Pacific Symposium on Biocomputing 12:421-432(2007)

MINING TANDEM MASS SPECTRAL DATA TO DEVELOP A MORE ACCURATE MASS ERROR MODEL FOR PEPTIDE IDENTIFICATION

YAN FU[1,2†], WEN GAO[3], SIMIN HE[1], RUIXIANG SUN[1], HU ZHOU[4], RONG ZENG[4]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China,
[2]Graduate University of Chinese Academy of Sciences, Beijing, China,
[3]Peking University, Beijing, China,
[4]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

The assumption on the mass error distribution of fragment ions plays a crucial role in peptide identification by tandem mass spectra. Previous mass error models are the simplistic uniform or normal distribution with empirically set parameter values. In this paper, we propose a more accurate mass error model, namely conditional normal model, and an iterative parameter learning algorithm. The new model is based on two important observations on the mass error distribution, i.e. the linearity between the mean of mass error and the ion mass, and the log-log linearity between the standard deviation of mass error and the peak intensity. To our knowledge, the latter quantitative relationship has never been reported before. Experimental results demonstrate the effectiveness of our approach in accurately quantifying the mass error distribution and the ability of the new model to improve the accuracy of peptide identification.

1. Introduction

Tandem mass spectrometry is playing an increasingly important role in current proteomics research [1]. In an experiment of tandem mass spectrometry, peptides digested from protein mixture are first protonized and isolated according to their mass-to-charge ratios. Peptide ions of a specific mass-to-charge ratio then undergo the low-energy collision-induced dissociation to break into fragment ions. Fragment ions are detected and their masses (or rather mass-to-charge ratios) and intensities are recorded. The ion intensity at one mass value forms an observed mass peak. All the mass peaks corresponding to the detected fragment ions of the same peptide constitute the experimental tandem mass spectrum of this peptide.

To identify the peptide corresponding to a tandem mass spectrum, database searching is the most widely used approach. Popular database searching tools are SEQUEST [2] and Mascot [3]. Another approach is the de novo sequencing, e.g. the Lutefisk [4], PEAKS [5] and PepNovo [6] algorithms. A third approach is the sequence tag query, e.g. the pioneer work by Mann and Wilm [7], and the recent GutenTag [8] and Popitam [9] algorithms.

[†] To whom correspondence should be addressed. E-mail: yfu@ict.ac.cn.

A key ingredient of peptide identification algorithms is the scoring function that measures the likelihood of a candidate peptide producing the experimental spectrum. In a peptide-scoring algorithm, observed mass peaks in the experimental spectrum are matched to the fragment ions predicted from a candidate peptide according to their mass values. Due to the imprecision of mass measurement, an error window on mass values is commonly used to tolerate mass match errors in a certain range. The error window plays a very important role in peptide-scoring algorithms. An error window inconsistent with the actual mass error distribution can lead to increased random matches or reduced true matches, thus degrading the performance of a peptide-scoring algorithm. Moreover, in the de novo or sequence tag approach to peptide identification, the allowed maximal mass error can greatly affect the number of candidate peptides or sequence tags.

Ion trap mass spectrometers have been quite attractive in proteomics research, due to their relatively high sensitivity and low cost. However, compared to higher-resolution mass spectra, such as the Q-TOF spectra, the mass error of ion trap spectra is in general much larger and is less exploited in the computational proteomics area. Therefore, this paper focuses on ion trap spectra. The mass error models assumed for ion trap spectra in current peptide identification algorithms are quite simple. The most common assumption is that the mass error is uniformly distributed within the $\pm\varepsilon$ error window around the theoretical mass value [2, 3, 6, 10-18]. For ion trap spectra, the width of error window ε is often empirically set to 0.5 u, e.g. [6, 11, 12]. Another assumption is the normal distribution of mass error [19-22].

Previous mass error models used in ion-trap spectra analysis and peptide identification algorithms can be characterized as follows:

1. The mass error is centered at zero,
2. The mass error is independent of both the mass and the intensity of fragment ions;
3. The parameters in the mass error distribution are empirically set.

