Proteomics 2003, *3,* 1597–1610 DOI 10.1002/pmic.200300486

Marc Gentzel
Thomas Köcher
Saravanan Ponnusamy
Matthias Wilm

European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

Preprocessing of tandem mass spectrometric data to support automatic protein identification

Liquid chromatography tandem mass spectrometry is a major tool for identifying proteins. The fragment spectra of peptides can be interpreted automatically in conjunction with a sequence database search. With the development of powerful automatic search engines, research now focuses on optimizing the result returned from database searches. We present a series of preprocessing steps for fragment spectra to increase the accuracy and specificity of automatic database searches. After processing, the correct amino acid sequences from the database can be related better to the fragment spectra. This increases the sensitivity and reliability of protein identifications, especially with very large genomic databanks, and can be important for the systematic characterization of post-translational modifications.

Keywords: Liquid chromatography / MASCOT / Protein identification / Quadrupole-time of flight / Tandem mass spectrometry PRO 0486

1 Introduction

Biological research relies more than ever before on the accuracy of protein identifications by mass spectrometry [1–5]. Electrospray MS/MS is one of the preferred methods of identifying proteins [6, 7]. The low-energy collision experiments from multiple-charged peptides generate fragment spectra which can often be related to peptide sequences from a database in a relatively reliable way. However, even if the sequence of a fragmented peptide is present in the database its identification can still be problematic. Numerous peptides similar to the one investigated can be present in the database or the fragment spectrum can contain very few and/or weak signals [8]. This is why recent research focuses on optimizing the search engine results, either by modifying the scoring function or by evaluating the results in a sophisticated way in order to distinguish better between correct and probably wrong peptides sequences [9–11]. When digested protein mixtures are analyzed using liquid chromatography electrospray tandem mass spectrometry (LC-ESI MS/MS), a large number of peptides can be investigated in a single experiment [12]. However, in contrast to manual analysis with nano-ESI ion sources, the time spent acquiring a fragment spectrum is preprogrammed and cannot be adjusted to the individual requirements of each peptide. The result is that a relatively large number of fragment spectra have a poorer quality than with manual nano-ESI investigations [13, 14]. This can compromise the accuracy of automatic sequence identifications [15, 16]. HPLC MS/MS experiments cannot be avoided since only they allow the throughput required by many biological experiments. Here, we demonstrate that simple data pretreatment methods can be used to improve the reliability of automatic database identifications using HPLC tandem mass spectrometric datasets.

We have developed a series of preprocessing steps with the expectation that processed spectra can be related to the database entries of the underlying peptides better than the unprocessed spectra can. The search engine that we used to evaluate the efficiency of the approach was MASCOT from Matrix Science (http://www.matrixscience.com) [8, 16, 17]. Our approach is clearly empirical. The precise way in which MASCOT scores peptide sequences has not been published. MASCOT is a high quality search engine, to our knowledge the only one freely accessible *via* the WWW that can use entire LC-MS/MS experiment files. It is one of the most popular search programs for protein identification using this type of data. Despite the empirical nature of our evaluation scheme, we think the approach is justified considering the importance of MASCOT as a research tool for the scientific community, the necessity of trusting the protein identities in face of the large number of identifications made in one experiment and the importance certain proteins may have for directing further biological research. We would have liked to test the effect of our preprocessing on the identification scores of other search engines, but, to our knowledge, there are no other programs available on the WWW which can use LC MS/MS experiment files.

Correspondence: Dr. Matthias Wilm, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany
E-mail: wilm@embl-heidelberg.de
Fax: +49-6221-387-306

Abbreviations: DTE, 1,4-dithioerythritol; **Q-Tof**, quadrupole time of flight

The purpose of preprocessing the data is to correct for imperfections in the automatic acquisition scheme of the mass spectrometer and to adapt the physical data as much as possible to the fragment masses calculated from the peptide sequences. The steps are: (1) centroiding; (2) joining of spectra which were generated from the same peptide but saved independently; (3) automatic calibration; and (4) deisotoping and deconvolution of the charge state. The algorithms used in the steps will be described in some detail. We decided to program the procedures ourselves to have reasonable flexibility and because the programs available to us did not perform sufficiently well. Individual comparisons will be mentioned between our solution and the solutions implemented by commercial software available in our laboratory.

2 Material and methods

2.1 Chemicals

All reagents were of analytical grade or better. Acetonitrile, water, Tris, formic acid, guanidinium-hydrochloride and 1,4-dithioerythritol (DTE) were purchased from Merck (Darmstadt, Germany). Ammonium bicarbonate and iodoacetamide were from Sigma-Aldrich (Deisenhofen, Germany). Trypsin (sequencing grade, modified and unmodified) and Lys-C were from Roche (Mannheim, Germany).

2.2 Sample preparation

In-gel digests were carried out following [18]. Briefly, stained gel pieces were washed three times with water:acetonitrile (1:1). After vacuum drying the gel pieces were rehydrated with 10 mM DTE in 0.1 M ammonium bicarbonate solution and incubated for 45 min at 56°C. Excess liquid was removed and replaced by 55 mM iodoacetamide solution. The alkylation reaction was carried out in the dark for 30 min at room temperature. Reagents were removed by three washes with water:acetonitrile (1:1) and gel pieces were dried *in vacuo*. Gel pieces were swollen with 0.05 M ammonium bicarbonate containing trypsin or Lys-C (12.5 ng/μL) and incubated at 37°C overnight. After recovery of the supernatant the gel pieces were extracted once with 25 mM ammonium bicarbonate, 50% acetonitrile and once with 5% formic acid. Combined supernatants were dried under vaccum. When the protein mixtures were digested in solution the sample was adjusted to 1.0 M guanidium-hydrochloride at pH 8.0, 10% acetonitrile in the original buffer (usually 3–20 mM Tris-HCl at pH 7.5–8.0). Trypsin was added to a concentration of 8.3 ng/μL and the sample was incubated at 37°C. After 24 h the same amount of trypsin or Lys-C was added and the sample was incubated a second time for 12 h.

