Normalization Regarding Non-Random Missing Values in High-Throughput Mass Spectrometry Data

Pei Wang, Hua Tang, Heidi Zhang, Jeffrey Whiteaker, Amanda G. Paulovich, and Martin Mcintosh

Pacific Symposium on Biocomputing 11:315-326(2006)

NORMALIZATION REGARDING NON-RANDOM MISSING VALUES IN HIGH-THROUGHPUT MASS SPECTROMETRY DATA

PEI WANG[§], HUA TANG, HEIDI ZHANG, JEFFREY WHITEAKER,
AMANDA G PAULOVICH, MARTIN MCINTOSH

Fred Hutchinson Cancer Research Center,
Seattle, WA, 98109
[§]*E-mail: pwang@fhcrc.org*

We propose a two-step normalization procedure for high-throughput mass spectrometry (MS) data, which is a necessary step in biomarker clustering or classification. First, a global normalization step is used to remove sources of systematic variation between MS profiles due to, for instance, varying amounts of sample degradation over time. A probability model is then used to investigate the intensity-dependent missing events and provides possible substitutions for the missing values. We illustrate the performance of the method with a LC-MS data set of synthetic protein mixtures.

1. Introduction

High-throughput mass spectrometry (MS) technology offers a powerful means of analyzing biological samples. The ability of MS to identify and precisely quantify thousands of proteins from complex samples is expected to broadly affect biology and medicine[3]. However, MS systems are subject to considerable noise and variability that is not fully characterized or accounted for. Thus, it is important and necessary to properly conduct data-preprocessing steps such as signal filtering, peak detection, alignment in time (and mass charge ratio), and amplitude normalization before reliable conclusions can be made from the data[1].

In this paper, we focus on the normalization step, and propose a probability model for intensity-dependent missing events in MS-based data sets. In MS experiments, the instrument may have trouble detecting the weak signals of low-abundance peptides. Even if the instrument detects the signal, the peak intensities may be too low to be distinguished from background noise during data processing. Therefore, the lower the ion abundance, the

more likely the peptide will be "missing" in the MS output data. Ignoring such non-random missing pattern may introduce significant bias into subsequent analyses. In this paper, we propose a novel probability model to describe the missing behavior, which accounts for this type of intensity-dependent missing events.

The rest of the paper is organized as follows: Section 2 provides a brief description of a data example illustrating the problem. Section 3 introduces a global normalization step, which adjusts systematic trends. The missing model, which represents our major contribution, is described in Section 4. Section 5 applies the proposed methods to an example data set and Section 6 is the conclusion.

2. Experiment and Data

In this section, we describe an experiment, in which replicates of two protein mixtures were analyzed on three consecutive days. We find that the samples processed in later days experienced higher levels of protein degradation due to, for instance, longer storage time as well as more freeze-thaw cycles[2]. Such variations are often unavoidable in real disease studies involving human samples.

2.1. *Sample preparation*

Two mixtures of proteins were assembled as part of an exploratory study to understand the performance of our MS instrument. One mixture (denoted as A) consisted of four proteins: bovine albumin, bovine transferrin, bovine alpha lactalbumin and bovine catalase. The other mixture (denoted as B) consisted of the same four proteins plus bovine beta lactoglobulin (proteins were selected based on their length and abilities to produce tryptic peptides). All five recombinant proteins were purified with reversed-phase high performance liquid chromatography (VisionWorkstation Applied Biosystems, Framingham, MA, USA). The collected protein fractions were dried in SpeedVac (Thermo Savant, San Jose, CA, USA). The purified proteins were denatured individually with 60% MeOH, reduced with 10 mM DTT at 60^oC for 1 hr, and alkylated with 50 mM iodoacetamide in the dark at room temperature for 30 min. The polypeptides were trypsinized for 6 hr at 37^oC with a protein/enzyme of 50/1.

2.2. *LC-MS system*

The LC-MS system comprised an 1100 Series Nanoflow LC system (Agilent Technologies, Palo Alto, CA, USA), a binary capillary pump, a C18 Symmetry NanoEase trapping column (Waters Corporation, Milford, MA, USA), a C18 PepMap nano LC column(LC Packings, Sunnyvale, CA, USA), and an LCT Premier time-of-flight mass spectrometer (Waters Corporation). The flow rates are 20 uL/min in the trapping column, and 400 nL/min in the LC column. The solvents were A(0.1% formic acid in water) and B (0.1% formic acid in acetonitrile). Linear gradient elution was applied from 0 to 40% B in 30 min. Mass spectra were acquired every 1.0 s with a 0.1 s interscan delay time. The instrument was mass-calibrated with a sodium formate solution prior to analysis.

