BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 8 2008, pages 1070–1077
doi:10.1093/bioinformatics/btn078

Gene expression

A noise model for mass spectrometry based proteomics

Peicheng Du[1,*], Gustavo Stolovitzky[1], Peter Horvatovich[2], Rainer Bischoff[2], Jihyeon Lim[3] and Frank Suits[1]

[1]IBM Computational Biology Center, P.O. Box 218, Yorktown Heights, NY 10598, USA, [2]University of Groningen, Antonius Deusinglaan 1, Postbus 196, 9700 AD Groningen, The Netherlands and [3]Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Received on September 10, 2007; revised on February 21, 2008; accepted on February 27, 2008

Advance Access publication March 18, 2008

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Mass spectrometry data are subjected to considerable noise. Good noise models are required for proper detection and quantification of peptides. We have characterized noise in both quadrupole time-of-flight (Q-TOF) and ion trap data, and have constructed models for the noise.

Results: We find that the noise in Q-TOF data from Applied Biosystems QSTAR fits well to a combination of multinomial and Poisson model with detector dead-time correction. In comparison, ion trap noise from Agilent MSD-Trap-SL is larger than the Q-TOF noise and is proportional to Poisson noise. We then demonstrate that the noise model can be used to improve deisotoping for peptide detection, by estimating appropriate cutoffs of the goodness of fit parameter at prescribed error rates. The noise models also have implications in noise reduction, retention time alignment and significance testing for biomarker discovery.

Contact: pdu@us.ibm.com

Supplementary information: Supplementary data are available at *Bioinfomatics* Online.

1 INTRODUCTION

Mass spectrometry is an increasingly important analytical tool for detection and quantification of proteins and peptides in biological samples. However, mass spectrometry is limited by the almost ubiquitous presence of noise. The specific aim of this work is to characterize measurement noise in mass spectrometry in order to improve protein and peptide detection and quantification. In this study, we focus on the measurement noise of proteins and peptides.

Not surprisingly, noise has not been the primary target of interest in previous studies. In most cases, noise is assumed to be Gaussian. Although noise models for microarray data have been developed (Tu *et al*., 2002; Weng *et al*., 2006), noise models for mass spectrometry data are still neglected. This is the case even when it is clear that mass spectra are noisy, and therefore peaks of biological importance may be masked. Among a small number of studies on noise characterization, Shin *et al*. (2004) hypothesized that three types of noise exist in MALDI TOF spectra: Johnson noise due to the electrical system, shot noise or Poisson noise due to the discrete nature of the ion signal, and chemical noise due to matrix ions. Anderle and coworkers developed a quantitative error model for protein liquid chromatography-mass spectrometry (LC-MS), and reported that the error and intensity have a linear relationship for time-of-flight (TOF) mass spectral data (Anderle *et al*., 2004). Because LC-MS involves many steps from sample preparation to data analysis, it is unclear how much error is contributed by each step. Blackler *et al*. (2006) studied the S/N ratio, which is crucial to protein identification, and stated that noise from ion trap spectrometers consists of a component independent of signal, a component linearly related to signal and a shot noise component proportional to the square root of the signal intensity (Blackler *et al*., 2006). However, they did not develop a quantitative noise model based on this conclusion.

A common problem requiring a noise model is in deisotoping, which refers to the recognition of peptides from mass spectral peaks due to heavy isotopes. Deisotoping is a fundamental task in proteomic data analysis. It involves fitting the intensity of observed isotopic distributions (OID) at certain mass to charge ratios to the expected isotopic distributions (EID). A widely used method of fitting OID to EID is to use a least-squares fit (Du and Angeletti, 2006; Horn *et al*., 2000; Leptos *et al*., 2006). Both ordinary least-squares fit and the closely related cross-correlation method (Higgs *et al*., 2005; Wang *et al*., 2003) implicitly assumes that noise is Gaussian and is equal for all peaks. Bellew and coworkers (Bellew *et al*., 2006) used the Kullback–Leibler distance (Kullback and Leibler, 1951) to measure the deviation of OID to EID. Alternatively, Kaur and O'Connor used the multinomial distribution of isotope patterns to fit OID to EID (Kaur and O'Connor, 2004, 2007) in order to estimate the number of ions and the isotopic ratios. In most cases, the assumed noise model was not verified from experiments. Additionally, noise models should also improve the detection of significant biological differences in applications such as biomarker discovery. Error models of microarray data have been used to improve the statistical testing in such data (Weng *et al*., 2006).

In this work, we study data from electrospray quadrupole time-of-flight (Q-TOF) and ion trap instruments, both of which are widely used. Since multiple spectra for a peptide are

*To whom correspondence should be addressed.

available in LC-MS or LC-MS/MS (from scans nearby in retention time), we choose to examine the noise in the isotopic clusters of peptides. LC-MS is a convenient way to study noise since it automatically provides related scans of each peptide during elution of chromatographic peaks.

2 METHODS

2.1 Experimental methods for collection of rat serum spectra from Q-TOF instrument

The experimental conditions for the mixture of the rat serum spectra have been described previously (Du and Angeletti, 2006). Briefly, normal serum is trypsin digested prior to injection. Subsequently, peptides are injected into a strong cation exchange column with step gradient at salt concentrations of 50, 60, 70, 80, 90, 100, 200, 300 and 600 mm. Each fraction is then injected into the C18 column and the LC-MS spectra are collected with a QQ-TOF mass spectrometer (QSTAR Pulsar, Applied Biosystems, Foster City, CA). The C18 column used is 75 μm i.d. × 25 cm (Dionex, Sunnyvale, CA). The flow rate is 250 nl/min. Nanoelectrospray Sources are used for electrospray ionization (ESI). TOF-MS scan is performed in the m/z range of 300–1800 with a scan time of 1 s. This dataset is referred to below as 'SCX' to indicate ion exchange is used for prefractionation.

