprocessing procedures including baseline subtraction by Loess and normalization were applied. A relative quantity was derived for each identified peak.

Twelve distributions, with different parameters simulating systematic bias (parameterized by the mean, μ) and noise (parameterized by the standard deviation, σ), were generated and added into the relative quantity

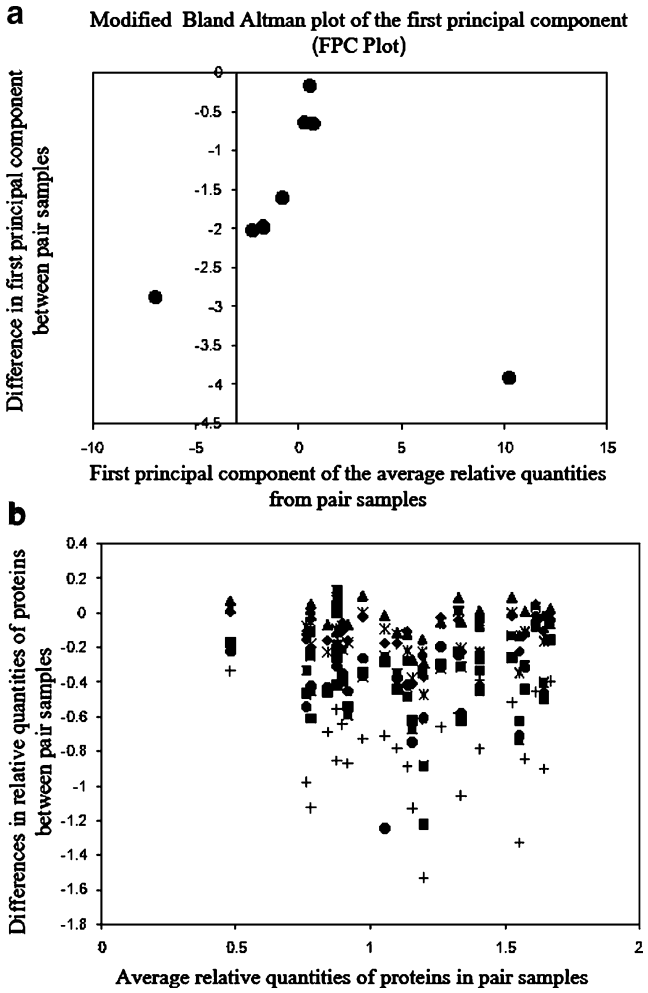

Fig. 1 **a** PC plot of first dimension (*FPC plot*) from cardiac LC-MS/MS iTRAQ data. **b** Difference in log(area) between replicates from all proteins vs. average of all log(area) from cardiac patient's LC-MS/MS iTRAQ data. A *unique symbol* is used for each patient

data. The distributions were normal distributions $\{(\mu = 0/2/4, \sigma = 2/4), (\mu = 0/2/4, \sigma = 2/4)\}$, exponential distributions $\{(\mu = \sigma = 0.5), (\mu = \sigma = 1)\}$, and bimodal distributions (mixture of two normal distributions with different means and standard deviations at different sections of m/z values). The FPC plots from two simulations are shown in Fig. 2. In the FPC plot of the normally distributed differences with (μ=0, σ=4), the differences in the first principal component scores between the sample and its simulated replicate are randomly scattered above or below 0. In the FPC plot of the exponentially distributed differences with $\mu = \sigma$=0.5, the differences in the first principal component scores tend to be significantly below 0.

The eigenvalues of the sample matrices were tested before the permutation test proceeded. In the principal component analysis, quantities were automatically normalized. The permutation tests were applied to differing sizes of samples (8, 15, and 30 samples) and to different distributions with different parameters in the simulated replicates. The comparison results are shown in Table 1.

Results of Global Permutation Reproducibility Testing

Sample Size and Sensitivity of the Test When the sample size was equal to eight, the permutation tests using maximum t statistics failed to detect the simulated bias; the permutation tests using the sign score statistics successfully identified bias for the normal distribution (μ=2, σ=2) and exponential distribution ($\mu = \sigma$=1 and $\mu = \sigma$=0.5), but failed for the other simulated distributions (Table 1).