2.3. *Data and problem*

The raw data is first processed using a program developed in our group,*msInspect*.[a], which includes modules for detecting and aligning peptide features. The output peptide array reports the intensities of all peptide features in each sample (an LC-MS experiment). Denote the intensity of the ith feature in the kth sample as y_i^k. If the ith feature is detected in the kth sample, then y_i^k is set to 0.

The total number of non-zero intensity peptides in each sample is summarized in Table 1. Clearly, more features were detected in experiments

Table 1. Number of peptide features in each sample. Mixture A consists of four proteins, while mixture B consists of five proteins.

Day 1	Sample Index	A1	B2	B3	A4		
	Feature Number	660	648	789	495		
Day 2	Sample Index	B5	B6	A7	B8	A9	A10
	Feature Number	609	339	386	492	384	413
Day 3	Sample Index	A11	B12	B13	A14		
	Feature Number	237	302	406	178		

Note: In the sample indexes, A=4 protein mixtures, B=five protein mixtures, and the number indicators the experimental order.

performed on day 1 than those performed on day 3.

If we further compare Sample A1 (with 660 features) and Sample A14

[a]Available at http://proteomics.fhcrc.org/CPL/home.html

(with 178 features), as illustrated in Figure 1, we see there is an overall decrease in intensity in sample A14 compared to sample A1.

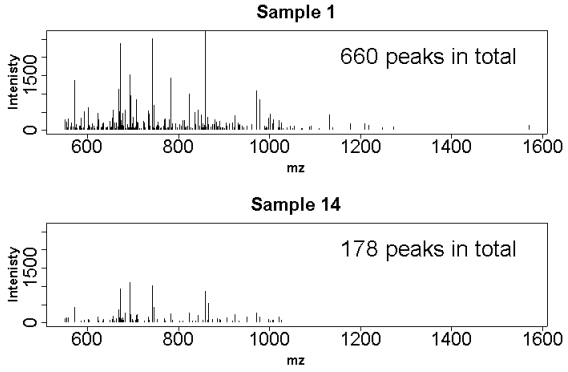

Figure 1. **Compare Sample A1 and Sample A14.** The two plots compare the intensities of all features in the two samples. The x coordinate is the mass charge ratio (mz), and the y coordinate is the intensity value. Each feature is represented by a vertical line at its mz position, with the length of the line equal to its intensity.

Given this kind of variation, it is crucial to normalize intensities before different samples can be properly compared.

3. Global Normalization

By globally normalizing signal intensities across multiple samples, we aim to identify and remove systematic variation arising because of differential amounts of sample loaded into the LC-MS system, protein degradation over time, or variation in the sensitivity of the instrument detector.

It is natural to assume that the sample intensities are all related by a constant factor[5]. A common choice for this re-scaling coefficient is the sample mean or median. This choice is based on the assumption that the number of features whose measurements change is few compared to the total number of features. So the distribution of the measurements of all the features should be roughly the same across different experimental runs[4].

However, in MS experiments, because of the limitation of detector sensitivity and the unavoidable instrument noise, ions below a certain intensity level may hardly be detected, which leads to non-random missing of peptide features in the result. Thus, it is not appropriate to use overall mean

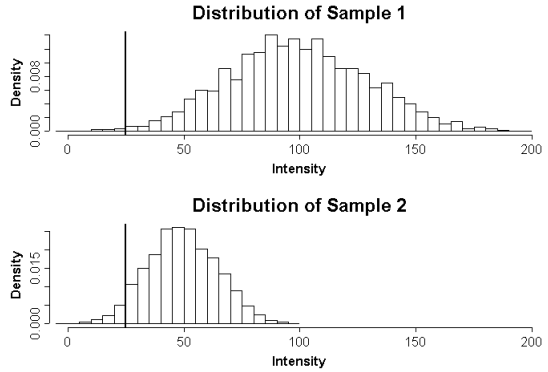

Figure 2. **The effect of non-random missing.** Suppose the overall peptide abundance of sample 1 are twice as great as the overall peptide abundance of sample 2. The histograms shows the true intensity distribution of all peptides in sample 1 and sample 2 respectively. The minimal detection level of the instrument is represented by the vertical line (features on the left side of the line can not be observed in the experiment). Using the mean or median intensity of the observed features in each sample leads to biased estimate of the scaling coefficients.

or median for re-scaling. This is illustrated in Figure 2. In order to avoid the possible bias due to non-random missing events, we propose to use the top L ordered statistics of feature intensities in each sample, where L is a parameter chosen by users.