A notable exception to (1) and (3) is the recent work due to Wan and Chen [21], in which the mean and standard deviation of the normally distributed mass errors are learned from training data. However, all existing error models have assumed so far that all mass errors in a given dataset of spectra come from an identical distribution regardless of ion masses and intensities. Although peptide-identification tools based on these simple error models often work well, a large proportion (eighty to ninety percentage) of spectra cannot be successfully interpreted in current proteomic experiments due to either known or unknown reasons. A mass error model lacking of enough accuracy has to be responsible for some of these un-interpreted spectra.

In this paper, we statistically investigate the distribution of mass errors of singly charged fragment ions in ion trap tandem mass spectra. By visualizing mass errors in various ways, we first illustrate that there is a linear correlation between the mass error and the ion mass, and there is an approximate log-log linearity between the standard deviation (SD) of mass error and the peak intensity. To our knowledge, the latter quantitative relationship has never been reported in the literature. Based on these observations, we model the mass error of a fragment ion by a conditional normal distribution, whose mean and SD are the functions of ion mass and peak intensity, respectively. We also propose an iterative algorithm, named PMED, to accurately estimate the parameter values in the conditional mean and SD functions. Experimental results demonstrate that the PMED algorithm converges very fast and the learned parameter values match real data very well. Experiment also shows that the new mass error model can considerably improve the accuracy of peptide identification.

The rest of the paper is organized as follows. Section 2 describes the used datasets of tandem mass spectra. In Section 3, we first qualitatively illustrate the distribution trends of mass errors and then propose the conditional normal model of mass error. The iterative parameter learning algorithm, PMED, is presented in Section 4. Section 5 gives experimental results. We finally conclude the paper and point out future work in Section 6.

2. Datasets

We have analyzed several datasets of ion trap mass spectra. However, due to the limited space, we report results on our own dataset in this paper. Results on several published datasets [23-26] are given in Supplementary Information online (http://www.jdl.ac.cn/user/yfu/pmed/index.html).

The steps to generate our dataset (denoted by SIBS dataset) are briefly described below. A total of 300 μg protein sample from whole-cell lysate of mouse liver were digested with trypsin. Five LC-MS/MS runs were performed on the digested mixture with a linear ion trap (Thermo Finnigan, San Jose, CA) using different concentrations in salt steps. The mass spectrometer was set so that one full MS scan was followed by ten MS/MS scans on the ten most intense ions from the MS spectrum. The acquired spectra were searched against the mouse database (SwissProt) using the SEQUEST program. The resulting assignments of database peptides to experimental spectra were filtered according to their Xcorr and DeltCn scores (Xcorr≥1.9 and 2.2 for [M+1] and [M+2] spectra, respectively, and DelCN≥0.1). In addition, to reduce duplicate peptides, only the spectrum of the largest Xcorr was retained among a certain number of consecutive MS/MS scans on the same peptide ion. This finally resulted in a total of 1,505 [M+1] and [M+2] spectra with high-confidence peptide assignments. [M+3] spectra were not included, since doubly charged fragment ions are often dominant in these spectra while our analysis focuses on the mass error of singly charged fragment ions.

3. Conditional Distribution of Mass Error

Our purpose is to study how the mass errors are distributed. Especially, we are interested in whether the mass error correlates with the ion mass and the peak intensity.

3.1. *Visualization of Mass Error Distribution*

To visualize and analyze the mass error, the mass peak produced by each expected fragment ion must be identified in advance. To this end, we first use a common strategy to match observed peaks to expected fragment ions - the most intense peak within the error window of $\pm\varepsilon$ around the theoretical mass value of an expected fragment ion is assigned to this fragment ion. This criterion for determining peak-ion matches certainly lacks accuracy, since the error window is set empirically and is fixed for all the fragment ions, regardless of their masses and intensities. Fortunately, we find that the training data obtained with the above criterion are adequate already for the qualitative analysis of mass error at this stage. In the next section, we will develop an iterative learning algorithm, based on the observations in this section, to quantify the mass error distribution and revise the criteria for determining peak-ion matches.

