Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum

Markus Anderle, Sushmita Roy, Hua Lin, Christopher Becker and Keith Joho

SurroMed, Inc., 1430 O’Brien Drive, Menlo Park, CA 94025, USA

Received on July 17, 2003; revised on June 18, 2004; accepted on July 26, 2004
Advance Access publication July 29, 2004

ABSTRACT
Summary: Using replicated human serum samples, we applied an error model for proteomic differential expression profiling for a high-resolution liquid chromatography-mass spectrometry (LC-MS) platform. The detailed noise analysis presented here uses an experimental design that separates variance caused by sample preparation from variance due to analytical equipment. An analytic approach based on a two-component error model was applied, and in combination with an existing data driven technique that utilizes local sample averaging, we characterized and quantified the noise variance as a function of mean peak intensity. The results indicate that for processed LC-MS data a constant coefficient of variation is dominant for high intensities, whereas a model for low intensities explains Poisson-like variations. This result leads to a quadratic variance model which is used for the estimation of sample preparation noise present in LC-MS data.

Contact: manderle@surromed.com

INTRODUCTION
Proteomics has been recognized as a major contributor to the effort to uncover disease pathways and biomarker discovery for drug development (Griffin et al., 2001). Protein differential expressions can be measured using high-resolution liquid chromatography-mass spectrometry (LC-MS), which produces a large set of relative intensities at chromatographic retention time and values of mass/charge (m/z). One goal of such studies is to identify a biomarker that serves as a diagnostic tool for disease progression and susceptibility, as well as for pharmacological response.

The use of MS technology in high-throughput proteomics faces several challenges in order to compare differentially expressed proteins using peak information, such as retention time, m/z and signal intensity. Software solutions for peak sensing, spectrum alignment, de-noising and de-isotoping needed to be implemented in order to successfully find biologically significant differences among samples. We have developed such a software and demonstrated its functionality for quantitative differential expression using human serum (Wang et al., 2003).

A major goal of differential proteomics is to identify differentially expressed components under different clinical conditions. Several studies report on the successful discovery of (m/z) patterns that are able to discriminate between patient samples using different classification algorithms (Petricoin et al., 2001; Lilien et al., 2003; Wu et al., 2003). To properly assign statistical significance to changes in component intensity it is important to understand the various sources of variability, especially when considering small sample sizes. Error models, which explain and quantify random variations have been developed for the analysis of differential gene expressions, measured with DNA microarray technology (Rocke and Durbin, 1995, 2001; Durbin et al., 2002).

Error models base their utility on the combination of variance estimates across intensities with similar average expressions, while ‘borrowing strength’ across intensities by modeling noise jointly. Different approaches exist which estimate variances for differential expressions as they vary across average intensities (Ideker et al., 2000), taking the entire set of expressions into account. Non-parametric techniques consider local sample averages across intensities to gain more robust estimates of the variance as a function of intensity. Examples of such approaches are the local windowed estimate and the local weighted sample average (Huang and Pan, 2002; Kamb and Ramaswami, 2001). The results of non-parametric methods are functional relationships, e.g. polynomial curves, that give an estimate of the variance as a function of intensity. Analytic approaches postulate an error distribution, taking into account different error structures depending on intensity, as demonstrated by Rocke and Durbin (2001), using a two-component error model to model the noise distribution.
model. This approach has the advantage that an interpretation of sources of variation and noise behavior is possible, e.g. a quadratic dependence of variance on mean intensity indicates a constant coefficient of variation (CV), or multiplicative noise.