2.3 Liquid chromatography and mass spectrometry

HPLC separations of peptides were performed on a capillary HPLC system consisting of the Ultimate HPLC system with UV detector, the Switchos II precolumn switching device and the Famos autosampler (LC Packings, Dionex, Idstein, Germany). Samples were loaded onto a C-18 precolumn (0.3 mm × 1.0 mm, PepMap C-18 from LC Packings or YMC C-18 ODS AQ from YMC Europe, Schermbeck, Germany) for desalting and preconcentrated with a flow rate of 20 μL/min in 0.2% formic acid, 2% acetonitrile (solvent A). Peptides were separated on a 75 μm × 15 cm column at 200 nL/min (PepMap C-18 or YMC C-18 ODS AQ) in a gradient of solvent A and 80% acetonitrile, 0.5% formic acid (solvent B).

The HPLC was interfaced on-line to the mass spectrometer (Q-Tof1; Micromass, Manchester, UK). The electrospray was generated from PicoTip 10 μm needles (New Objectives, Woburn, MA, USA) with a spraying voltage between 2800–3500 V using a cone voltage of 40 V at the mass spectrometer. The quadrupole-time of flight machine (Q-TOF) was operated using MassLynx V 3.4 (Micromass). The following parameters set was used for the automatic adjustment of tandem mass spectrometric conditions: selection of double and triple charged precursors, threshold at 10 ion counts/s, scan time 3 s, interscan delay time 0.1 s, detection window 2.5 Da, up to four components selected simultaneously, precursor ion exclusion for 90 s. The collision energy applied was static and dependent on the ions' *m/z* values and their charge states.

2.4 Data processing

LC tandem mass spectrometric data were exported from MassLynx 3.4 with minimal processing. The peptide filter to eliminate poor fragment spectra was switched off. Background subtraction and smoothing were disabled. The mean peak width was set to one channel and peak centroiding was minimized by setting the threshold of peak centroiding to 100% of their height. The dataset was saved as a MASCOT-compatible pkl file in ASCII format. This file was imported into IGOR Pro (WaveMetrics, Lake Oswego, OR, USA) for further processing. Within IGOR Pro four processing steps were performed: (a) fragment spectra originating from the same peptide but acquired as independent datasets were joined; (b) all spectra were centroided by applying a centroiding function that takes the resolution of the mass spectrometer into account; (c) the entire LC MS/MS run was calibrated using fragment spectra from autolysis products of trypsin; (d) all fragment spectra were deisotoped and charge state

deconvoluted. Finally, the fragment spectra were exported preserving the original pkl file structure and submitted to a database search using MASCOT *via* the world wide web (http://www.matrixscience.com).

2.5 Database searches

The fragment spectra were searched against the NCBI nonredundant database without any restrictions in the choice of the organism using MASCOT on-line. Carbamidomethylated cysteines, oxidized methionines, histidines or tryptophans and deamidations were set as variable modifications. Only tryptic peptides were considered. Up to two missed cleavage sites of trypsin in the peptides were allowed. The fragment mass tolerance was set to $\pm$0.15 Da, the precursor mass tolerance to $\pm$150 ppm. The instrument type was set to ESI-QUAD-TOF. To confirm some peptide sequences a sequence tag based approach using PeptideSearch was used (http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html) [19, 20]. Again, only tryptic peptides were considered, allowing up to two missed cleavage sites. Here, cysteine was set to carbamidomethylated cysteine. No other modification was considered. The mass tolerance for fragments and precursors was 150 ppm.

3 Results and discussion

3.1 Data preprocessing

3.1.1 Peak centroiding

The first step in preprocessing the data is peak centroiding. The aim is for every isotope in a mass spectrum to be represented by exactly one data point. The challenge is to do this over the entire dynamic range of the mass spectrum. Very intense and very weak peaks should be centroided in the same way. Large peaks should never be pulled apart into two separate peaks and small peaks should never disappear. When the MASCOT search algorithm determines a score for a peptide a subset of peaks of the entire spectrum is used and is compared to the calculated masses of all the fragment ions considered. The subset of peaks MASCOT selects in various rounds of scoring depends on the intensity of the fragments. However, the intensity threshold for peak picking changes over the spectrum. At the upper end of the spectrum peaks of only one or two ion counts can be selected and if they correspond to y- or b-ions can considerably influence the overall score of the sequence evaluated. This peak picking algorithm is well adapted to the characteristics of fragment spectra. MS/MS spectra are nearly free of noise. Only isolated single counts can be considered noise events. This is why very low intensity peaks in a fragment spectrum are often real and correspond to fragment ions generated with low efficiency. In manual interpretations of spectra they are successfully used to confirm an identified peptide.