We use monoisotopic masses of amino acid residues to calculate the theoretical mass and set ε to 0.5 u. Without loss of generality, we illustrate the analysis results only for y ions in this paper. Results for other fragment ion types, e.g. b ions, show a similar trend to y ions and are not given. Figure 1 gives the frequency histogram of mass errors. It shows that the mass error has a bell-shaped distribution. A similar trend is also observed on other datasets (See Figures S1-1, S2-1, S3-1 and S4-1 in Supplementary Information). In addition, depending on the instrument calibration, the center of the mass error distribution may deviate from zero (See Figure S3-1 for example in Supplementary Information). This is called systematic error.

Figure 2 plots all the mass errors of y ions against their corresponding ion masses. From Figure 2, we can see that the mass errors display a trend of descending linearly with increasing ion masses, although, at a given value of ion mass, the mass errors spread fairly abroad. For well calibrated instruments, such a phenomenon may not be apparent (See Figure S4-2 in Supplementary Information for example). However, we did observe the linear relationship between the mass error and the ion mass on several real datasets (See Figures S1-2, S2-2 and S3-2 in Supplementary Information). This relationship has rarely been taken into account by peptide identification algorithms.

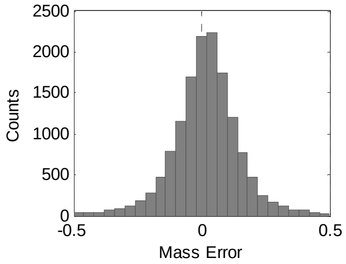

Figure. 1. Frequency histogram of mass errors.

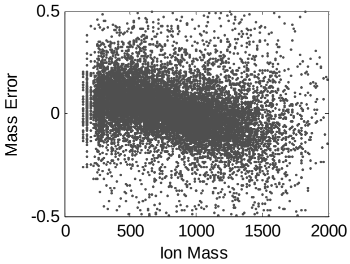

Figure. 2. Mass errors versus ion masses.

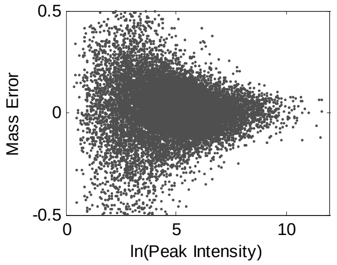

Figure. 3. Mass errors versus the logarithms of their corresponding peak intensities.

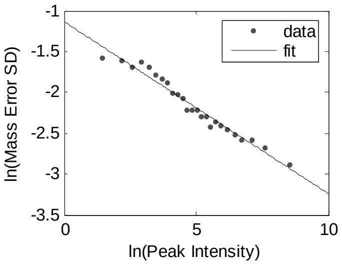

Figure. 4. Log-log plot of the standard deviation (SD) of mass errors versus the peak intensity.

Figure 3 plots all the mass errors of y ions against the logarithms of their corresponding peak intensities. Raw intensities are used here. It shows that the mass error distribution is dramatically correlated with the peak intensity. The more intense the peaks are, the more concentrated the mass errors tend to be. This is intuitively understandable - the more ions are detected, the more accurate the measured mass value should be.

Further analysis reveals that the logarithm of the standard deviation (SD) of mass errors goes down approximately linearly as the logarithm of the peak intensity increases (see Figure 4). The data in Figure 4 are obtained by grouping the intensities into a number of bins and calculating the mass error SD of fragment ions falling in each bin (sampling equal number of data points from each bin results in the same phenomenon). On other datasets, a similar trend is also observed (See Figures S1-3, S1-4, S2-3, S2-4, S3-3, S3-4, S4-3 and S4-4 in Supplementary Information).

To the best of our knowledge, the above quantitative relationship between the mass error and the peak intensity has never been discussed in the literature. It provides a new important constraint for determining peak-ion matches; that is, scaled rather than fixed size of error window should be used for peaks of different intensities.