A second Q-TOF dataset is generated for noise model testing. The experimental conditions are mostly the same, except that fractionation is performed by 1D-PAGE. Six gel spots are trypsin digested separately. Digested sample from each spot is then injected into a QSTAR XL hybrid LC-MS/MS system, which is a similar Q-TOF instrument from the same vendor. This Q-TOF dataset is subsequently referred to as 'GEL' to indicate that 1D-PAGE is used for prefractionation.

2.2 Experimental methods for collection of Cytochrome C spectra from ion trap instrument

Experimental conditions for the collection of Cytochrome C spectra have been described previously (Horvatovich *et al*., 2007). Briefly, ion trap data is acquired on a capillary LC-MS system (Agilent, Palo Alto, California, USA). Trypsin-digested horse heart Cytochrome C is analyzed with an HPLC system coupled online to a MSD-Trap-SL ion trap mass spectrometer (Agilent) with enhanced scan resolution, 5500 m/z per second scan speed, ICC target: 30 000, max. accumulation time: 15 000 μs, scan range: 100–1500 m/z. The spectra are acquired without rolling average and saved in profile mode.

2.3 Collection of peaks from the spectra

For the Q-TOF dataset 'SCX', protein identification is performed by analyzing the resulting MS/MS spectra using Mascot (Matrix Science, London, UK). The search parameters are: peptide mass tolerance: 100 ppm; fragment mass tolerance: 0.4 Da and maximum number of missed cleavages allowed: one. Only peptides with Mascot scores of at least 50 and e-values of below 0.01 are selected regardless of salt fractions. For each selected peptide, isotopic peak clusters of the peptides are taken from scans which satisfy the following criteria: (i) the scan is within ± 13 scans of the elution peak apex of that peptide; (ii) the first four isotopic peaks are all stronger than 30 counts and (iii) there is no obvious overlapping in either the m/z or time dimension with other peptides by visual inspection. The criteria are intended to select strong and clean peaks which are not or only weakly affected by the baseline (under five counts) or by peaks of other peptides, in order to characterize the noise in the peptide peaks. Finally, a total of 3276 peaks, i.e. 1092 isotopic clusters (three peaks per cluster) from 99 peptides are obtained for noise modeling of the Q-TOF spectra. All but three isotopic clusters are doubly charged. The rest are triply charged. The dataset 'GEL' is processed similarly. Only the 'SCX' dataset is used for fitting the noise model.

Similarly, peaks are collected from the ion trap spectra of a tryptic digest of horse Cytochrome C from eight replicate LC-MS analyses. Instead of performing a Mascot search, peaks are simply matched to a peptide list from an *in Silico* tryptic digestion. Eventually a total of 4176 peaks, i.e. 1392 isotopic clusters from 71 peptides in different replicate runs are collected and used for noise modeling of the ion trap spectra. All selected ions are singly charged to avoid peak overlapping due to limited resolution of the quadrupole ion trap mass analyzer.

2.4 Q-TOF data: fitting a multinomial and Poisson model

The OID should follow a multinomial distribution where each outcome is in fact an isotopic peak with a different number of extra neutrons. This is because the probability of each outcome (i.e. EID) can be calculated given the atomic composition and the abundance of relevant elements with the polynomial method of Yergey (1983). Among these elements, the natural abundance of ^{13}C has been reported to be in a tight range (Beavis, 1993), and it can be assumed to be 1.11% for the sake of noise analysis. Denote the probability of the first three isotopic peaks by p_1, p_2 and p_3 in the EID, then p_1, p_2 and p_3 can be calculated from the peptide sequence. Let the number of ions in these peaks in the OID be n_1, n_2 and n_3, which are the peak intensities for the Q-TOF instrument used, and denote the total number of peptide ions of all isotopic forms by n_{pep}, then the probability of observing the OID can be calculated using the multinomial distribution as follows:

$$P_{\text{EID}}(n_1, n_2, n_3; n_{\text{pep}}) = n_{\text{pep}}! \prod_{i=1}^{4} \frac{p_i^{n_i}}{n_i!}$$

where $n_4 = n_{\text{pep}} - n_1 - n_2 - n_3$ and $p_4 = 1 - p_1 - p_2 - p_3$, and p_1, p_2 and p_3 are calculated according to Yergey's method assuming that all elements (i.e. C, N, O, S, H) are of natural abundance, and $n_{\text{pep}} > n_i$ for $i = 1, \ldots, 4$. The adjustable parameter n_{pep} can be estimated by maximizing P_{EID} according to the principle of maximum likelihood.

However, the multinomial model alone is not sufficient to explain the OID. Because the ion count in a unit time is governed by Poisson statistics, the observed peak intensity n_i can be approximated as a sample from the Poisson distribution with a mean of t_i, which in turn is a sample from the multinomial distribution with parameters p_1, p_2, p_3 and n_{pep}. Therefore, the probability of observing intensity n_i given t_i for the i-th isotopic peak is

$$P_{\text{OID}}(n_i | t_i) = \frac{e^{-t_i}(t_i)^{n_i}}{n_i!}$$

For a given set of parameters p_1, p_2, p_3 and n_{pep}, there are multiple possible values of t_i, therefore the probability of observing the OID with peaks n_1, n_2 and n_3 is a summation over all possible values of t_i as follows:

$$P_{\text{OID}}(n_1, n_2, n_3; n_{\text{pep}}) = \sum_{t_1+t_2+t_3+t_4=n_{\text{pep}}} n_{\text{pep}}! \prod_{i=1}^{4} \frac{e^{-t_i}(t_i)^{n_i}}{n_i!} \frac{p_i^{t_i}}{t_i!} \ldots \quad (1)$$

where the adjustable parameter n_{pep} can be estimated by maximizing the above probability according to the maximum likelihood principle.