To achieve this level of peak centroiding we use a relationship describing the peak width as a function of the *m/z* value of the ion. The resolution of a Q-TOF mass spectrometer is determined mainly by the geometry of the instrument. Therefore, it does not vary over time and the function describing the peak width is constant. For our Q-TOF1 we use the function $pw(\times) = 0.08 + 0.0004 \times$, for the peak width pw and the *m/z* value $\times$. Each spectrum is represented by separate mass (x_i) and intensity (y_i) value sets. It is important to note that the spectrum is not projected onto an equally spaced x-axis. The resolution of a TOF mass spectrometer is so high that the stepwidth of such an x-axis needs to be very small to avoid reducing the accuracy of the measurements. The small stepwidth increases the size of the datasets and requires considerably more processing time. All the algorithms described in this article are written for discrete $(x_i, y_i)_{i=1,\ldots,n}$ datasets to gain processing speed. An additional advantage of these algorithms is that they can be applied easily to spectra acquired with higher resolution mass spectrometers. The calculation time depends on the number of different ions detected, the number of (x_i,y_i) pairs, but not primarily on the resolution of the spectra. For every fragment spectrum $(x_i,y_i)_{i=1,\ldots,n}$ two additional arrays are calculated to support the centroiding procedure, the peak width array (pw_i) and a projected intensity array (t_i) with:

$$t_i = \sum_{\substack{j \text{ with} \\ x_i \le x_j \le (x_j + pw_i)}} y_j \qquad (1)$$

The intensity array (t_i) represents an estimate of the intensity of a centroided peak at every point (x_i,y_i) of the spectrum. It is calculated as a forward projection. Its values tell the programme whether x_i could be the starting point of a rising peak. Using these arrays the peak centroiding becomes a linear process. As soon as the intensity array (t_i) passes a parameterized threshold value t in t_j the maximum of the rising peak is determined so as to position a centroiding window symmetrically around it with the width $pw(x_{max})$ and x_{max} being the x-location of the maximum. A new centroided mass spectrum is built from data points (x_c, y_c), x_c the weighted average of the *m/z* values in the centroided interval and y_c the integrated intensity:

$$y_c = \sum_{x_{max} - pw(x_{max})/2 \le x_i \le (x_{max} + pw(x_{max})/2)} y_i \qquad (2a)$$

and

$$x_c = \frac{\sum_{x_{max} - pw(x_{max})/2 \le x_i \le (x_{max} + pw(x_{max})/2} x_i y_i}{\sum_{x_{max} - pw(x_{max})/2 \le x_i \le (x_{max} + pw(x_{max})/2)} y_i} \qquad (2b)$$

The flexibility of the algorithm, its applicability to intense continuous spectra and, without any change in the parameter settings, partially centroided weak fragment spectra or completely centroided data lies in the way the peak maximum is determined. When the intensity crosses the threshold in x_j, the spacing of adjacent data points is compared with the peak width $pw(x_j)$ to detect whether this dataset has already been centroided using this algorithm. If so, the maximum in the window (x_j, $x_j + pw(x_j)/2$) is considered the peak maximum. Otherwise, the peak maximum is the intensity maximum in the window (x_j, $x_j + pw(x)$). A simple test is performed to find whether the detected maximum is on a flank of a peak. If not, the peak centroiding procedure starts. The lower border of the centroiding window is compared with the upper border of the preceding centroiding procedure to detect partially overlapping peaks and to split overlapping intensities between the two adjacent peaks.

The advantages of such an algorithm are evident. It uses specific information about the mass spectrometer, its peak width distribution over the mass scale, and can centroid peaks independent of their overall intensity because the first step is the determination of the location of a peak maximum before the centroiding window is positioned. The built-in test of whether the spacing of data points suggests that the spectrum consists of centroided peaks allows it to be applied to centroided data a second time with only minor effects on the data. This feature is used in the spectrum joining and deconvolution algorithm (see Section 3.1.2). Figure 1 shows an example. We compared this algorithm with to centroiding algorithms embedded in commercial applications, BioMultiView (MDS-Sciex, Toronto, Canada) and MassLynx (Micromass). Both programms gave very good, comparable results. However, they cannot be applied to data already centroided without changing their parameters. The LC-tandem mass spectrometric datasets exported from MassLynx in a pkl format are already partially centroided. Our centroiding algorithm is adapted to this kind of data whereas the commercial centroiding algorithms expect continuous data. The result is that individual peaks are sometimes eliminated because they are too narrow.