For the simple case of two samples, denote the intensity measurements of one sample as $X = (x_1, x_2, ..., x_n)$ and of the other sample as $Y = (y_1, y_2, ..., y_m)$, whose order statistic can be represented as $x_{(1)} > x_{(2)} > ... > x_{(n)}$ and $y_{(1)} > y_{(2)} > ... > y_{(m)}$ respectively. Then, for a chosen number $L(L < \min(n, m))$, the scaling coefficient of X versus Y can be estimated as $\lambda = \sum_{i=1}^{L} x_{(i)} / \sum_{j=1}^{L} y_{(j)}$ or more robustly,

$$\lambda = \text{median}(x_{(1)}, ..., x_{(L)}) / \text{median}(y_{(1)}, ..., y_{(L)}). \quad (1)$$

For the case of $K(K > 2)$ samples, denote the intensity measurements of the kth sample as $X^k = (x_1^k, x_2^k, ..., x_{n_k}^k)$. For a given number $L(L < \min(\{n_k\}_{k=1}^K))$, define the population median as

$$\mu_0 = \frac{1}{K} \sum_k \text{median}(x_{(1)}^k, x_{(2)}^k, ..., x_{(L)}^k).$$

Then the scaling coefficient for the kth sample is

$$\lambda^k = \frac{1}{\mu_0} \text{median}(x_{(1)}^k, x_{(2)}^k, ..., x_{(L)}^k) \quad (2)$$

4. Model of Missing Events

We can make inferences on the missing events of one sample based on the information from other samples. The idea is illustrated in Figure 3. Suppose Sample 1 and Sample 2 are identical mixtures, but due to experimental factors, the overall peptide abundance of Sample 1 is smaller than the overall peptide abundance of Sample 2. Peptide 1 cannot be observed in Sample 1 because its intensity falls below the minimum detectable level. However, based on the intensities of those peptides observed in both samples (*e.g.* Peptide 2), the scale difference of the overall abundances between Sample 1 and Sample 2 can be estimated. Therefore, the "missing" intensity of Peptide 1 in Sample 1 can be reasonably approximated with the intensity measured in Sample 2 divided by a scale coefficient.

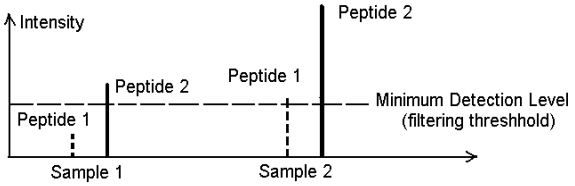

Figure 3. **Missing model.** The heights of the vertical bars indicate the true intensities of different peptides in the two samples. The dashed vertical lines represent Peptide 1, while the solid vertical lines represent Peptide 2. The horizontal dashed line indicates the minimal detection level of the instrument.

More general, we use a probability model to describe such missing events, which is described in below.

4.1. *Probability Model*

In one sample, we introduce a latent variable z_i for the ith peptide, which indicates whether this peptide exists in the sample or not:

$$z_i = \begin{cases} 1, & \text{if } i\text{th peptide exists in the sample;} \\ 0, & \text{if } i\text{th peptide does not exist in the sample.} \end{cases} \tag{3}$$

Given $z_i = 1$ (the ith peptide exists in the sample), the abundance of this peptide x_i can be deemed as a random variable:

$$x_i \begin{cases} = 0, & \text{if } z_i = 0, \\ \sim f_i, & \text{if } z_i = 1, \end{cases} \tag{4}$$

where f_i is the density function of some probability distribution. It is reasonable to assume that z_i and $x_i|z_i = 1$ are independent with each other.