With N independent isotopic clusters ($N = 1092$ in 'SCX'), the probability p_{total} of observing all isotopic clusters is the product of all P_{OID}:

$$p_{\text{total}} = \prod_{i=1}^{N} P_{\text{OID}}(n_{1,i}, n_{2,i}, n_{3,i}; n_{pep,i}) \ldots \quad (2)$$

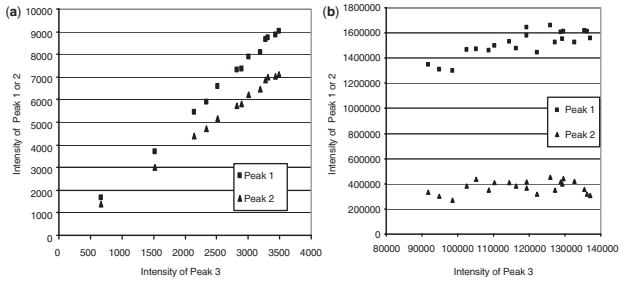

Fig. 1. Representative data from Q-TOF and ion trap instruments (**a**) Intensity of the first and second isotopic peaks versus the third isotopic peak in multiple scans of Q-TOF spectra for peptide VKDFATVYVDAVK with monoisotopic mass of 1453.8 Da. (**b**) Intensity of the first and second isotopic peaks versus. the third isotopic peak in multiple scans of ion trap spectra for peptide IFVQK with monoisotopic mass of 633.4 Da.

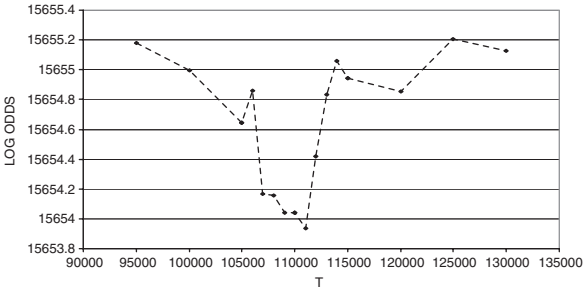

Fig. 2. Maximum likelihood estimation of T for the Q-TOF dataset 'SCX'. T is estimated to be 111 000 at the point of minimum odds as a global minimum. Only regions around the minimum are shown.

2.5 Q-TOF data: detector dead-time correction

The pure multinomial model described above assumes that isotopic peaks in a cluster are independently detected. The assumption of independent detection is often violated due to the detector 'dead-time effect', known to cause suppression of heavier ion counts by lighter ions that arrive at the detector slightly earlier in TOF instruments (Chernushevich *et al.*, 2001). The model can be modified to correct the dead-time effect prior to the multinomial and Poisson fit. Denote the intensity of the i-th isotopic peak by h_i, the intensity of the $(i-1)$-th isotopic peaks by h_i-1, where $i=0$ for the monoisotopic peak and $h_i-1=0$ when $i-1<0$, and the corrected intensity by $h_i^{'}$. To relate $h_i^{'}$ with the observed peak intensities h_i, we postulate the semi-empirical relation

$$h_i^{'} = \frac{h_i \log\left(1-(h_{i-1}+0.5h_i)/T\right)}{(h_{i-1}+0.5h_i)\log(1-1/T)}\cdots \tag{3}$$

where T is a constant that is related to the total number of detectors for each instrument, and $(h_{i-1}+0.5\ h_i)$ is the measured number of ions in the $(i-1)$-th isotopic peak plus half the measured number of the ions in the i-th isotopic peak (because all ions in the $(i-1)$-th peaks and the leading half of the i-th peak can suppress the i-th peak itself). It can be shown that according to Equation (3) the measured peak intensity is always smaller than the actual number of ions, that is, $h_i < h_i^{'}$. Furthermore, as the number of ions incoming to the detector increases, the value of $(h_{i-1}+0.5\,h_i)$ saturates to a constant, T. T is unknown and therefore can be fitted to obtain corrected intensities that maximize P_{total} according to Equation (2). Because P_{total} can only be calculated for Q-TOF data, detector dead-time correction is only applied to the Q-TOF dataset 'SCX' and not to the ion trap data.

2.6 Ion trap data: estimating intensity dependent noise

The noise in ion trap data is much larger than that in the Q-TOF data (Fig. 1) and appears to come from sources other than the variability in isotopic distribution alone. Indeed, a multinomial fit is not feasible because the resulting P_{total} is unrealistically close to zero. In addition, the reported counts in the intensities are no longer the actual number of ions (MacCoss *et al.*, 2001). Therefore, the goal is to characterize noise as a function of intensity.

The noise can be estimated as follows. For a pair of isotopic peaks with observed peak intensities, h_1 and h_2, we calculate the probability of an ion going into each isotopic peak, p_1 and p_2 using Yergey's methods assuming all elements to be of natural abundance. Denote n_{pep} as the total number of peptide ions in the isotopic peak cluster. Then we have two simple equations:

$$h_1 = n_{\text{pep}} p_1$$

$$h_2 = n_{\text{pep}} p_2$$

Assuming that the errors in h_1 and h_2 are roughly equal, the least-squares solution for n_{pep} is the following:

$$n_{\text{pep}} = \left(\frac{h_1 p_1 + h_2 p_2}{p_1^2 + p_2^2}\right)$$

The above equation is based on the assumption that the errors in h_1 and h_2 are roughly equal otherwise a weighted fit might be necessary. Since we assume that the error is a function of the signal intensity for ion trap data, similarity in peak intensities h_1 and h_2 implies that their errors are also similar. In practice, a subset of peak pairs is used for which p_1/p_2 is in the range of [1/2, 2], which implies the true peak intensities are roughly similar. At the same time, the range allows a number of peak pairs to be collected to build statistics.