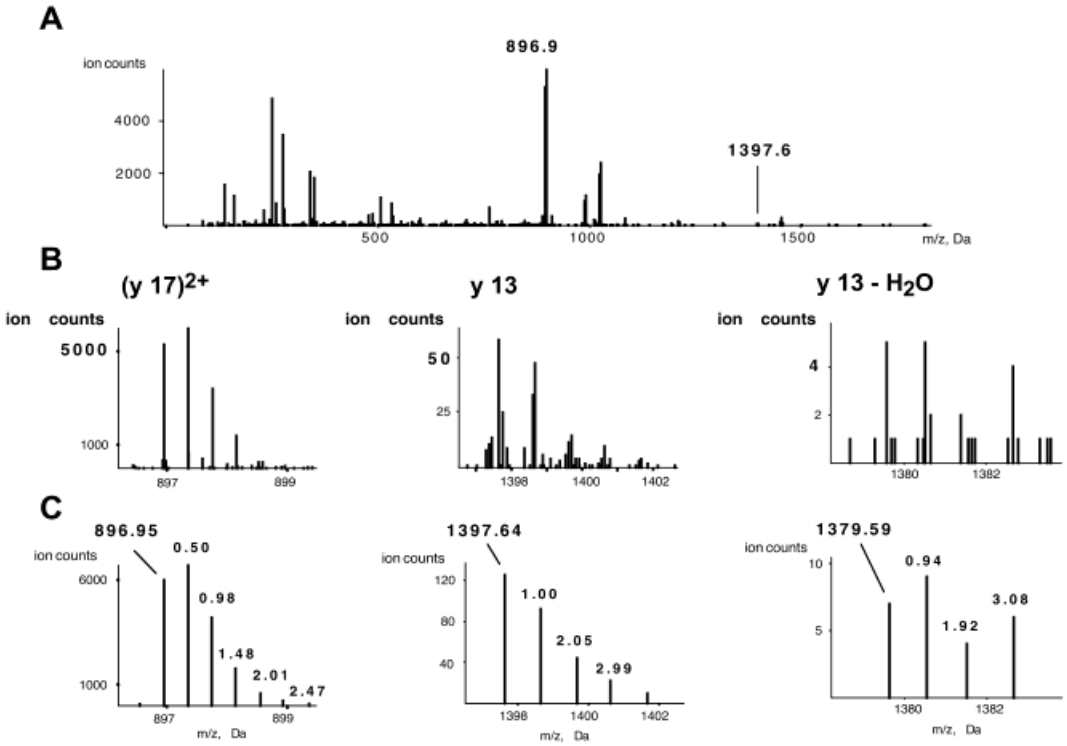

Figure 1. Fragment spectrum of the triple-charged ion of the peptide YIAWPLQGWQATFGGGDHPPK. Panel A shows the complete spectrum, panel B selected ions before centroiding as exported from MassLynx, and panel C the centroided spectrum. The centroiding algorithm correctly centroids the isotopes to a single peak independent of their intensity. Recentroiding of already centroided data does not perturb the spectrum. The same result is achieved when applying the method to more continous data of even higher intensity obtained by directly copying spectra out of MassLynx without using the pkl-export routines.

3.1.2 Spectrum joining

The spectrum joining algorithm identifies and eliminates fragment spectra of identical peptides in the LC-MS/MS experiment file that were acquired independently. The acquisition software MassLynx uses dynamic exclusion to avoid repetitive selection of the same precursor. After a fragment spectrum of a peptide is acquired its *m/z* value is excluded from selection for a programmable period of time. This time usually corresponds to the average elution time of a peptide from the chromatography column. However, some peptides have longer elution times due to their higher than average abundance. They can be selected a second time for fragmentation. In such a case the majority of the peptide has been sprayed already, and in many cases the second fragment spectrum has a much lower quality. By joining it with the more abundant spectrum in the same datafile we reduce the number of dubious low-score identifications when the fragment spectra are compared to peptide sequences.

Every fragment spectrum in the LC-MS/MS file is classified by the mass of its precursor. The algorithm starts by sorting the fragment spectra building groups whose precursor masses are in common intervals of 1.2 Da. Comparisons between the spectra are done only within each group. By this classification, two precursors closer than 1.2 Da can still be in two different, adjacent groups and will therefore never be compared to each other. However, our experience is that since we introduced the spectrum joining routine into our data preprocessing we hardly ever encounter the same peptide identified several times. The spectrum comparisons are relatively calculation intensive, so we did not introduce the more flexible scheme of floating mass intervals for grouping the precursors.

Comparisons between two spectra are done in a series of steps.First, fingerprint spectra are generated. Second, the fingerprint spectra are split into two parts which are compared independently of each other. Two partial spectra are judged to be identical based on specific characteristics of a correlation function between them. The complete spectra are joined if both parts are considered to be identical. The comparison follows simple logical rules in the sense that if spectrum A equals spectrum B and spectrum C, then spectrum B and C are not compared with each other since they are identical. Spectrum comparisons are done by calculating their correlation as a function of a relative mass shift Δm to each other. If the spectra correlate best for a mass shift of 0 Da, then they are considered identical and are joined. A similar correlation algorithm is used by SEQUEST to identify a peptide in an automated database search [15].

The generation of the correlation function $c(\Delta m)$ requires a large number of calculations. The algorithm is optimized to reduce the time that is necessary to calculate $c(\Delta m)$. In a first step the two spectra $(x_i,y_i)_i$ and $(x_i,y_i)_2$ are synchronized in x with a parameterized stepwidth s. Synchronization means that the existing x-values are slightly shifted so that they fall onto multiples of the stepwidth. First we use 0.1 Da as stepwidth but we saw no decline in overall performance when using up to 0.4 Da. The number of calculations required does not depend much on this parameter but only on the number of datapoints in the spectra compared. A larger stepwidth parameter increases the tolerance of shifting mass calibrations.

In a second step, fingerprint spectra are generated from the original spectra. Each spectrum is divided in 100 Da wide windows. For the fingerprint spectrum all peaks higher than the tenth most intense peak in every 100 Da window are considered. The MASCOT search engine seems to follow a similar, but probably more developed, approach when comparing peptide sequences to fragment spectra [16]. The fingerprint spectra are used to calculate the correlation function. The fingerprint spectra usually contain not only fewer data points than the original spectra, which allows faster calculation of the correlation function, but they also seem to be more representative of the spectrum. Using them eliminated false positive comparisons when applying the algorithm to abundant fragment spectra.

The fingerprint spectra were divided into two equal parts for the upper and lower mass ranges. The two complete spectra are considered to be identical only if both parts are identical to their respective counter parts. The splitting of the spectra into two parts is necessary to increase the specificity of the comparisons. The lower mass halves of fragment spectra are sometimes considered identical when the entire spectrum is not. Most of the ion counts are recorded in the lower part, so their similarity can outweigh the differences in the upper part of the spectrum. This problem is effectively avoided by splitting the spectra into two and asking both halves to match independently. The sequence tag algorithm for protein identification inherently makes use of this feature of fragment spectra [19]. A sequence tag is generated from the upper part of the fragment spectrum. Its informational content is so high that it is one of the most specific ways to identify a peptide in a database search.

The decision whether two partial spectra $(x_i,y_i)_1$ and $(x_j,y_j)_2$ are similar enough to be considered as coming from the same peptide is based on the correlation function $c(\Delta m)$:

$$c(\Delta m) = \sum_{\substack{i,\,j \text{ with} \\ (x_{i_1} - \Delta m/2) = (x_{j_2} + \Delta m/2)}} y_{i_1} y_{j_2} \qquad (3)$$

This correlation function is calculated for mass shifts Δm of -25 Da up to 25 Da in steps of 0.5 Da between the two spectra. For a given mass shift it represents the sum of the product of all mass-matched fragment intensities in the two spectra. Figure 2 shows an example. The two spectra are considered identical if the correlation function has its absolute maximum at a mass shift of 0 Da and if this maximum is at least 1.5 times higher than any other of its values outside the Δm interval (–2,2). Within this interval the isotopes of large fragment ions generate a nonrandom positive correlation if the spectra are identical.