3.2. *Conditional Normal Model of Mass Error*

Based on the above observations, we model the mass error E_f of a fragment ion *f* as a random variable following a normal distribution, whose mean and SD are determined by the theoretical mass value *M*(*f*) and the observed raw peak intensity *I*(*f*), respectively; that is,

$$E_f \sim n\left(\mu\left(M(f)\right), \sigma^2\left(I(f)\right)\right), \tag{1}$$

where

$$\mu\left(M(f)\right) = u \cdot M(f) + v\ , \tag{2}$$

$$\sigma\left(I(f)\right) = b \cdot \left(I(f)\right)^a, \tag{3}$$

and *u*, *v*, *a* and *b* are parameters to be determined. The conditional mean and SD functions (2) and (3) directly follow from the observations in Section 3.1 (Figures 2 and 4). Notice that by fixing the value of parameter *u* (or *a*) at zero, the mass error mean (or SD) becomes unconditional on the ion mass (or peak intensity), which leads to variant formats of the conditional normal model.

4. Iterative Parameter Learning Algorithm

Generally speaking, given a dataset of spectra of known peptide sequences, the parameter values in the conditional mean and SD functions can be roughly learned from the mass-error data generated in Section 3. However, since such training data are derived from a less accurate peak-ion matching criterion, the accuracy of parameter estimation could be accordingly affected. This problem may become significantly serious, when actual mass errors are distributed out of the expected range. To overcome this difficulty, we develop an iterative algorithm for more accurate parameter estimation.

Intuitively, the algorithm, which we name PMED (Peaks' Mass Error Model), for learning the values of parameters, *u*, *v*, *a* and *b* is performed in the following iterative manner. In one step, according to the mass error distribution determined by the current parameter values, probable peak-ion matches are selected to generate a training dataset. In another step, the parameters are re-estimated on this training dataset. These two steps are carried out alternately until the learned parameter values do not change any more. By this iterative procedure, the learned parameter values are expected not to be sensitive to the prior assumption on mass error distribution.

Let {<S_1, P_1>, <S_2, P_2>, …, <S_N, P_N>} denote a set of tandem mass spectra labeled with corresponding peptide sequences. A spectrum S_i is a set of peaks, $\left\{s_{i1}, s_{i2}, \ldots, s_{im_i}\right\}$, each associated with a mass value $M(s_{ij})$ and an intensity value $I(s_{ij})$. A peptide P_i is a sequence of amino acid residues. Let $\left\{f_{i1}, f_{i2}, \ldots, f_{in_i}\right\}$ denote the set of expected fragment ions of peptide P_i, each associated with a theoretical mass value $M(f_{ik})$.

Algorithm PMED

Input: A set of tandem mass spectra labeled with peptide sequences, {<S_1, P_1>, <S_2, P_2>, …, <S_N, P_N>}.

Output: Estimated parameter values in the mass error distribution i.e. u, v, a and b in Equations (2) and (3).

Step 1. Initialize the values of u, v, a and b in the conditional mean and SD functions (Equations (2) and (3)), according to the prior knowledge about the mass error distribution.

Step 2. For each possible combination of i, j and k, compute the z-score z_{ijk} of the mass error of fragment ion f_{ik}, under the assumption that f_{ik} produced peak s_{ij}, based on the current values of u, v, a and b:

$$z_{ijk} = \frac{\left(M\left(s_{ij}\right) - M\left(f_{ik}\right)\right) - \mu\left(M\left(f_{ik}\right)\right)}{\sigma\left(I\left(s_{ij}\right)\right)}, \tag{4}$$

where $\mu(M(f))$ and $\sigma(I(s))$ are as defined in Equations (2) and (3), respectively.

Step 3. Generate the training dataset D by selecting those peak-ion matches whose absolute z-scores are smaller than a given threshold z_t:

$$D = \left\{\left(s_{ij}, f_{ik}\right) \middle| \left|z_{ijk}\right| < z_t\right\}. \tag{5}$$

Step 4. Update the values of u, v, a and b with the maximum likelihood (ML) estimates for them based on the training dataset D; that is,

$$\langle u, v, a, b\rangle \leftarrow \underset{\langle u,v,a,b\rangle}{\arg\max} \prod_{\left(s_{ij}, f_{ik}\right)\in D} p_{ijk}, \tag{6}$$

where

$$p_{ijk} = \frac{1}{\sqrt{2\pi}\sigma\left(I\left(s_{ij}\right)\right)} \exp\left(\frac{\left|\left(M\left(s_{ij}\right) - M\left(f_{ik}\right)\right) - \mu\left(M\left(f_{ik}\right)\right)\right|^2}{-2\left(\sigma\left(I\left(s_{ij}\right)\right)\right)^2}\right). \tag{7}$$

Step 5. Terminate the algorithm and return the learned values of u, v, a and b if they remain stable; otherwise, go to Step 2.