With n_{pep} estimated, for the i-th isotopic peak, the true intensity should be $n_{\text{pep}}\ p_i$ and the error can be estimated as

$$Err_i = h_i - n_{\text{pep}} p_i = h_i - p_i\left(\frac{h_1 p_1 + h_2 p_2}{p_1^2 + p_2^2}\right)\cdots \tag{4}$$

3 RESULTS AND DISCUSSION

3.1 Q-TOF data: fitting a multinomial and Poisson model with detector dead-time correction

The Q-TOF dataset 'SCX' used in our noise estimate consists of 1092 isotopic clusters with three isotopic peaks in each cluster giving a total of 3276 peaks. The first step of the fit is to find the best-fit parameter, T, for dead-time correction. Figure 2 shows the convergence at the best-fit T that corresponds to the lowest log odds, defined as $-\log(p_{total})$. The best-fit T is found to be 111 000. The convergence at the best-fit value for T supports the validity of dead-time correction.

Next, the errors are calculated by plugging the best-fit T into the model. Figure 3 shows the error as a function of intensity for all peaks, with the first, second and third isotopic peaks in blue, red and yellow, respectively. Figure 3a shows the error without dead-time correction. An ideal residual plot should show only random errors and be free of systematic patterns. Most data points on Figure 3a seem to be random and roughly symmetric with respect to the x-axis. However, in the

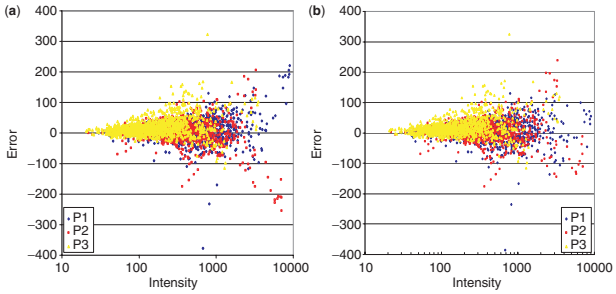

Fig. 3. Error versus observed intensity with and without dead-time correction for Q-TOF dataset 'SCX' in log scale. Error is calculated as the observed peak intensity minus the fitted peak intensity. The blue, red and yellow symbols represent the first, second and third isotopic peak, respectively. (**a**) Plot of error without dead-time correction shows the model underestimates the first isotopic peak and overestimates the second isotopic peak at high intensity. (**b**) Plot of error with dead-time correction at $T = 111\,000$ shows improved fit at high-intensity region.

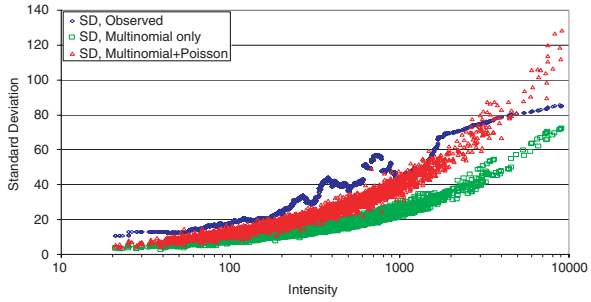

Fig. 4. Comparison of observed SD to those derived from the multinomial model and from the multinomial and Poisson model for Q-TOF dataset 'SCX' in log scale. The observed SDs are in blue, those derived from the multinomial model are in green and those derived from the multinomial and Poisson model are in red. The observed SD are calculated by first sorting peaks by intensity, then the SD of the ith peak is calculated as $\mathrm{SD}_i = \sqrt{\sum_{k=i-200}^{i+200} \mathrm{Error}_k^2/401}$, where Error_k is the error for the kth peak. The SD for the multinomial model is calculated as $\sqrt{n_{\mathrm{pep}} \times p \times (1-p)}$, where n_{pep} is the number of peptide ions from the best-fit multinomial model and p is the fraction of peptide ions in the isotopic peak calculated with Yergey's method. The SD for the multinomial and Poisson model is calculated by *Monte Carlo* simulation.

high-intensity region there is a clear pattern that the first isotopic peak (in blue) is underestimated and the second isotopic peak (in red) is overestimated by the model. Together with the clear convergence of log odds as a function of T in Figure 2, the error plot in Figure 3 indicates that either the first isotopic peak is somewhat enhanced or the second isotopic peak is suppressed. The latter is more likely because detector dead-time effect is known to cause suppression of peaks by ions that arrive at the detector slightly earlier (Chernushevich *et al.*, 2001).

Figure 3b shows the error plot with dead-time correction. The error plot is clearly improved in the high-intensity region, where points from all isotopic peaks are roughly symmetrically distributed about the x-axis, and there is no apparent bias for any isotopic peak. There are several points that appear to be outliers, which could be due to false protein identification by Mascot, overlapping peaks or solvent ions.

Figure 4 shows the SD in log scale calculated from the observed error (shown in Fig. 3b with dead-time correction) in blue, versus the SD calculated from the multinomial model in green, and a combined multinomial and Poisson model in red. According to the multinomial model, for a peak within an isotopic cluster that has a total of n_{pep} ions, assuming the fraction of the peak in the cluster is p calculated by Yergey's method, the intensity of this peak will vary in replicate measurements with a mean of $n_{\mathrm{pep}}p$ and a variance of $n_{\mathrm{pep}}p(1-p)$, simply because of sampling error in the multinomial process. However, it is obvious that the observed error is significantly larger than that prescribed by the multinomial model, i.e. the green squares.

By realizing that the ion counts contain shot noise characterized by a Poisson distribution, we can construct a noise model of multinomial and Poisson which can explain most of the observed noise, as shown in Figure 4. The analytical form for the variance in the multinomial and Poisson model is $n_{\mathrm{pep}}p + n_{\mathrm{pep}}p(1-p)$. It can also be estimated by *Monte Carlo* simulation in repeated random draws from a multinomial process followed by a Poisson process. As shown in Figure 4, the multinomial and Poisson model explains most of the observed noise, although a fraction of the observed noise is still of unknown source. Note that the observed plateau of the observed SD in blue for intensities above 4000 is likely an artifact, due to insufficient number of peaks in that region and the averaging effect of weaker peaks to the left. It is interesting to note that the actual SD is slightly larger than the model estimate, indicating that there can be extra sources of variability which are not considered in the present model. A more complex model may fit better to the data than the Poisson does, such as the Gamma distribution which has one more parameter. However, the Poisson distribution is preferred because it is regarded as a reasonable approximation of the detection process (Chernushevich *et al.*, 2001; Senko *et al.*, 1995).