We made two surprising observations when testing different correlation functions. First, the factor 1.5 is sufficient to identify two identical spectra and second, the attempt to consider the intensity distribution of peaks in a spectrum was counterproductive. When developing the spectrum joining algorithm we aimed at optimizing the specificity of recognizing identical spectra without compromising the sensitivity. False positives were effectively eliminated by requiring the upper and lower part of the spectrum to match independently. With this condition, the factor 1.5 for the peak maximum at 0 Da relative shift of the correlation function was sufficient. It has to be considered that our datasets are no bigger than 500 fragment spectra. It is reassuring to know that we can easily increase the specificity in spectral comparisons by increasing this factor with larger datasets.

It came as a surprise to us that the consideration of relative peak intensities did not increase specificity, at least not beyond the level we had achieved already, without reducing the sensitivity. When comparing two spectra visually the intuitive approach is to look for peak locations and the relative peak intensities in the two spectra, compare their outline and decide whether they are similar. The visual inspection is probably based more on relative peak intensities than on identical peak locations. We considered matching of peak intensities mathematically by modifying the correlation function. The two spectra were each normalized separately to the average of their five most abundant peaks and the correlation function $c(\Delta m)$ was calculated as:

$$c(\Delta m) = \sum_{\substack{i,\ j \text{ with} \\ (x_{i_1} - \Delta m/2) = (x_{j_2} + \Delta m/2)}} y_{i_1} y_{j_2} e^{-((Y_{i_1} - Y_{j_2})^2 - a)/b} \quad (4)$$

with $(x_i,y_i)_1$ $(x_j,y_j)_2$ the two spectra and Y_{i1} and Y_{j2} the normalized intensity values. The parameters a and b define the maximal positive enhancement in the correlation factor for matching intensities ($Y_{i1} = Y_{j2}$) and the width at which the peak intensity matching factor drops to half its maximal value. The net effect when comparing entire spectra was that we lost several correctly matched spectrum pairs whatever parameters a and b we tried before eliminating false positives. Apparently, the relative peak intensities vary too much from spectrum to spectrum in contrast to the peak locations (see Fig. 2). This is particularly true for low-level fragment spectra when individual isotopes consist of only four ions or sometimes even fewer. These intensities are simply too low to reflect reliably a specific peak intensity even within one LC experiment when the fragmentation conditions and the mass spectrometer remain unchanged. When two spectra are judged to reflect the same peptide the datasets are joined, sorted on *m/z* values and the new spectrum is recentroided. All spectrum joining is documented in a log-file and the original spectra are exported as individual files for documentation purposes before they are combined.

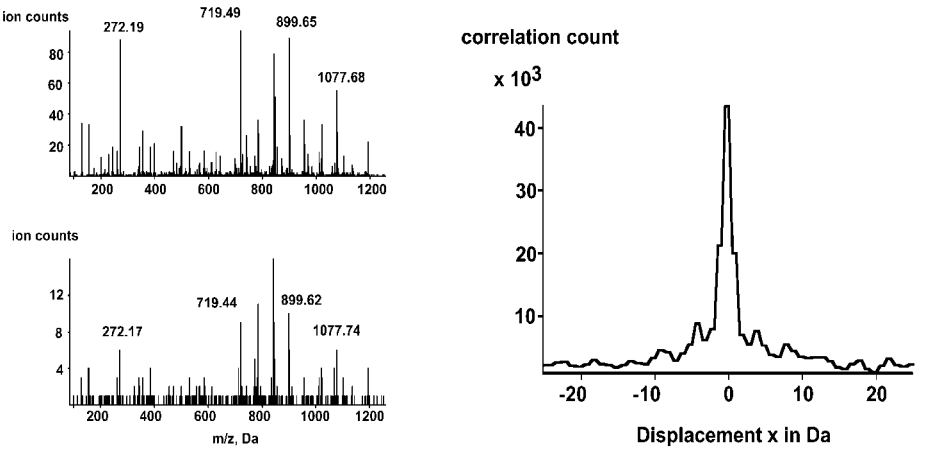

Figure 2. Fragment of the triple-charged ion of the peptide GVDLDQLLDMPNNQLVELMHSR. The two spectra had been identified as being from the same precursor using the correlation approach on the entire spectrum. The correlation function shows a clear and absolute maximum at a relative mass shift of 0 Da. The intensity ratios of the fragments vary considerably between the two spectra, in contrast to the peak locations.