In the analysis of microarray data from gene chips (mRNA gene expression profiling) to discover differences between conditions, many approaches have been proposed to use error models and noise analysis for the assignment of significant fold changes (Chen et al., 1997, 2002; Ideker et al., 2000; Newton et al., 2001). In the context of making statistical inferences from expression levels where arbitrary fold-change cut-offs are replaced by parametric (e.g. t-test) and non-parametric (e.g. wilcoxon test) tests, error models have an important application. In order to assign differential significance, one has to estimate the variance of components, which is problematic for small sample sizes. Because the number of replicates frequently is low, standard statistical techniques, such as t-tests, may need to be modified to be valid (Baldi and Long, 2001). When applying standard statistical tests to differential data analysis one must consider that t-tests, for example, assume normality and constant variance across samples. These assumptions may be inappropriate for some of the components despite any given transformation. The application of these tests also can not take advantage of the data when a correction for heterogeneity between samples is applied. Furthermore, it is also essential to correct for the possibility of a high number of false positives resulting from multiple comparisons when thousands of components are investigated for their significance. Otherwise, if a typical P-value of 0.05 were used to signify differential expression for individual genes between two groups, one would expect to find 50 positives for every 1000 components under examination, even though none of these genes are differentially expressed, assuming independence. Conceptually similar methods such as Significance Analysis of Microarrays (SAMs) have been developed to deal with this issue (Tusher et al., 2001). Although many error models show great improvement over a simple fold-change cut-off, component-specific random fluctuations are usually neglected which leads to increased numbers of Type I and Type II errors in the statistical analysis of differential expressions.

Several methods have been proposed to estimate and model the variance of intensity components as a function of component mean under a specific transformation. The logarithmic function transforms intensity ratios into differences and helps centralizing intensity distributions. In the context of error models, the log function transforms multiplicative error terms into additive effects and therefore stabilizes the variance of high-intensity MS components. The importance of variance-stabilizing transformations derived from error models has been shown in Durbin et al. (2002) and Huber et al. (2002). We have shown that variance-stabilization can also be observed for high-intensity MS intensity components, after the application of the log transformation.

Rocke and Durbin (1995) developed a two component error model for gas chromatography-MS measurements. However, the quantification of sample preparation and instrument variation of MS components built from chromatographic retention time and m/z information has not yet been considered. An approach to find adaptive significance thresholds that consider variations as a function of peak intensity and lead to differential protein detection across intensity levels has not yet been reported for the MS data. Such error models are now a common tool for helping to validate differentially expressed genes. Here, we show that the effects of systematic sources of variation in protein expression data can be minimized by normalization, and formulate an error model that quantitatively identifies sources of measurement noise. It is essential to quantify variability in a production environment, for example, to estimate the variation from sample preparation versus pure equipment. Future work will focus on different sources of biological variation.

In this paper, we describe in detail the reproducibility experiments using identical (pooled) samples as part of a quality control process. Based on these experiments, we introduce global and local error models as a function of local peak intensity and quantify strong and weak sources of variation.

QUALITY CONTROL EXPERIMENTS
Sources of variation and sample flow
Variability in experimental measurements is introduced by the sample preparation process and analytical equipment performance. Sample preparation involves several steps, such as the removal of abundant proteins [human serum albumin and immunoglobulin G (IgG)] and the reduction, alkylation and tryptic digestion of the proteome (Wang et al., 2003), all of which contribute to some degree to sample preparation noise. Equipment variations due to LC-MS instrument uncertainties include for example LC separation, electrospray ionization, ion suppression and ion detection.

The sample flow was designed to distinguish and quantify the two independent sources of variation: sample preparation and LC-MS equipment. Figure 1 shows a schematic illustration of the sample preparation process. We compared two cases, both beginning with pooled human serum. A total of 20 identical samples were run on the LC-MS platform with the samples individually processed, and 20 samples were re-pooled and then split after sample processing but before introducing them to the equipment.