3.2 Evaluation of noise in the ion trap data

From Figure 1b it is obvious that the relative noise level in the ion trap data is much larger than that in the Q-TOF data. Unlike Q-TOF data, the error in ion trap intensity is calculated using Equation (4), and plotted in Figure 5a. The noise of each peak can then be estimated by binning peaks of similar intensity and calculating the SD from the intensity error. Figure 5b shows the log–log graph of signal-to-noise ratios versus signal intensity, where signal is the peak intensity. The solid line shows the relationship for the Cytochrome C data. Apparently the trend is roughly parallel to the dashed line, which shows the theoretical square root relationship predicted by a Poisson model. The trend in Figure 5b is insensitive to changes of window sizes for binning.

Our results are consistent with previous findings regarding noise in ion trap instruments. Blackler *et al.* (2006) found that noise in quadrupole ion trap instruments, LCQ and LTQ (both from Thermo Finnigan), is Poisson limited in which the measured signal-to-noise ratio versus noise relationship is parallel

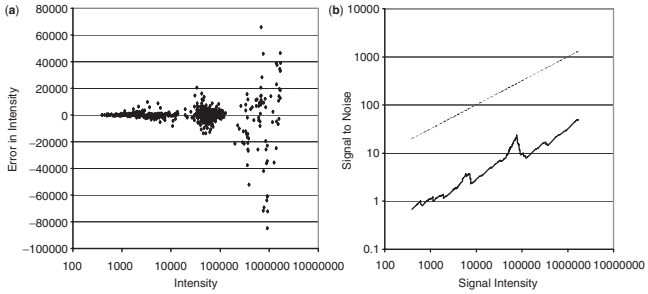

Fig. 5. Plot of error versus intensity for the ion trap data. (**a**) Error in intensity versus intensity. Error in intensity is calculated according to Equation (4). (**b**) Log–log graph of signal-to-noise ratios versus intensity. The signal-to-noise ratio is the ratio of peak intensity to the SD of the peak, which is calculated in the same way as in Figure 4, except with a sliding window size of 201. Results are plotted as a solid line. The dashed straight line shows the square root of signal intensity versus signal intensity, as predicted by the Poisson noise model.

to that predicted by the Poisson model. According to Blackler *et al.*, the distance between the solid line and the dashed line in Figure 5b represents the difference between signal intensity and the actual number of ions. The distance is roughly 10-fold for LCQ, and 2–3-fold for LTQ. Our result in Figure 5b shows the distance between the lines is also approximately one order of magnitude, which is more similar to the LCQ than to the LTQ. Alternatively, Li *et al.* (2006) found that the noise in LTQ is proportional to the 0.7–0.8th power of the intensity. Our result is therefore more consistent with a Poisson noise model.

The importance of our result on the ion trap data is 3- fold. First, to our knowledge this is the first study in the literature on the noise model of a MSD-Trap-SL instrument from Agilent, which shares wide use along with LCQ and LTQ. Second, compared to Blackler *et al.*'s results, our study shows Poisson noise is dominant on a different ion trap instrument, and by a different method of analyzing the isotopic patterns instead of repeated injections. Third, the ion trap noise model complements the Q-TOF noise model and allows comparison of ion trap noise to Q-TOF noise. The ion trap and Q-TOF instruments are important for proteomics; together, they account for most of the protein identifications in the HUPO Plasma Proteome Project (Omenn *et al.*, 2005).

3.3 The noise model improves deisotoping

The noise models developed have many applications that can improve the statistics of mass spectrometry results. One important application is deisotoping, which involves fitting the intensity OID at certain mass to charge ratios to the EID. A widely used method of fitting OID to EID is to use a least-squares fit (Du and Angeletti, 2006; Horn *et al.*, 2000; Leptos *et al.*, 2006). A typical measurement of the goodness of fit is the Pearson correlation coefficient (below referred to as r). The problem of using r is that a good cutoff has to be chosen to distinguish peptide peaks from non-peptide ones.

In this section, we first show that a single r cutoff is not optimal for all peptides of all intensities. Instead, the cutoff should be intensity and mass specific. We then proceed to show that the noise model can be used to estimate r cutoffs for

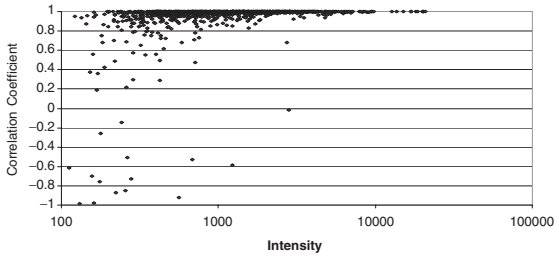

Fig. 6. Correlation coefficients versus intensity for all peptides in the Q-TOF dataset 'SCX'. The intensity is defined as the sum of the first four isotopic peaks in a cluster and is plotted in log scale.