In Step 2, the absolute value of a z-score (or standard score) measures the deviation of a mass error from its expected value under the current assumption about the conditional mean and SD of the mass error distribution. If a peak-ion match is not due to chance, the deviation should be within a reasonable range. In Step 4, the ML estimates for *u*, *v*, *a* and *b* are not analytically tractable but can be numerically resolved efficiently. During the learning process, the parameters *u* and/or *a* can be fixed at zero to obtain variant versions of the error model. We implemented the PMED algorithm in MATLAB.

The PMED algorithm, as an iterative ML estimator, was inspired by the Expectation-Maximization (EM) algorithm. Although it is not a rigorous EM algorithm, experiments in the next section demonstrate its convergence.

5. Results and Discussions

The results given in this section are obtained on the SIBS dataset described in Section 2. Those obtained on other datasets are presented in Supplementary Information online.

5.1. *Parameter Learning*

The values of *u*, *v* and *a* are initialized to zero, and the value of *b* is initialized to 0.1. These initialized parameter values reflect the weakest prior assumptions about the mass error distribution - centered at zero (*v*=0), independent of the ion mass (*u*=0), and independent of peak intensity (*a*=0). Such assumptions are most common in current peptide identification algorithms.

Figure 5 depicts the learned values for *u*, *v*, *a* and *b* against the number of iterations. The learning process converges after four iterations and takes less than one minute. We made small changes to the initialized parameter values and found that the learning results are not sensitive to the initialized values. The learned results are also quite stable, when the z-score threshold z_t in Equation (5) is set to about three.

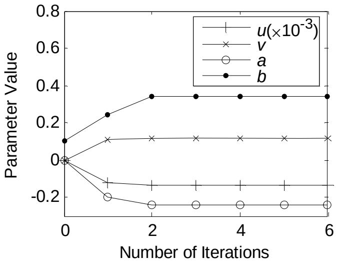

Figure. 5. Learned parameter values versus the number of iterations

It is also shown in Figure 5 that the learned parameter values after the first iteration, which correspond to the direct ML estimates on the initial training data, are significantly different from the finally learned values. This justifies the necessity of the PMED algorithm.

Figures 6 and 7 plot the learned conditional mass error mean and SD respectively. We can see that the learned results of the PMED algorithm are quite consistent with real data. In the case that the parameters *u* and/or *a* are fixed at zero, other parameters can also be accurately learned (results not given).

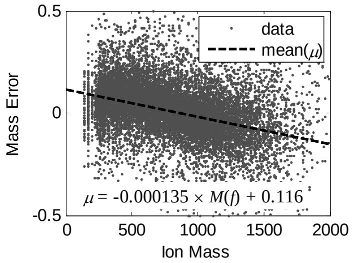

Figure. 6. Learned conditional mean (dashed line) of the mass error distribution, plotted together with mass error data (dots)

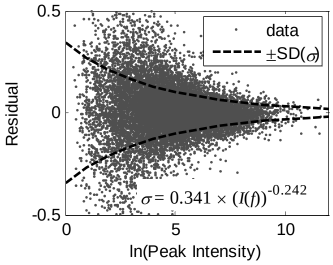

Figure. 7. Learned conditional standard deviation (dashed curve) of the mass error distribution, plotted together with residuals (dots) of mass errors from the learned mean.

Due to the differences in instrument type, setup and calibration, the learned parameter values vary with instruments (See Figures S1-6, S1-7, S21-6, S2-7, S3-6, S3-7, S4-6 and S4-7 in Supplementary Information).