When the sample size was equal to 15, the permutation tests using maximum t statistics successfully detected the bias with the normal distribution (μ=2, σ=2), the exponential distribution with $\lambda = 1(\mu = \sigma = 1)$ and $\lambda = 2(\mu = \sigma = 0.5)$, and the bimodal distribution $(\mu = 1, \sigma = 1(m/z < 1,000)\mu = 2, \sigma = 2(m/z > 1,000))$. However, it failed to detect the bias with the normal distribution (μ=2, σ=4), the bimodal distribution $(\mu = 1, \sigma = 2(m/z < 1,000)\mu = 2, \sigma = 4(m/z > 1,000))$,

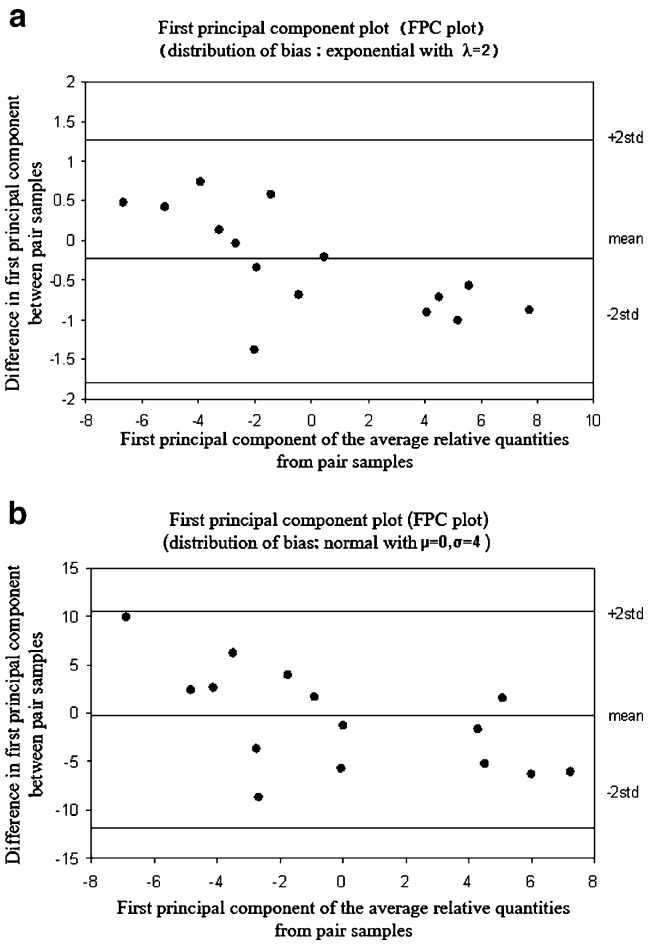

Fig. 2 **a**, **b** FPC plots of two different simulated systematic biases

Table 1 Simulated replicates having added bias and noise of different distributions