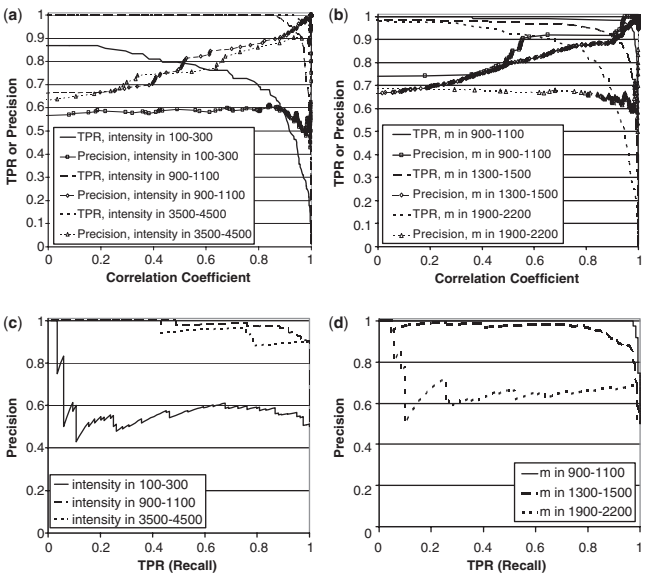

Fig. 7. The effect of intensity and mass on the TPR and precision in deisotoping using correlation coefficient r on the dataset 'SCX'. (**a**) TPR and precision are intensity dependent. (**b**) TPR and precision are mass dependent. The legend 'm in 900–1100' means that the mass is in the range of 900–1100 Da. (**c**) The precision–recall curve corrresponding to (a). (**d**) The precision–recall curve corrresponding to (b).

given error rates, or vice versa, for any mass and intensity combination.

Figure 6 shows that r depends on intensity for all annotated peptides in the Q-TOF dataset of 'SCX'. Peptides with stronger intensities generally have larger r, which is consistent with the multinomial and Poisson noise model that predicts the signal-to-noise ratio generally increases with intensity. While the r for most peptides at intensities around 10 000 is greater than 0.95, a cutoff of 0.95 apparently rules out most peptides with intensities around 200. The calculation of r is independent from the noise model.

The effect of intensity on the error rates of deisotoping using r is shown in Figure 7a. The error rates are estimated as follows. First, a dataset is created which contains each of the 1092 isotopic clusters in the 'SCX' dataset, and exactly one permuted version for each of original 'true' isotopic cluster. The original clusters are labeled as 'true' and the permuted clusters are

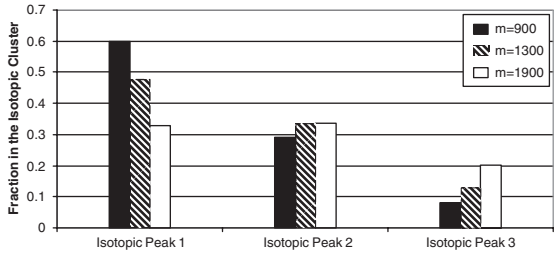

Fig. 8. Comparison of the expected isotopic distributions for masses of 900, 1300 and 1900 Da. The averaging model (Senko *et al.*, 1995) is used for the expected distribution at each mass.

labeled as 'false'. The dataset is referred to below as 'SCX-PMU'. Note that in such a dataset, 'true' and 'false' are equally likely. Then the true positive rate (TPR), or recall, can be calculated as the fraction of 'true' isotopic clusters which are above a certain *r* cutoff among all 'true' ones. The precision can be calculated as the fraction of 'true' clusters among all clusters which are above a certain *r* cutoff. The TPR and precision for each *r*-value are calculated for each of the three intensity regions, 100–300, 900–1100 and 3500–4500, respectively, by using subsets of 'SCX-PMU' which contain only clusters in the corresponding intensity region. The TPR and precision as a function of *r* are shown in Figure 7a, where solid, dashed, and dotted lines represent each of the three intensity regions from weak to strong. The lines with or without markers represent precision and TPR, respectively. Apparently, the general trend is that both TPR and precision decrease as intensity decreases. The corresponding precision–recall curve is also shown in Figure 7c, though it does not contain the *r* cutoff.

Appropriate *r* cutoffs should also be mass specific. Similarly, it can be shown that peptide masses also affect the error rates of deisotoping using *r*, as shown in Figure 7b. With the same dataset of 'SCX-PMU', the TPR and precision for each *r*-value are calculated for three mass regions, 900–1100, 1300–1500 and 1900–2200 Da, respectively, by using subsets of 'SCX-PMU' which contain only clusters in the corresponding mass regions. Clearly, both TPR and precision decrease as mass increases for the masses used. The explanation is that the isotope patterns of lower masses have more 'information content' than those with higher masses. Specifically, the ratios of the first two isotopic peaks for masses 900, 1300 and 1900 Da are approximately 2.1, 1.4 and 1.0, as shown in Figure 8. It is more likely to obtain two peaks with a ratio of 1.0 than to have two peaks with the ratio of 2.1 by random chance.

The Q-TOF noise model can be used to estimate *r* cutoffs corresponding to given error rates, or to estimate error rates for a given *r*-value. This is feasible because *r* directly reflects the noise. To show this with actual data, four mass and intensity regions are chosen first such that each region contains at least 36 isotopic clusters for meaningful error rate estimation. The two mass regions of 1000–1400 and 1400–1900 Da, and the two intensity regions of 100–350 and 500–800 form four mass and intensity regions. Higher intensities are not selected due to the lack of enough peptides with high intensity. An additional reason is that, the noise model is more useful for deisotoping at low intensities where the signal-to-noise ratio is low. For each