Suppose the minimum detectable level of the instrument is d. Then, given the value (x_i, z_i, d), the observed abundance y_i of this peptide satisfies

$$y_i|(x_i, z_i, d) = \begin{cases} 0, & \text{if } z_i = 0; \\ 0, & \text{if } z_i = 1 \text{ and } x_i < d; \\ x_i, & \text{if } z_i = 1 \text{ and } x_i \geq d. \end{cases} \tag{5}$$

We say that a missing event happens to the ith peptide if the ith peptide exists in the sample but no signal has been detected (denoted as $M_i = \{z_i = 1, y_i = 0\}$). We are interested in the probability of missing event when no signal is observed, *i.e.* $P(M_i|y_i = 0)$, which can be calculated as follows:

$$\begin{aligned} P_d((z_i = 1, y_i = 0)|y_i = 0) &= \frac{P_d(z_i=1, y_i=0)}{P_d(y_i=0)} \\ &= \frac{P_d(y_i=0|z_i=1)P(z_i=1)}{P_d(y_i=0|z_i=1)P(z_i=1)+P_d(y_i=0|z_i=0)P(z_i=0)} \\ &= \frac{P_d(x_i<d|z_i=1)P(z_i=1)}{P_d(x_i<d|z_i=1)P(z_i=1)+P(z_i=0)}, \end{aligned} \tag{6}$$

where

$$P_d(y_i = 0|z_i = 1) = P_d(x_i < d, z_i = 1|z_i = 1) = P_d(x_i < d|z_i = 1)$$

and

$$P_d(y_i = 0|z_i = 0) = P_d(z_i = 0|z_i = 0) = 1$$

comes from Equation (5); $P(z_i)$ does not depend on d.

In addition, if $P(x_i > d|z_i = 1) > 0$, we have

$$P(z_i = 1) = \frac{P_d(z_i = 1, x_i > d)}{P_d(x_i > d|z_i = 1)} = \frac{P_d(y_i > 0)}{P_d(x_i > d|z_i = 1)}. \tag{7}$$

Therefore, given the detectable level parameter d, the distribution function f_i, and the observed abundance y_i, we can estimate the probability $P(M_i|y_i = 0)$ with Equation (6) and (7).

Moreover, a natural choice for imputing the intensity of a missing peak is $E(x_i|y_i = 0)$, which can be calculate as

$$\begin{aligned} E(x_i|y_i = 0) &= E(x_i|y_i = 0, z_i = 1)P(z_i = 1|y_i = 0) \\ &\quad + E(x_i|y_i = 0, z_i = 0)P(z_i = 0|y_i = 0) \\ &= E(x_i|x_i < d, z_i = 1)P(z_i = 1|y_i = 0) + 0 \\ &= E(x_i|x_i < d, z_i = 1)P(M_i|y_i = 0). \end{aligned} \tag{8}$$

Note $E(x_i|x_i < d, z_i = 1)$ only depends on the detector level parameter d and the distribution function f_i.

4.2. *Model Fitting*

4.2.1. *Detectable level d*

A reasonable estimate of the parameter, d, is the background noise level in each MS profiles, since those peaks with height below this value can not be confidently distinguished from noise signals. For a set of profiles from the same instrument, we assume that the same detectable level. Hence, we estimate d using all raw profiles. After the global normalization described in section (3), the detectable level of the kth profile becomes $\widetilde{d^k} = \frac{\hat{d}}{\lambda^k}$, where λ^k is the normalization scale coefficient in Equation (2).

4.2.2. *Abundance distribution f_x*

A. When Biological Replicates Available

For K replicates of the same biology samples, we assume that $\{\frac{x_i^k}{\lambda^k}|z_i^k = 1\}_{k=1}^K$ are independently identically distributed as $N(\mu_i, \sigma_i^2)$ for some parameter μ_i and σ_i, where x_i^k is the true abundance of the ith peptide in the kth profile.

Since $x_i^k|z_i^k = 1$ and z_i^k are independent from each other, it is easy to see that $y_i^k|(y_i^k > 0)$ and $x_i^k|(x_i^k > d, z_i^k = 1)$ are equal in distribution. Thus,

$$\begin{aligned} \frac{y_i^k}{\lambda^k}|(y_i^k > 0) \sim \widetilde{f_i^k}(t) &= \frac{P(x_i^k/\lambda^k \in dt, x_i^k > d, z_i = 1)}{P(x_i^k > d, z_i = 1)} \\ &= \frac{\varphi_{\mu_i,\sigma_i}(t)}{P(x_i > d|z_i = 1)}, \text{ for } t > \widetilde{d^k}. \end{aligned} \tag{9}$$

where φ_{μ_i,σ_i} is the density function of $N(\mu_i, \sigma_i^2)$.