Data presented are median(P25,P75)		Parametric version		Non-parametric version	
Distribution of the systematic bias	Size of the samples and its replicates	Distribution of the permutated p values	Distribution of the test statistics	Distribution of the permutated p values	Distribution of the test statistics
Normal					
μ=2 σ=4	30	0.02 (0.01,0.06)	3.76 (3.42,4.19)	–	–
μ=0 σ=4	15	0.51 (0.30, 0.67)	2.18 (1.95, 2.57)	0.59 (0.41, 0.82)	0.00 (−0.08, 0.08)
μ=2 σ=2		0.02 (0.009, 0.04)	4.01 (3.63, 4.65)	0.001 (0.001, 0.001)	−0.84 (−0.96 ,−0.71)
μ=2 σ=4		0.15 (0.06, 0.34)	2.94 (2.50, 3.47)	0.005 (0.001, 0.03)	−0.42 (−0.48, −0.32)
μ=0 σ=4	8	0.36 (0.13, 0.57)	2.34 (1.92, 3.17)	0.64 (0.42, 0.90)	0.00 (−0.14,0.14)
μ=2 σ=2		0.18 (0.10, 0.31)	2.97 (2.42, 3.54)	0.02 (0.004, 0.08)	−0.65 (−0.81, −0.50)
μ=2 σ=4		0.32 (0.19,0.53)	2.50 (1.98, 3.01)	0.50 (0.24, 0.64)	−0.21 (−0.35, −0.07)
Exponential					
$\lambda = 1(\mu = \sigma = 1)$	15	0.01 (0.005, 0.03)	4.48 (3.88, 4.98)	0.001 (0.001,0.001)	−1.03 (−1.14, −0.89)
$\lambda = 2(\mu = \sigma = 0.5)$		0.007 (0.002,0.02)	4.70 (4.24, 5.37)	0.001 (0.001, 0.001)	−1.17 (−1.29, −1.01)
$\lambda = 1(\mu = \sigma = 1)$	8	0.12 (0.05, 0.19)	3.21 (2.86, 4.22)	0.01 (0.001, 0.03)	−0.73 (−1.01, −0.63)
$\lambda = 2(\mu = \sigma = 0.5)$		0.07 (0.04, 0.13)	3.61 (3.16, 4.56)	0.002 (0.001, 0.005)	−0.98 (−1.17, −0.81)
Bimodal					
$\mu = 1, \sigma = 2(m/z \leq 1,000)\mu = 2, \sigma = 4(m/z > 1,000)$	30	0.03 (0.008, 0.08)	3.68 (3.33, 4.27)	–	–
$\mu = 2, \sigma = 4(m/z \leq 1,000)\mu = 4, \sigma = 8(m/z > 1,000)$		0.03 (0.01, 0.08)	3.68 (3.28, 4.02)	–	–
$\mu = 1, \sigma = 1(m/z \leq 1,000)\mu = 2, \sigma = 2(m/z > 1000)$	15	0.03 (0.01, 0.08)	3.68 (3.28, 4.02)	0.001 (0.001, 0.001)	−0.96 (−1.08, −0.80)
$\mu = 1, \sigma = 2(m/z \leq 1,000)\mu = 2, \sigma = 4(m/z > 1,000)$		0.12 (0.06, 0.32)	3.07 (2.55, 3.51)	0.004 (0.001, 0.01)	−0.42 (−0.52, −0.34)
$\mu = 2, \sigma = 4(m/z \leq 1000)\mu = 4, \sigma = 8(m/z > 1,000)$		0.14 (0.06, 0.29)	3.00 (2.61, 3.48)	0.005 (0.001, 0.03)	−0.40 (−0.50, −0.31)
$\mu = 1, \sigma = 1(m/z \leq 1,000)\mu = 2, \sigma = 2(m/z > 1,000)$	8	0.18 (0.08, 0.37)	2.88 (2.39, 3.56)	0.02 (0.003, 0.14)	−0.65 (−0.81, −0.43)
$\mu = 1, \sigma = 2(m/z \leq 1,000)\mu = 2, \sigma = 4(m/z > 1,000)$		0.26 (0.14, 0.55)	2.67 (1.95, 3.21)	0.48 (0.23, 0.65)	−0.21 (−0.35, −0.07)
$\mu = 2, \sigma = 4(m/z \leq 1,000)\mu = 4, \sigma = 8(m/z > 1,000)$		0.36 (0.19, 0.54)	2.32 (2.03, 2.87)	0.32 (0.23, 0.62)	−0.28 (−0.35, −0.14)

Table 2 Summary of discordance index and median percent change from simulated bias

Distribution of bias with different parameter	Number of samples–replicates grouped in the same cluster	Discordance index (% of samples–replicates that failed to group in the same cluster)	Median % change across all features
Normal (n=30)			
μ=0 σ=4	27	0.10	0.4% [−24%, 21%]
μ=2 σ=2	27	0.10	67% [16%,288%]
μ=2 σ=4	23	0.23	78.2% [17.1%, 313.2%]
Exponential (n=30)			
$\lambda = 1(\mu = \sigma = 1)$	30	0.0	43% [8.6%, 137%]
$\lambda = 2(\mu = \sigma = 0.5)$	30	0.0	21% [4.3%, 68%]
Bimodal (n=30)			
$\mu = 1, \sigma = 2(m/z \leq 1,000)\mu = 2, \sigma = 4(m/z > 1,000)$	26	0.13	65%[8.8%,191%]
$\mu = 2, \sigma = 4(m/z \leq 1,000)\mu = 4, \sigma = 8(m/z > 1,000)$	10	0.67	130% [18%, 381%]

and the bimodal distribution $(\mu = 2, \sigma = 4(m/z < 1,000)\mu = 4, \sigma = 8(m/z > 1,000))$. The tests using the sign score statistics successfully detected all of the simulated bias.

When the sample size was equal to 30, both test statistics successfully identified all the simulated bias. The sensitivity of the reproducibility is affected by the sample size.

Variance and Sensitivity When the variation of the sample increased, the sensitivities of both test statistics were weakened. In the simulations, when the coefficient of variation of a normally distributed difference was greater than 1, the permutation test using maximum t statistics was not sensitive with sample sizes <30.

Discordance Index and Median Percentage Change In the bias assessment, all 30 samples, including both replicates, were entered in a cluster analysis, and 30 clusters were formed by the Ward method [13]. Table 2 summarizes sample details and grouping of replicates in the same cluster.

A high discordance index can be caused by a high degree of bias with high variation. The simulation results show that the discordance index is not sensitive to bias with small magnitude and large variation. However, the discordance index is an interesting way to summarize the data and provides extra information about outlying samples.