Materials and methods
Pooled human serum used in the proteome analyses was purchased from Sigma-Aldrich (St Louis, MO). Affinity beads for albumin and IgG removal were from ProMetic Biosciences (Cambridge, UK). All other general reagents were purchased.
IgG-depleted serum was denatured by 6 M guanidine hydrochloride digested as follows. The human serum albumin and IgG-depleted serum proteome was denatured, reduced, alkylated and trypsin-digested to affinity beads (Prometic Biosciences) for human serum albumin and IgG removal. The albumin- and IgG-depleted human serum was diluted with 25 mM phosphate-buffered saline (PBS) buffer (pH 6.0) before it was applied to affinity beads (Prometic Biosciences) for human serum albumin and IgG removal. The albumin- and IgG-depleted serum proteome was denatured, reduced, alkylated and trypsin-digested as follows. The human serum albumin and IgG-depleted serum were denatured by 6 M guanidine hydrochloride, reduced by 10 mM dithiothreitol at 37°C for 4 h, and alkylated with 25 mM iodoacetic acid/NaOH at room temperature for 30 min in dark. The denaturant and reduction-alkylation reagents were removed from the mixtures by buffer exchange against 50 mM (NH₄)₂CO₃ at pH 8.3 using a 5 kDa molecular weight cut-off spin filter (Millipore, Billerica, MA). Modified trypsin (Promega Corp., Madison, WI) of 1% weight equivalence of the proteins was then added to the mixtures with incubation at 37°C. Samples were then diluted and analyzed by reversed-phase LC-MS using a high-resolution time-of-flight (TOF) mass spectrometer instrument (LCT, Waters, Micromass) coupled with a capillary LC system (Cap 1100; Agilent). An aliquot of 20 μl of ~20 μg of tryptic peptides per sample were injected onto a capillary column. Gradient elution of the proteome sample was achieved using 100% solvent A (0.1% formic acid in H₂O) to 40% solvent B (0.1% formic acid in acetonitrile) over 100 min. Quantification was based on the reproducibility of analyte signals as well as their linear dependence on concentration (Wang et al., 2003).

**MS data processing and quantification**

Each mass spectrum at a given elution time (scan), typically spanning one to a few seconds, was first processed by removing a modeled baseline by subtraction from the original spectrum. The mass spectrum peaks were recovered by smoothing the data, which also served as a noise filter. After peaks were identified, isotopic patterns were assigned, also known as de-isotoping (Wang et al., 2003). An intensity value was then recorded for the molecular ion. A condensed peak list was generated by linking together mass spectral peaks that appear in neighboring scans within one sample, as expected for a chromatographic peak elution, by defining a m/z window appropriate for the mass spectrometer. Using an arbitrarily chosen reference sample, the LC-MS retention times were adjusted for small variations between samples, and intensities were normalized in a global fashion (one separate normalization constant for each sample). Corresponding mass spectral peaks found in several samples became MS components of the study, and a peak intensity was recorded for each sample in the study. The details of this process are outlined below.

The peak intensities were determined after smoothing by using either the peak area or the maximum peak height in ion counts. Both approaches have been used successfully, and a maximum peak height has been used with the TOF mass spectral data in this study. Furthermore, it is possible to sum over all isotopes or use the intensity of the monoisotopic peak; in this work, we used the monoisotopic peak intensity. Mass spectral peaks in adjacent scans, if within the user-selected m/z window, were assigned as originating from a single ion species. From this process, a condensed peak list was created for each sample. A typical m/z window is ±0.10 Da for the TOF data. The retention time for the condensed peak list was associated with the maximum signal. Small temporal deviations over the elution profile were corrected by dynamic realignment with respect to a reference scan in order to correct for instrument sensitivity drift over time and/or sample-to-sample concentration variations. Additional normalization was performed by choosing one sample as a reference and individually normalizing all other files relative to the reference sample, one sample at a time. For each sample’s normalization, a set of pairs of corresponding peaks from the two condensed peak lists were chosen based on close proximity in retention time and m/z. The single normalization constant for each sample was taken as the median of the ratios of intensities for all components between the sample in question and the reference sample. At this point in the algorithm, component building took place for the study by scanning the condensed de-isotoped peaks listed according to their retention time and m/z values. This process assigns peaks that fall within the user-adjustable retention time and m/z windows to the same MS component in each sample it appears. In the component building process, a threshold was user-selected (commonly at about 20 ion counts). Using a relatively conservative threshold resulted in a more reliable and robust component building process. In this study, we considered MS components that were built after this threshold was applied.
ERROR MODELS

Data driven techniques
>An example for the quantitative noise analysis is given by Tu et al. (2002), where a noise distribution is estimated for pairs of replicated experiments. Starting from the noise distribution function, a quantitative model for the dependence of the variance from mean intensity is developed and used to estimate the strength of different noise sources. An technically similar approach was used in the Locally Pooled Error (LPE) test (Jain et al., 2003), where local estimates of variances were pooled together on sub-intervals of equal percentile ranges of intensities. We also used the LPE approach as a non-parametric technique to complement an analytical model which is based on two components that contribute to the overall variance structure.