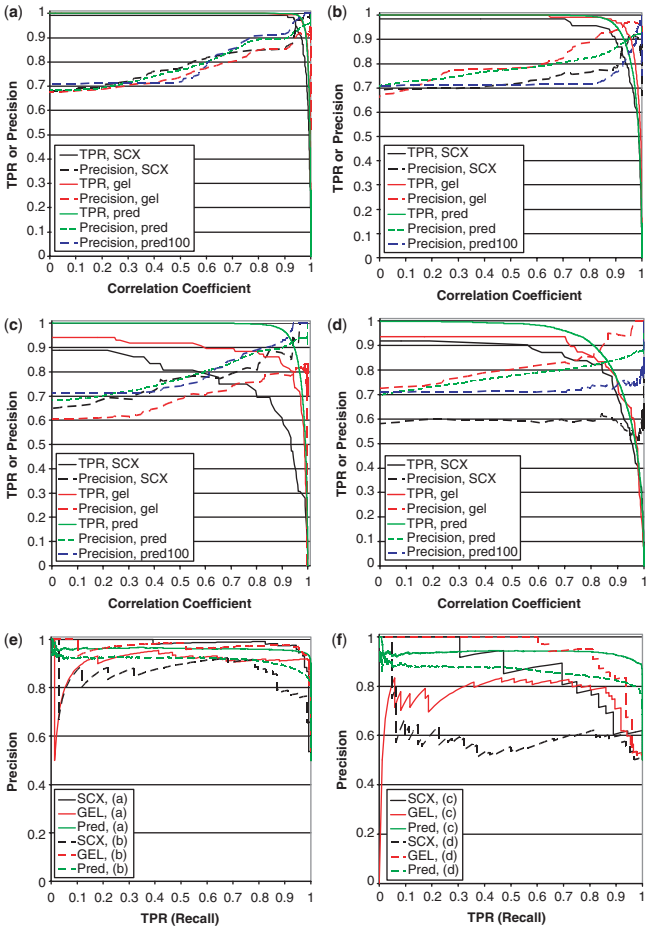

Fig. 9. Comparison of observed and predicted TPR and precision for datasets 'SCX-PMU' and 'GEL-PMU'. 'SCX' means the dataset is 'SCX-PMU'. 'Gel' means the dataset is 'GEL-PMU'. 'Pred' means the rates are predicted by simulation with the multinomial and Poisson noise model, with a total of 5000 simulated isotopic clusters. 'Pred100' means the rates are predicted with the first 100 simulated isotopic clusters. (**a**) The region of mass in 1000–1400 Da and intensity in 500–800. (**b**) The region of mass in 1400–1900 Da and intensity in 500–800. (**c**) The region of mass in 1000–1400 Da and intensity in 100–350. (**d**) The region of mass in 1400–1900 Da and intensity in 100–350. (**e**) and (**f**) are the corresponding precision–recall curves for the above Figure (a)–(d).

mass and intensity region, TPR and precision are calculated from subsets of the 'SCX-PMU' dataset with the specified mass and intensity. Likewise, those rates are also calculated for subsets of the 'GEL-PMU' dataset in those regions. Note that the 'GEL-PMU' dataset is not used to fit the noise model. The number of isotopic clusters in each region ranges from 36 to 150, with an average of about 100. Finally, TPR and precision are predicted from the multinomial and Poisson noise model for each *r* cutoff at the given mass and intensity by *Monte Carlo* simulations of 5000 isotopic clusters.

In Figure 9a–d, rates for 'SCX-PMU', 'GEL-PMU' and prediction are in black, red and green, respectively. The blue lines are the precision predicted with the first 100 simulated isotopic clusters. The difference between the green and blue

dashed lines in the same figure is due to sampling error, i.e. small number of isotopic clusters in each region. Figure 9a displays the results for the region with mass in 1000–1400 Da and intensity in 500–800. The predicted TPR and precision in green closely match with those from 'SCX-PMU' and 'GEL-PMU'. Note that the predicted TPR and precision are somewhat more optimistic than the actual rates, because the multinomial and Poisson processes are inherent parts of the ion statistics, and therefore represent the lower noise limit. Figure 9b for the region with mass in 1400–1900 Da and intensity in 500–800 is similar to Figure 9a. Figure 9c and d are for the lower intensity region of 100–350. The differences between the green and the blue dashed lines in each figure serve as examples of the sampling error. It is largely due to sampling errors that the precision from 'SCX-PMU' is lower than the prediction by ~0.2 in Figure 9d, and occasionally the actual rates are higher than the predicted rates. The corresponding precision–recall curves are in Figure 9e–f.

Results in Figure 9 show that the noise model can be used to estimate *r* cutoffs given an error rate. Instead of choosing the cutoffs arbitrarily or improperly, having the appropriate cutoffs would minimize the error rates for optimal deisotoping results. For example, to reach 90% precision, the required *r* cutoffs for the four mass and intensity regions in Figure 9a–d are 0.91, 0.95, 0.90 and 1.00, respectively, which can be directly read from the dashed green lines. Alternatively, given an *r* cutoff for a mass and intensity region, the TPR and precision can be estimated. For instance, the solid green line in Figure 9d shows that the TPR is only 50% at the *r* cutoff of 0.968 for that region. In addition, when precision is over 90%, TPR is almost zero for the same region, as shown in Figure 9f. These would be difficult to estimate without a noise model, or with a Gaussian noise model that assumes noise is equal for all peaks. A Poisson noise model or multinomial model alone would underestimate the noise. Note that even though we use *r* to demonstrate the usefulness of the noise model, the approach should also work for other similarity measures between OID and EID in deisotoping because the similarity is directly affected by the noise level.

An actual example of how the noise model improves deisotoping is shown in Figure 10 for an isotopic cluster from 'GEL'. The cluster has a mass of 1149.58 Da, an intensity of 230 and an *r*-value of 0.93. The traditional method fails to recognize it as a peptide at the precision of 90% because the *r* is less than the cutoff of $r = 0.97$, found from the 'SCX-PMU' dataset at 90% precision. In practice, the traditional method often uses arbitrary cutoffs for lack of annotated datasets. With the noise model-based method, which predicts a cutoff of 0.90 at the same precision (shown as the dashed green line in Fig. 9c), the cluster is recognized as a peptide. It makes sense because at the low intensity of 230, the deviation from the expected isotopic distribution is expected to be larger than that when the intensity is strong. The deviation is mass and intensity specific, and is captured by the noise model.