Discussion

This paper proposes a method of assessing the global reproducibility of mass spectral data rather than focusing on the reproducibility of single selected candidate proteins or peptides. A multivariate reproducibility assessment is useful to assess overall performance and identify problematic candidate proteins or peaks. Using principal component analysis, high dimensional correlated spectral data are reduced to lower dimensions and projected into orthogonal space. Random matrix theory provides a basis for testing the underlying correlation pattern of proteins to eliminate non-significant principal components from further analysis. A permutation reproducibility test can be used to identify systematic bias and adjust for multiple testing. If bias is identified, further analysis of the principal components can identify problematic proteins or peaks using maximum t test statistics or sign score statistics. The strategy of combining dimension reduction with permutation testing utilizes all the information effectively.

From the simulation experiments, it was found that a sample size of 30 will have greater statistical power to detect simulated bias than a sample size of 15 or 8. The size and variation of samples have significant impacts on the sensitivity of the assessment.

A large-scale reproducibility study using LC-MS/MS that assesses the real day-to-day operation and patient variation is needed. This study would be important before applying the proteomic technology in daily clinical laboratory practice. The reproducibility assessment in a clinical proteomic experiment is complex. It involves early phase assessment for reproducibility of laboratory technique and the late clinical phase assessment for reproducibility of patients' day-to-day physiological conditions. For the examples used in this study, the reproducibility of quantification post-protein identification was assessed. However, the proposed method can be applied to specific sources of variation including intra/inter-run reproducibility and day-to-day variability.

A limitation of the current study is that the sensitivity of eigenvalue testing is affected by the sample size. When the sample size is small, the eigenvalue test combined with the

traditional scree plot may be a better way to identify the main pattern of protein profiles.

In conclusion, this paper suggests extensions of reproducibility methods from the single-dimension assessment to a higher dimension assessment and demonstrates that this systematic approach to reproducibility is useful and workable.

Acknowledgments The first author benefited from useful discussions with Dr. Kathy Ruggerio. The authors also benefited from the open access to the proteomic data of National Cancer Institute. Dr. Patrick Gladding, Ms. Irene Zeng, and Dr. Ralph Stewart received from Green Lane Research and Education Trust funding for the cardiac ischemic clinical proteomic study.

References

1. Hale JE, Gelfanova V, Ludwig JR, Knierman MD. Application of proteomics for discovery of protein biomarkers. Briefings in Functional Genomics and Proteomics. 2003;2(3):185–93.
2. Mcguire NJ, Overgaard J, Pociot F. Mass spectrometry is only one piece of the puzzle in clinical proteomics. Briefings in Functional Genomics and Proteomics. 2008;7(1):74–83.
3. Lyne R, Burns G, Mata J, et al. Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. BMC Genomics. 2003;4:27.
4. Tan PK, Downey TJ, Spitznagel EL Jr, et al. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 2003;31(19):5676–84.
5. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC Bioinformatics. 2007;8 (Suppl. 9):S20.
6. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering pattens observed in analyses of microarray data. Bioinformatics. 2002;18 (11):1462–9.
7. Semmes O, Feng Z, Adam BL, Banez LL, Bigbee WL, Campos D, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. Clin Chem. 2005;51(1):102–12.
8. Pelikan R, Bigbee WL. Intersession reproducibility of mass spectrometry profiles and its effect on accuracy of multivariate classification models. Bioinformatics. 2007;23(22):3065.
9. Chong PK, Gan CS, Pham TK, Wright PC. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: implication of multiple injections. Journal of Proteome Research. 2006;5(5):1232–40.
10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.
11. Bejan AI. Largest eigenvalues and sample covariance matrices. Tracy–Widom and Painleve II: computational aspects and realization in S-Plus with applications. Mathematics Subject Classification. 2005;1991.
12. Onatski A. The Tracy–Widom limit for the largest eigenvalues of singular complex Wishart matrices. Ann Appl Probab. 2008;18 (2):470–90.
13. Rencher AC. Methods of multivariate analysis. New York: Wiley; 2002.
14. Good P. Permutation, parametric, and bootstrap tests of hypotheses: a practical guide to resampling methods for testing hypotheses. New York: Springer; 2005.
15. Wheldon MC, Anderson MJ, Johnson BW. Identifying treatment effects in multi-channel measurements in electroencephalographic studies: multivariate permutations tests and multiple comparisons. Aust N Z J Stat. 2007;49(4):397–413.
16. Neubert K, Brunner E. A studentized permutation test for the nonparametric Behrens–Fisher problem. Comput Stat Data Anal. 2007;51(10):5192–204.