For the simple case where $\sigma_i \ll \left|\widetilde{d^k} - \mu_i\right|$, we can approximate $P(x_i^k > d|z_i^k = 1)$ with $I\left(\widetilde{d^k} < \mu_i\right)$. It follows

$$\widetilde{f_i^k} \approx \varphi_{\mu_i,\sigma_i}, \text{ when } \widetilde{d^k} < \mu_i. \tag{10}$$

Thus, the mean intensity of the ith peptide can be estimated as the average of the observed signals:

$$\widehat{\mu}_i = \frac{\sum_k y_i^k/\lambda^k}{\sum_k I(y_i^k > 0)}. \tag{11}$$

Together with Eq.(7), we have

$$\widehat{P}(z_i = 1) = \frac{\sum_k I(y_i^k > 0)}{\sum_k I(\hat{\mu}_i > \widetilde{d^k})}. \tag{12}$$

Therefore

$$\widehat{P}(M_i^k|y_i^k=0)=\begin{cases}\hat{P}(z_i=1), & \text{if } \hat{\mu}_i<\widetilde{d^k},\\ 0, & \text{if } \hat{\mu}_i>\widetilde{d^k}.\end{cases} \tag{13}$$

And then,

$$\widehat{E}(x_i^k|y_i^k=0)=\begin{cases}\hat{\mu}_i\hat{P}(z_i=1), & \text{if } \hat{\mu}_i<\widetilde{d^k},\\ 0, & \text{if } \hat{\mu}_i>\widetilde{d^k}.\end{cases} \tag{14}$$

with Eq.(8), Eq.(11) and Eq.(12).

B. When Biological Replicates Not Available

Because the biological samples are limited, a large number of MS replicates are not always available for each sample. In such cases, a natural solution is to use the nearest K "neighbor samples" as pseudo replicates to fit the missing model. Here nearest K "neighbors" refers to the K closest profiles to the target profile under certain distance metrics (*i.e.* L_2 norm). However, if the missing rate is relatively high, the distance measured with the raw data could be misleading. Thus, we propose the following iteration procedure to try to recover the true "neighborhood" structure:

(1) Begin with K=N, where N is the total number of samples. Denote the original peptide array data matrix as Pep^0.
(2) (a) Based on Pep^{N-K}, calculate the distance between each two samples.
(b) For each sample, estimate the missing features by using its nearest K neighbors. Denote the new peptide arrays as Pep^{N-K+1}.
(c) K=K-1.
(3) Repeat step 2 until $K=K_0$, where K_0 is a pre-selected number.

If we aim to separate the samples into two clusters, a possible choice for K_0 is $N/2$.

5. Result

5.1. *Global normalization*

The scale coefficients of global normalization are estimated with the top 80 order statistics of each sample according to Eq.(2). Fig.4 shows the relationships between the top 80 order statistics of Samples $11-14$ (the four samples on the third day) and the top 80 order statistics of Sample 1.

Table3 shows the scale coefficients for the four pairs of samples in Fig.4. Compared to the estimators derived with the order statistics, the estimators

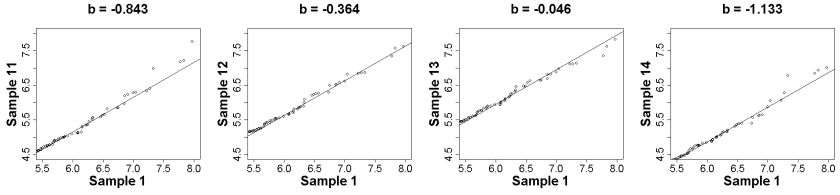

Figure 4. **The top 80 order statistics of Sample** $11 - 14$ ***v.s.*** **Sample 1.** The y and x coordinates represent the log intensities of the 80 most abundant features in the corresponding samples. The good linear relationship with slope= 1 justifies the assumption that the sample intensities are related by a constant factor (model $x = \lambda y$ is equivalent to model $log(x) = log(y) + b$, where b and λ are parameters).

derived with overall medians dramatically overestimate the scale change between these sample pairs. This demonstrates the necessariness of using the top order statistics to conduct the global normalization when non-random missing is a concern in the study.

Table 2. Scale Coefficients *v.s.* Sample 1.

Sample Index	11	12	13	14
λ (based on the order statistics)	0.43	0.69	0.95	0.32
λ (based on overall median)	1.19	1.10	0.99	1.09

5.2. *Study of Missing Events*

We consider 12 of the 14 samples whose non-zero features are at least 10% of the total.

5.2.1. *Supervised analysis*

Treating all 4-protein samples as replicas and all 5-protein samples as replicas, using Equation (13) we can estimate the total number of possibly missing features $\sum_i I(\widehat{P}(M_i^k|y_i^k = 0) > 0)$ for each sample. The result is shown in Table 3. Again, we can see that the missing trend is more severe in some samples than in others. Ignoring such trend may bring unexpected bias into downstream analysis.