Analytic techniques
>The theoretical description of observed heteroscedasticity in the intensity data is the goal of analytical error model approaches, where a correction for factors due to protein-specific heterogeneity is neglected. In order to assign significance to fold-changes, multiplicative and additive error models have been proposed by Rocke and Durbin (2001) and Durbin et al. (2002). Analytic approaches also lead to the formulation of variance stabilization transformations, such as the ars� function (Huber et al., 2002), after it was demonstrated that variances of log ratios of intensities $[\log(I_1/I_2)]$ also depend on signal intensity (Rocke and Durbin, 2001), as shown in Figure 3. Baggerly et al. (2001) derived such an error model from the biological mechanisms of the microarray while using replicates to calibrate parameters in the model. We applied a quadratic variance dependence to describe the high-intensity variation, where a constant CV is applicable. This leads to a multiplicative description of the intensity distribution, which is numerically supported by the non-parametric LPE technique, as shown below.

Global error model
>Figure 2 shows the scatter plot for 3763 intensity pairs of samples 17 versus 9 for the pooled set of proteomic expression intensities. The scatter is approximated reasonably well by the bisector line, which indicates that systematic variations were minimized during the MS data processing step by normalization and chromatographic retention time alignment. When the log-transformed data are investigated, it becomes evident that the variance spread is not uniform but depends on the mean log-intensity. The RI (ratio by intensity) scatter plot of log ratios versus the average of log intensities has become an important diagnostic plot for detecting intensity-dependent biases (Dudoit et al., 2002). In the RI plot (Figure 3), the difference from log intensity is evident as a function of average log intensity for samples 17 and 9, and supports the observation of a variance dependence on component mean.

We first investigated the global component variance structure and argue that measurement noise varies linearly with component mean resulting in a constant CV for components having high intensities. This translates into an approximately
Fig. 4. The dependence of component-wise standard deviation as a function of component mean can be approximated globally with a linear relationship. This is characterized by a constant CV, indicated by the line with slope CV\textsubscript{pooled} = 0.114.

constant error variation for the log-transformed data. This global observation is consistent with many error model approaches which assume multiplicative errors:

\[ x \approx \mu \epsilon. \]  \hspace{1cm} (1)

For observed intensities \( x \) with true mean intensity \( \mu \) and error \( \epsilon \) this results in a variance dependence

\[ \sigma^2_x \approx \mu^2 \sigma^2_\epsilon, \]  \hspace{1cm} (2)

with a global intensity CV \( \equiv \sigma_\epsilon/\mu = \sigma_\epsilon \).

For purposes of statistical analysis, e.g. mean comparisons using ANOVA, the logarithm transformation converts multiplicative errors into additive errors. If variances are proportional to signal intensity on the original scale, the application of the log transformation will result in a constant variance across the range of signal intensity on the logarithm scale. The logarithmic transformation of intensities \( y \) results in an additive error:

\[ y = \log(x) \approx \log(\mu) + \eta. \]  \hspace{1cm} (3)

The global linear dependence in the \( \sigma \) versus \( \mu \) plot is shown in Figure 4, where 2% or 75 MS components were dropped due to inflated CVs, probably a result of misclassified peaks during the component building process. In addition, one can observe a flattening of the trend for intensities over 4000 counts, which is due to saturation in the LC-MS detector. In Figure 5, the LPE variance estimations (circles) clearly indicate a quadratic behavior of the sample variance on the intensity mean. This observation is supported by the close fit of the LPE data to the global variance model (2).

In addition to the \( \sigma \) versus \( \mu \) dependence in Figure 6 for the pooled samples, the approximated linear fit for the
individually (ind) processed samples are shown. The slopes of the two linear curves approximate the two CVs as $CV_{\text{pooled}} = 11.4\%$ and $CV_{\text{ind}} = 20.2\%$. In detail, the quadratic fit to the pooled data resulted in $\sigma^2_{\text{pooled}} = 0.013$ and $\sigma^2_{\text{ind}} = 0.041$. Assuming independence of the two sources of variation (sample preparation and LC-MS instrument variation), an estimation of the sample preparation variation is given by $\sigma^2_{\text{ind}} - \sigma^2_{\text{pooled}} = 0.028$, resulting in a $CV_{\text{sample prep}} = 16.7\%$.