3.1.3 Automatic calibration

Quadrupole-time of flight mass spectrometers generate fragment spectra with very high resolution and high accuracy when correctly calibrated. Our instrument, the Q-Tof1 machine from Micromass, has a nominal mass resolution of m/Δm of 6000, later models have a resolution of 10 000 or more. Most often their accuracy is limited by an imperfect calibration. Temperature shifts in the laboratory of 2°C can be sufficient to limit the accuracy of the mass assignments. High accuracy is important to limit the number of possible candidates in a database search and thereby increase the search specificity. This increase in search specificity is reflected in a decrease in the score threshold value for a relevant sequence match. The threshold value depends on the number of peptide sequences considered as possible hits in the database search and on threshold probability. If the threshold for relevance is set to 0.05, meaning that only 1 out of 20 hits with relevant scores should be a random hit, then the treshold score is calculated as $-10 \cdot \log_{10}(0.05/N)$ where N is the number of sequences within the mass tolerance window. The threshold factor decreases by 10 for every 10-fold decrease of the number of peptides considered [8]. We recalibrate our data during preprocessing in a fully automatic way to become independent of the actual calibration of the mass spectrometer. Fragment spectra of trypsin autolysis products are used as internal standards. Reference fragment spectra and associated mass lists are read in from an external library. Using the same procedures as in the spectrum joining algorithm, matching fragment spectra in the LC-MS/MS dataset are sought for and used for calibration. The measured masses of the first isotopes of all peaks corresponding to the reference mass list are retrieved from the uncalibrated fragment spectra and compared with the reference masses. Mass deviations are plotted against the *m/z* values and a linear fit is calculated. This line is used to recalibrate the entire LC-MS/MS dataset. We could adapt the calibration curve to any polynomial function but linear recalibration was sufficient for our mass spectrometer. The remaining mass deviations are recalculated after recalibration. If some of the masses still show a deviation of more than 0.1 Da from the expected values, they are eliminated from the calibration list and the calibration is repeated. Finally, a report file is generated containing graphical representations of mass deviations against *m/z* values before and after calibration, the number of reference masses used, highest and lowest *m/z* reference values and a statistical analysis of the remaining mass deviations (maximum, average and standard deviation). These numbers give a very good estimate of the accuracy of the data. They can be used as lower limits and estimates of data accuracy when submitting the spectra to a database search.

3.1.4 Spectrum filtering

Known fragment spectra can be filtered from the LC run. Samples often contain peptides from common contaminants like keratin or unidentifiable modified peptides from the enzyme. Once these spectra are known they can be incorporated into a library. The library spectra are compared to the spectra in the current LC run and if identities are found they are extracted from the LC run and saved as a separate pkl file. The changes to the LC run are documented to keep track of the eliminated spectra. It is quite useful to exclude known contaminants from the datafile to restrict the database search as much as possible to relevant fragment spectra. By eliminating known spectra of modified peptides of the enzyme, possible misinterpretations of these often intensive fragment spectra are avoided.

3.1.5 Deisotoping and charge state deconvolution of the spectra

The ultimate purpose of preprocessing of the fragment spectra is to increase the specificity, sensitivity and accuracy of automatic database identifications. The deisotoping and charge state deconvolution is done to reduce the complexity of spectra to the point that every fragment is represented by only one datum. The two processes are linked because the charge state of every fragment has to be determined for correct deisotoping. Once the isotopes have been removed, the charge state of any fragment can no longer be recognized by the isotopic spacing in the spectrum. This is why the charge state deconvolution is performed simultaneously with the deisotoping.

The determination of the charge state for highly charged ions can be quite difficult, in particular in cases of overlapping ions with different charge states and noisy data [21]. We consider only charge states up to four. The charge state for a given peak is determined by the spacing of the following isotopes. The highest charge state considered for a particular ion is the smallest charge state of a group of three values. The three values are the charge state of the precursor ion, the highest charge state the ion under investigation could have without increasing its neutral mass beyond the mass of the entire peptide and four. Different charge states are tested by counting down to one. From a calculated fragment mass of 400 Da upwards two isotopes are expected to be present, from a

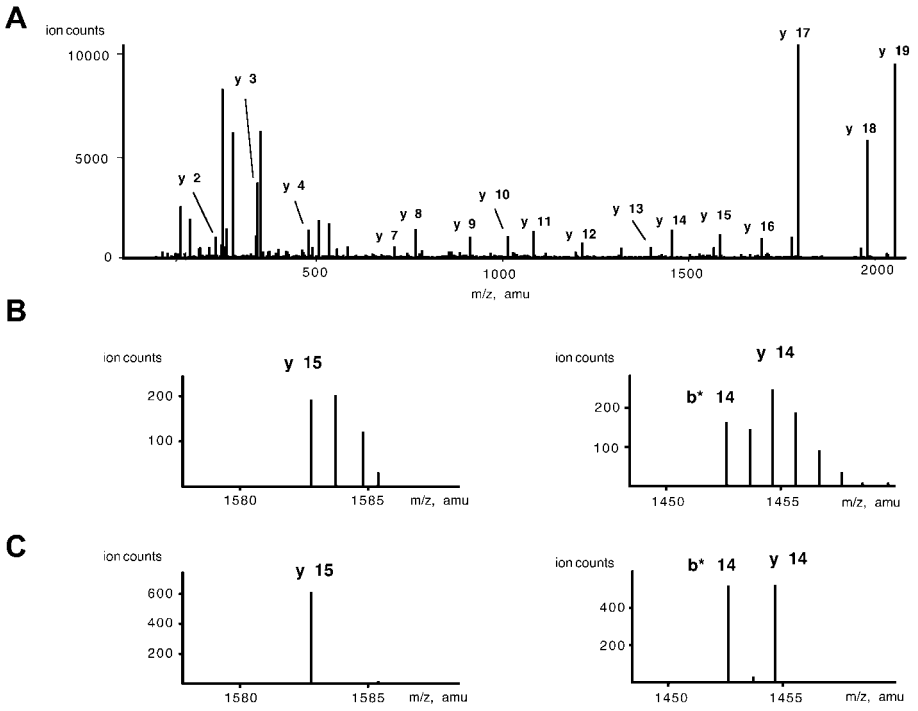

Figure 3. Fragment spectrum of the triple-charged ion of the peptide YIAWPLQGWQATF GGGDHPPK. Panel A shows the entire spectrum, panel B selected, already centroided ions and panel C the deisotoped fragments. Deisotoping is efficient. Partially overlapping isotopic envelopes, like the one of the internal b14 ion and the y14 ion are correctly treated. Deisotoping cannot be perfect because the real isotopic distribution of the ions may not correspond precisely to the one of the model peptide with an averaged amino acid composition and because the peak intensities of low abundance ions do not reflect the expected isotope intensities correctly.

mass of 1800 Da three and from a mass of 2800 Da four isotopes. A charge state is determined if the expected isotopes are found at their corresponding *m/z* values within a tolerance of $\pm$0.05 Da.