Table 3. Number of peptide features in each sample.

Sample	A1	B2	B3	A4	B5	B6	A7	B8	A9	A10	B12	B13
Missing	8	0	0	8	0	187	118	63	116	69	116	17

5.2.2. *Unsupervised analysis*

The goal here is to use the MS profiles to recover the 4-protein and 5-protein group labels for each sample. First, based on the data after global normalization, we perform hierarchical clustering analysis using the R[b] function *hclust* with complete linkage. The dendrogram is illustrated in the top plot of Figure 5. The two main sub-clusters are separated according to when the MS experiments were conducted (the first four samples were processed on day 1 while the others on the day 2 and 3).

Next, we use the iterative procedure described in section 4.2.2 to substitute the possible missing measurements with their expected values, and perform the hierarchical clustering on the resulting data. The new dendrogram is illustrated in the bottom plot of Figure 5, in which the 4-protein samples and the 5-protein samples are correctly clustered into two groups. This suggests that properly modelling the missing events would prevent the analysis from being driven by experimental variation rather than biological variation.

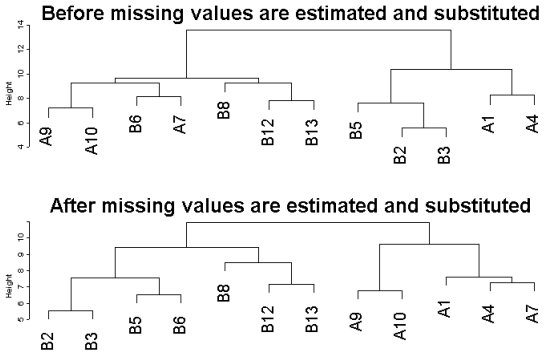

Figure 5. **Tree Structures of Unsupervised Hierarchical Clustering.** Each leaf in the tree represents one sample. A = 4 protein mixture; B = 5 protein mixture.

[b] R is a free statistics software, which can be downloaded at: http://www.r-project.org/

6. Conclusion

In this paper, we have shown that ignoring the intensity-dependent missing events in MS experiments may result in severe biases in the data analysis. To address this problem, we developed a probability model for the missing events and implemented a few normalization schemes to remove the negative effects. The missing rate estimates can also be used as a quality control of the data.

In the probability model, given that one peptide exists in the sample, a normal density is used to approximate the distribution of the intensity of this peptide. This approximation is supported by the synthetic data example: the Kolmogorov-Smirnov distance between $N(0,1)$ and the observed distribution of intensities (centered to $mean = 0$ and scaled to $sd = 1$) is 0.0392, which corresponds to a p-value of 0.1742.

When we estimate the missing values with nearest-neighbor scheme, the iteration number need to be carefully controlled to avoid problem of over-fitting.

Acknowledgments

We would like to thank three referees and Andrea E Detter for the comments that improved this manuscript. This work was funded by National Cancer Institute contract $\sharp 23XS144A$. HT was partially supported by NIH-CA86368. HZ, JW and AP were partially supported by philanthropy from Listwin Foundation/Canary Fund, Paul G. Allen Family Foundation and Keck Foundation.

References

1. J. Listgarten and A. Emili, *Molecular and Cellular Proteomics* **4.4**, 2005.
2. B.L. Mitchell, Y. Yasui, C.I.Li, A.L.Fitzpatrick and P.D.lampe, *Cancer Informatics* **1(1) 25-31**, 2005.
3. M. Man and R.Aebersold, *Nature* **422**, 2003.
4. J. Quackenbush, *Nat. Genet.* **32**, 2002.
5. A. Sauve and T. Speed, *Proceedings Gensips*, 2005.
6. M. Wagner, D. Naik and A. Pothem, *Proteomics* **3**, 1692-1698, 2003.
7. K.A. Baggerly, J.S. Morris, J. Wang, D. Gold, L.C.Xiao and K.R. Coombes, *Proteomics* **3**, 1667-1672, 2003.
8. M. Anderle, S. Roy, H. Lin, C. Becker and K. John, *Bioinformatics* **20**, 3575-3582, 2004.
9. R. Tibshirani, T. Hastie, and et.al. *Bioinformatics* **20**, 3034-3044, 2004.
10. W. Wang, H. Zhou,and et.al. *Anal. Chem.* **75**, 4818-4826, 2003.