**Local error model** As evidenced in Figure 3, the constant CV and linear $\sigma$ dependence on mean intensity does not hold for smaller intensity values due to additional noise sources dominant for such intensity ranges. This is also consistent with the two component error model introduced by Rocke and Durbin (2001), where the increase in variance at log expression levels is explained by constant background noise. In Tu et al. (2002), this noise component is explained due to Poisson processes present in measurement and readout devices in addition to the sample preparation process.

We also found that the addition of a linear term to the quadratic variance model fits the data well for smaller intensity ranges:

$$\sigma^2_x = \alpha \mu^2 + \beta \mu.$$  

(4)

For the CV of a MS component with intensity $x$, we obtained the following model:

$$CV_x = \sqrt{\alpha \mu^2 + \beta \mu / \mu} = \sqrt{\alpha + \beta / \mu}.$$  

(5)

The standard deviation $\sigma_x$ is now also proportional to the second term in (4) as the square root of $\mu$, which is indicative of Poisson-like noise. This is consistent with the exponential dependence for the log-transformed data, derived in Tu et al. (2002), which can be explained by the ion statistics of the signal detector at low-intensity levels:

$$\sigma^2_x = \alpha + \beta \exp(-y).$$  

(6)

Fitting model (4) to the raw pooled data results in $\sigma_x = 0.013 \mu^2 + 0.89 \mu$ with $\alpha = 0.013$ and $\beta = 0.89$. Respectively, a fit to the individually processed data results in $\alpha = 0.040$ and $\beta = 4.80$. The result of the fit to equation (5) can be observed in Figure 7 where the component-wise variance of the log-transformed data is shown together with the fitted exponential model. Figure 8 displays the dependence of the CV on component intensity, together with the CV curve resulting from plotting Equation (4) with the fitted parameters. Additionally, we included the mean CV on 100 sub-intervals of equal percentiles in both CV plots. As indicated by the CVs calculated on equal percentiles, the model follows the experimental data reasonably well.

Since the sample preparation process depends on multiple factors, its influence on the error model must be viewed as an rough approximation to the underlying variation of chemical processes involved in the sample preparation.

**SIGNIFICANCE FOR FOLD CHANGES**

A typical application of error models, apart from transforming the data to constant variance, is the estimation of the component variances for statistical tests, e.g. $t$-tests for samples under different clinical conditions. It has been shown that one can significantly increase the power to detect true differences compared to the Welch $t$-test, used in combination with adjustments for multiple testing by employing variance estimates from error models (Huang and Pan, 2002). In our design of replicated samples, we only compared the individually processed samples to each other by randomly dividing the group of 20 samples into two groups of 10 samples each. Here, we tested the null hypothesis $H_0$ for each component $i$ to have the same mean for both subgroups. We used the $t$-statistic with a significance level of $\alpha = 0.05$ and recorded the number of false rejections for 3763 components. The standard $t$-test resulted in 31 false positives, whereas the $t$-test with model-estimated variances resulted in 220 false positives, a more reasonable result because the observed Type I errors were closer to the specified nominal level $\alpha = 0.058$, compared to 0.008 for the standard $t$-test.

**DISCUSSION**

In detecting differential proteomic expressions an important issue is the computation of a robust estimate of the variability of observed proteomic components. With small number of samples under different clinical conditions the commonly used sample variance estimator often exhibits large variability. We applied a two component error model, together with a sample averaging technique to calculate locally pooled
application of model-estimated variances to the detection of differences, we have observed a more reasonable number of Type I errors compared to sample estimated variance, which indicates that tests which use model-estimated variances have increased power.

ACKNOWLEDGEMENTS

We would like to thank Sophia Chen for the preparation of mass spectrometry samples and the preparation of a sample work plan for the experiments conducted for this study.

REFERENCES