3.4 Other applications of noise models

The noise models can also predict ultimate limits on the reproducibility of LC-MS experiments. The reproducibility

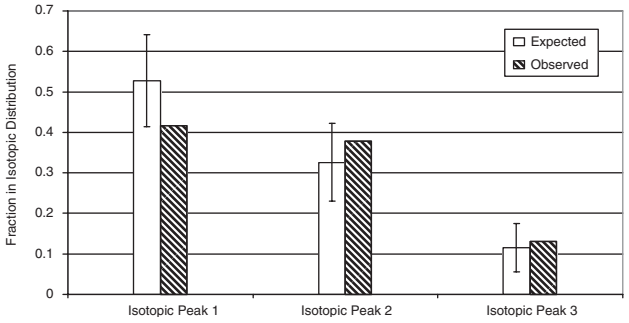

Fig. 10. An illustrative example of an observed peptide isotopic distribution compared to the expected distribution. Error bars represent 95% confidence interval, or 1.96 SD. SD is calculated from the multinomial and Poisson noise model, and normalized by the total intensity. Error on the first peak is 1.96 SD, which is somewhat large but is expected because the multinomial and Poisson processes are inherent parts of the ion statistics and therefore represent the lower noise limit. The peptide sequence is identified as VFSQQADLSR with a MASCOT score of 74, with a mass of 1149.58 Da and an intensity of 230.

of LC-MS has been measured using coefficient of variation (CV), which is defined as the ratio of SD to the mean. According to the described noise model, the CV for Q-TOF spectra is ultimately limited by the multinomial sampling error and by the Poisson noise. The limitation cannot be removed by other steps in the experiments because the noise is inherent. Additionally, the noise models predict that the elution time apex of the same peak may fluctuate among replicate runs, which puts an ultimate limit on the reproducibility of retention times. Because of the inherent noise, isotopic peaks of the same peptide may not have the same elution time apex. This is important because a cluster of peaks with the same elution time apex is often used as the criterion to look for peptide features in LC-MS (Du *et al*., 2007; Wang *et al*., 2003).

The noise models can also be used to guide statistical tests to find peaks or peptides that are differentially expressed between biological samples, based on measured intensities. Such 'label-free' approaches are frequently used in biomarker discovery. In a significance test such as a *t*-test, SD is estimated from the data, which takes away one degree of freedom and may not be a reliable estimate. Noise predicted from a noise model may be used instead of the SD, or as priors for estimating the SD of intensities.

4 CONCLUSION

We have developed a model that characterizes Q-TOF noise with a multinomial and Poisson model. This model explains most of the observed noise, whereas a only multinomial model proved to be inadequate. In addition, dead-time correction significantly improves the fit at high-intensity regions of Q-TOF data. We also find that the ion trap instrument has larger noise than Q-TOF, and the ion trap noise is roughly proportional to Poisson noise, consistent with previous reports for other ion trap instruments. We demonstrate that the noise model can be used to improve deisotoping for peptide detection, by estimating appropriate cutoffs of the goodness

of fit parameter at the prescribed error rates. Our findings also have implications in noise reduction and in LC-MS data analysis for biomarker discovery.

ACKNOWLEDGEMENTS

We thank SCIEX/AB for prompt technical support. Work related to data processing and data analysis in the Analytical Biochemistry group (Groningen) is supported by the following grants: the Dutch Cancer Fund-KWF (RUG 2004-3165), the Netherlands Proteomics Centre (Bsik 3015) and the Netherlands Bioinformatics Centre (BioRange 2.2.3).

Conflict of Interest: none declared.

REFERENCES

Anderle,M. *et al*. (2004) Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics*, **20**, 3575–3582.

Beavis,R.C. (1993) Chemical mass of carbon in proteins. *Anal. Chem*., **65**, 2.

Bellew,M. *et al*. (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, **22**, 1902–1909.

Blackler,A.R. *et al*. (2006) Quantitative comparison of proteomic data quality between a 2D and 3D quadrupole ion trap. *Anal. Chem*., **78**, 1337–1344.

Chernushevich,I.V. *et al*. (2001) An introduction to quadrupole-time-of-flight mass spectrometry. *J. Mass Spectrom*., **36**, 849–865.

Du,P.C. and Angeletti,R.H. (2006) Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Anal. Chem*., **78**, 3385–3392.

Du,P. *et al*. (2007) Data reduction of isotope-resolved LC-MS spectra. *Bioinformatics*, **23**, 1394–1400.

Higgs,R.E. *et al*. (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res*., **4**, 1442–1450.

Horn,D.M. *et al*. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom*., **11**, 320–332.

Horvatovich,P.L. *et al*. (2007) Chip-LC-MS for label-free profiling of human serum. *Electrophoresis*, **28**, 4493–4505.

Kaur,P. and O'Connor,P.B. (2004) Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment. *Anal. Chem*., **76**, 2756–2762.

Kaur,P. and O'Connor,P.B. (2007) Quantitative determination of isotope ratios from experimental isotopic distributions. *Anal. Chem*., **79**, 1198–1204.

Kullback,S. and Leibler,R. (1951) On information and sufficiency. *Ann. Math. Stat*., **22**, 8.

Leptos,K.C. *et al*. (2006) MapQuant: Open-source software for large-scale protein quantification. *Proteomics*, **6**, 1770–1782.

Li,Q.H. *et al*. (2006) Analysis of the stochastic variation in LTQ single scan mass spectra, Rapid Commun. *Mass Spectrom*., **20**, 1551–1557.

MacCoss,M.J. *et al*. (2001) Evaluation and optimization of ion–current ratio measurements by selected-ion-monitoring mass spectrometry. *Anal. Chem*., **73**, 2976–2984.

Omenn,G.S. *et al*. (2005) Overview of the HUPO Plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, **5**, 3226–3245.

Senko,M.W. *et al*. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom*., **6**, 229–233.

Shin,H. *et al*. (2004) *Towards a Noise Model of MALDI TOF Spectra*. American Association for Cancer Research (AACR) Advances in Proteomics in Cancer Research, Waikoloa, Hawaii.

Tu,Y. *et al*. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl Acad. Sci. USA*, **99**, 14031–14036.

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

Weng,L. *et al*. (2006) Rosetta error model for gene expression analysis. *Bioinformatics*, **22**, 1111–1121.

Yergey,J.A. (1983) A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys*., **52**, 13.