In a second step the overall intensity of the isotopic envelope is measured. The algorithm considers the possibility that any of the isotope peaks may represent an overlap of several different ions with different charge states. First, the expected isotopic distribution of an average peptide of the calculated mass is determined [22]. The average peptide is a peptide with an amino acid composition corresponding to the statistical distribution of amino acids in the nonredundant database. The intensities of the first three isotopes of the fragment ion in the spectrum are measured and normalized with the expected intensities of the average peptide. The overall intensity of the isotopic envelope of the fragment in the mass spectrum is the minimum value of the first two normalized isotope intensities or the first three isotope intensities, if the third isotope has an expected intensity of at least 20% of the first isotope and more than four ion counts. Since the determination of the overall intensity of the isotopic envelope is a lower limit estimation, the determined value is multiplied by 1.2 to allow for a slight overcompensation of existing isotopes in some cases. It should be noted that peaks in fragment spectra are often very weak and that their intensity values have a considerable statistical error [23].

After the determination of the charge state and the overall intensity of given peak a new data point in the deconvoluted, deisotoped spectrum is generated. It is placed at the *m/z* value of the single charged fragment ion with an intensity of all its isotopes added up using the measured overall intensity of the isotopic envelope and the expected isotopic distribution of a peptide with an averaged composition. Finally, the calculated isotopic envelope of this averaged peptide is subtracted from the spectrum. The procedure is repeated once on the same peak for lower charge states to take overlaps of differently charged fragments into account before searching the next remaining peak in the spectrum. Finally, the spectrum is recentroided using the internal centroiding algorithm. Figure 3 shows an example.

3.2 Protein identification

3.2.1 Evaluation of data preprocessing on a controlled sample

All the data preprocessing steps are undertaken to improve the automatic protein identification mechanism. When acquiring fragment spectra with the nano-ESI ion source we use the peptide sequence tag algorithm for protein identification. A short stretch of amino acids is read from the fragment spectrum. The sequence of two

or three amino acids together with their mass location within a peptide of a given mass is used to search the database [19, 20]. This algorithm has the advantage of being relatively specific when identifying peptides but it requires manual interpretation of the spectra and is therefore not applicable to large sets of data acquired with an HPLC interfaced to a mass spectrometer. For automatic protein identification we use the MASCOT search engine (http://www.matrixscience.com) since to our knowledge it is the only automatic search engine for fragment spectra available on the WWW [16, 17]. To evaluate the effect of data preprocessing we digested seven gel-separated proteins, mixed the peptides and subjected the sample to an HPLC-MS/MS experiment. We exported the acquired data from MassLynx as a pkl file involving as little processing as possible. From the exported data we generated two different datasets to evaluate the effect of preprocessing on protein identification, one "original" dataset and one centroided and deconvoluted/deisotoped dataset. To avoid obvious false scoring values for the original dataset we recalibrated the exported spectra, joined identical fragment spectra and subtracted one ion count from all ion peaks. One ion count is the minimal noise level in MS/MS spectra. By eliminating all peaks which consist of only one ion count false assignments of fragment ions to these peaks are avoided.

3.2.2 Overview of all the submitted spectra

From 136 fragment spectra 82 could be assigned to the seven proteins we had digested, nine spectra were from keratin and trypsin peptides and 45 spectra remained unassigned. Of these 45 spectra, 21 contained so few peaks that any unambiguous interpretation, sequence tag based or automatic, is probably impossible. Sequence tags consisting of two or three amino acids could be assigned to 16 of the remaining 24 spectra in a semiautomatic way. The semiautomatic sequence tag algorithm requires the manual assignment of one fragment ion to start the tag. The sequence tag itself is generated automatically. The ability to assign an amino acid sequence tag to a fragment spectrum using a tolerance of 0.08 Da is a good indication that the underlying precursor is indeed a peptide. When using PeptideSearch to find the sequence with these tags, in only 4/16 cases was a putative sequence returned using 150 ppm mass tolerance as for the automatic database searches. None of these four peptides could be confirmed when comparing the proposed sequence to the fragment spectrum. The confirmation of a proposed sequence is a standard procedure when using tag based protein identifications. The database-retrieved sequences are compared to the fragment spectra to find additional fragment ions expected from the amino acid sequence like b- or internal b-ions when the sequence contains a proline. One of the four peptide sequences was confirmed indirectly as belonging to one of the digested proteins, an aldolase peptide. The other 15 fragment spectra of the 16 that allowed the construction of a sequence tag were most likely generated from peptides which carry modifications or do not correspond to tryptic peptides.

To summarize the findings: of 136 fragment spectra, 91 were correctly assigned to their respective peptides by an automatic database search using MASCOT. One peptide could not be identified by MASCOT but was identifiable by the sequence tag approach. Fifteen fragment spectra were obviously generated from peptides but could not be associated to a sequence either by a sequence tag approach or by MASCOT, presumably because they belong to nontryptic or modified peptides. Eight spectra could not be interpreted in any way and 21 spectra contained so few ions that any interpretation appears impossible. The one peptide that could be identified using a sequence tag but not by using MASCOT was KELSDIAHR. This peptide is small and its fragment spectrum is weak. The largest y-ion after centroiding and deisotoping has an intensity of 18 counts. Even after preprocessing, the spectrum contains a considerable number of noise ions. For the evaluation of data preprocessing, we used only the 91 spectra that could be assigned unambiguously.