5.2. *Application to Peptide Identification*

To test the usefulness of our proposed conditional normal model of mass error for improving the accuracy of peptide identification, we use a simple peptide-scoring function defined on mass match errors. Given a mass error model, the score of a candidate peptide is the sum of probability densities of all mass match errors. This scoring function is in fact a weighted version of the SPC (Shared Peak Counts) with each peak-ion match weighted by the probability density of the corresponding mass match error. Notice that the high intensity of a matched peak does not necessarily mean a high score. In fact, the situation can be the contrary if the mass match error (residual from learned mean) is large. This is illustrated in Figure 8.

Several mass error models are compared, including the uniform distribution, the normal distribution, and several variants of conditional normal distribution. Parameter values in each model are either set empirically or learned from data.

In the latter case, five-fold cross validation is used for performance evaluation. Further, when parameters are learned from data, they may be either conditional or fixed.

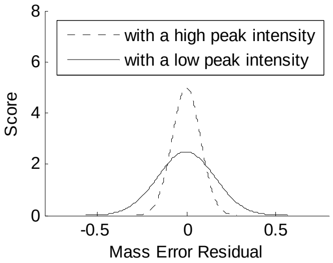

Figure. 8. The score of a match is determined by both the mass match error and the peak intensity.

Spectra are searched using the pFind program [17, 27, 28] against a large database containing 127,432 protein sequences (SwissProt database of all species entries, appended with the peptide sequences of test spectra). For simplicity, b and y ion series are predicted. Trypsin is used for theoretical digestion with up to two missed cleavage sites allowed.

Table 1 compares the search results on the SIBS dataset using the above defined peptide-scoring function equipped with various mass error models. Percentages of spectra with the correct peptide sequence ranked top one and top ten are used to measure the identification accuracy.

Table 1. Comparison of search results with various mass error models

Mass error model	Parameters		Top1(%)[1]	Top10(%)[2]
	μ (mean)	σ (SD) or ε (width)		
Uniform	0	$\varepsilon = 0.3$	87.0	95.9
		$\varepsilon = 0.5$	75.6	90.1
		$\varepsilon = 0.7$	60.0	78.8
Normal	0	$\sigma = 0.3/z_t$	73.2	90.4
		$\sigma = 0.5/z_t$	87.1	97.3
		$\sigma = 0.7/z_t$	90.2	98.3
	Fixed/learned	Fixed/learned	90.2	98.3
		Conditional	97.8	99.7
	Conditional	Fixed/learned	91.6	98.4
		Conditional	98.1	99.7

[1] Percentage of spectra with the correct peptide sequence ranked top one.
[2] Percentage of spectra with the correct peptide sequence ranked top ten.

It is shown in Table 1 that compared to the unconditional uniform and normal models, either the introduction of conditional mean or the introduction of condition SD can independently increase the identification accuracy. The improvement caused by the introduction of conditional SD is particularly significant - seven percentage points in Top1 performance. The best results are obtained with the fully conditional normal model. On the Top10 performance, the conditional normal model is also superior to unconditional models. On other datasets, the increases are remarkable too (See Tables S1, S2, S3 and S4 in Supplementary Information).

6. Conclusions

The proposed mass error model and the associated parameter learning algorithm provide an automated method for quantifying the mass error distribution of fragment ions in ion trap tandem mass spectra. Compared to previous mass error models, the new model has several advantages:

1. Systematic error of mass measurement is taken into account;
2. Fragment ions of different masses and intensities are of different mass error distributions;
3. Parameters can be automatically learned from data.

Experiments demonstrated the effectiveness of the parameter learning algorithm and the usefulness of the new mass error model for peptide identification. In the future, we expect to develop more sophisticated peptide scoring functions to take full advantage of the new mass error model.

The analysis in this paper is limited to singly charged fragment ions. Due to the disturbance of isotopic peaks in the low-resolution ion trap spectra, the mass errors of doubly charged fragment ions are more complex and the analysis of them is more challenging and will be our future work.

Acknowledgments

This work was supported by the National Key Basic R&D Program of China (Grant No.: 2002CB713807) and National Key Technologies R&D Program of China (Grant No.: 2004BA711A21). We gratefully acknowledge Dr. Andrew Keller from the Institute for Systems Biology for valuable comments on an early version of the paper. We also thank Prof. Runsheng Chen, Dr. Dongbo Bu, Jingfen Zhang, Quanhu Sheng, Jie Dai and many others from Chinese Academy of Sciences for helpful discussions.

References

1. R. Aebersold and M. Mann, *Nature* **422**, 198 (2003).

2. J. K. Eng, A. L. McCormack and J. R. Yates, III, *J. Am. Soc. Mass. Spectrom.* **5**, 976 (1994).
3. D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell, *Electrophoresis* **20**, 3551 (1999).
4. J. A. Taylor and R. S. Johnson, *Anal. Chem.* **73**, 2594 (2001).
5. B. Ma, K. Z. Zhang, C. Hendrie, C. Z. Liang, M. Li, A. Doherty-Kirby and G. Lajoie, *Rapid Commun. Mass Spectrom.* **17**, 2337 (2003).
6. A. Frank and P. Pevzner, *Anal. Chem.* **77**, 964 (2005).
7. M. Mann and M. Wilm, *Anal. Chem.* **66**, 4390 (1994).
8. D. L. Tabb, A. Saraf and J. R. Yates, III, *Anal. Chem.* **75**, 6415 (2003).
9. P. Hernandez, R. Gras, J. Frey and R. D. Appel, *Proteomics* **3**, 870 (2003).
10. D. Fenyo, J. Qin and B. T. Chait, *Electrophoresis* **19**, 998 (1998).
11. V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath and P. A. Pevzner, *J. Comput. Biol.* **6**, 327 (1999).
12. M. Havilio, Y. Haddad and Z. Smilansky, *Anal. Chem.* **75**, 435 (2003).
13. R. G. Sadygov and J. R. Yates, III, *Anal. Chem.* **75**, 3792 (2003).
14. J. Colinge, A. Masselot, M. Giron, T. Dessingy and J. Magnin, *Proteomics* **3**, 1454 (2003).
15. D. L. Tabb, L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin and J. R. Yates, III, *Anal. Chem.* **75**, 1155 (2003).
16. J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth and S. P. Gygi, *Nat. Biotechnol.* **22**, 214 (2004).
17. Y. Fu, Q. Yang, R. Sun, D. Li, R. Zeng, C. X. Ling and W. Gao, *Bioinformatics* **20**, 1948 (2004).
18. M. Bern and D. Goldberg, *The Ninth Annual International Conference on Research in Computational Molecular Biology* 357 (2005).
19. V. Bafna and N. Edwards, *Bioinformatics* **17**, S13 (2001).
20. N. Zhang, R. Aebersold and B. Schwikowski, *Proteomics* **2**, 1406 (2002).
21. Y. Wan and T. Chen, *The Ninth Annual International Conference on Research in Computational Molecular Biology* 342 (2005).
22. J. H. Oh and J. Gao, *The 5th IEEE Symposium on Bioinformatics and Bioengineering* 161 (2005).
23. A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett and E. Kolker, *Omics* **6**, 207 (2002).
24. V. Mayya, K. Rezaul, Y. Cong and D. Han, *Molecular & Cellular Proteomics* **4**, 214 (2005).
25. J. Peng, J. E. Elias, C. C. Thoreen, L. J. Licklider and S. P. Gygi, *J. Proteome Res.* **2**, 43 (2003).
26. J. T. Prince, M. W. Carlson, R. Wang, P. Lu and E. M. Marcotte, *Nat. Biotechnol.* **22**, 471 (2004).
27. D. Li, Y. Fu, R. Sun, C. Ling, Y. Wei, H. Zhou, R. Zeng, Q. Yang, S. He and W. Gao, *Bioinformatics* **21**, 3049 (2005).
28. J. Zhang, W. Gao, J. Cai, S. He, R. Zeng and R. Chen, *IEEE/ACM T. Comp. Biol. Bioinfo.* **2**, 217 (2005).