3.2.3 Influence of preprocessing on the scoring of the correctly identified fragment spectra

The change that preprocessing makes to the data is considerable. The original dataset contains 129 553 data points, the processed one only 19 786 or 15.3% of its original size, improving the database search time correspondingly. However, the purpose of preprocessing the fragment spectra is not primarily to decrease the amount of data, but to increase specificity and accuracy in the identification process. Every peptide identification algorithm compares masses calculated from peptide sequences with fragment masses found in the spectrum. Therefore, accuracy and specificity of the identification should improve if the fragment spectrum contains only one data point *per* fragment at the *m/z* value of the first isotope. The processed spectra are optimized towards this goal. The effect on the identification procedure is visible if the absolute scores before and after preprocessing are compared. The influence the preprocessing has on the specificity is reflected by the score value difference between the correct and the best incorrect hit.

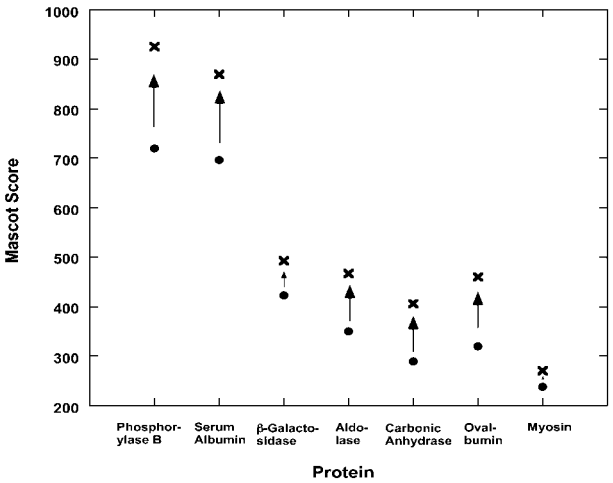

Figure 4. MASCOT score of seven model proteins before and after data preprocessing. For all proteins the score improves when processed spectra are submitted to the database search; for myosin by 34, the smallest increase, and for phosphorylase B by 206. This means that the probability that the proteins were identified by chance decreased by a factor of 2500 for myosin and a factor of 4×10^{20} for phosphorylase B.

3.2.4 Influence on the accuracy of the identification

Figure 4 shows the scores of the seven digested proteins before and after preprocessing. The protein scores increase by 34 (minimum) to 206 (maximum). The average increase is 122.7. The MASCOT score is defined as s = $-10 \times \log_{10}(P)$,P being the probability that the identification occurred by chance. This means that an increase in score of 30 is equvalent to a 1000-fold decrease of the probability that the identification occurred randomly. Protein scores are the sum of peptide scores. A more detailed analysis can be gained by looking at the peptide scores. Figure 5 shows a summary. Figure 5A shows the scores of all correctly identified peptides before preprocessing. Four peptides have a score of 0; they are not identified using the original spectra. Figure 5B shows the MASCOT scores after centroiding and Fig. 5C after centroiding and deisotoping. The peptide scores generally increase. Four additional peptides are correctly identified only after their spectra have been preprocessed. No peptide which was correctly identified before processing was lost. This is remarkable considering the overall reduction in data. It demonstrates that significant information is not lost by the procedure.

Some peptides in the lower mass range lose score upon processing. This is related to the way MASCOT uses to determine the scoring value, which is not always completely transparent. The peptide that loses 20 points in its scoring value is the peptide SSGTSYPDVLK from trypsin. Before processing its score is 77.8, after processing 57.1. Both values are so high that the identification is considered to be safe by MASCOT. When comparing the identified fragments, MASCOT reports that more fragment ions had been identified in the processed spectrum (20/47 fragments, y1-y10, b2, b6, [y1-NH_3], [y6-NH_3], [y7-H_2O], [y8-H_2O], [b2-H_2O]]–[b5-H_2O], than in the unprocessed (11/94) fragments, y1, y3, y5–10, b2, [y6-NH_3]$^{2+}$, [y9-H_2O]$^{2+}$). The average mass deviations for the matched fragments are slightly lower for the processed spectrum (about 0.05 Da, 56 ppm) than for the unprocessed (about 0.05 Da, 75 ppm). Obviously, the high score for the unprocessed spectrum is not primarily caused by the number of fragment ions that could be found but probably by the small number of peaks that had to be considered to find them. For the processed spectrum the 51 most intense peaks were considered to find the 20 matches. For the unprocessed spectrum only the most intense 15 peaks were taken into account to match 11 of them to the expected fragment masses. The total number of fragment masses considered is different because for the processed spectrum the charge state of the precursor is set to one, eliminating all the double-charged fragment ions from consideration. It can be assumed that several factors influence the overall score. For this spectrum the extreme ratio of 11 fragments found when 15 peaks were considered is probably dominating the outcome.

In the case with the second highest loss we see different factors influencing the score. It concerns a peptide from albumin, EAC*FAVEGPK. Its score before processing is 58, after processing 49. Again, both scores are so high that the identification is considered to be safe. For the unprocessed spectrum 9/84 fragment ions were found using the 67 most intense ions, for the processed 11/43 fragment ions were found using the 23 most intense ones. The 11 matched fragment ions cover the same 9 fragment ions that were found in the unprocessed spectrum. If the scores were dominated only by the number and the kind of matched fragment ions and the number of peaks considered to find the matches, the processed spectrum should score much better. The difference in scoring may be due to the different mass deviation of the fragment ions. On average, the mass deviation of the matched fragments for the unprocessed spectrum was 33 ppm and for the processed 42 ppm. The absolute mass deviations are well within the expected limits (around 0.05 Da) but the centroiding procedure removed data points which were nearly identical to the theoretical mass of the y9 ion and placed the centroided peak further away.

The precise scoring algorithm of MASCOT has never been published, so conclusions drawn from individual examples remain speculative. MASCOT is a very valuable

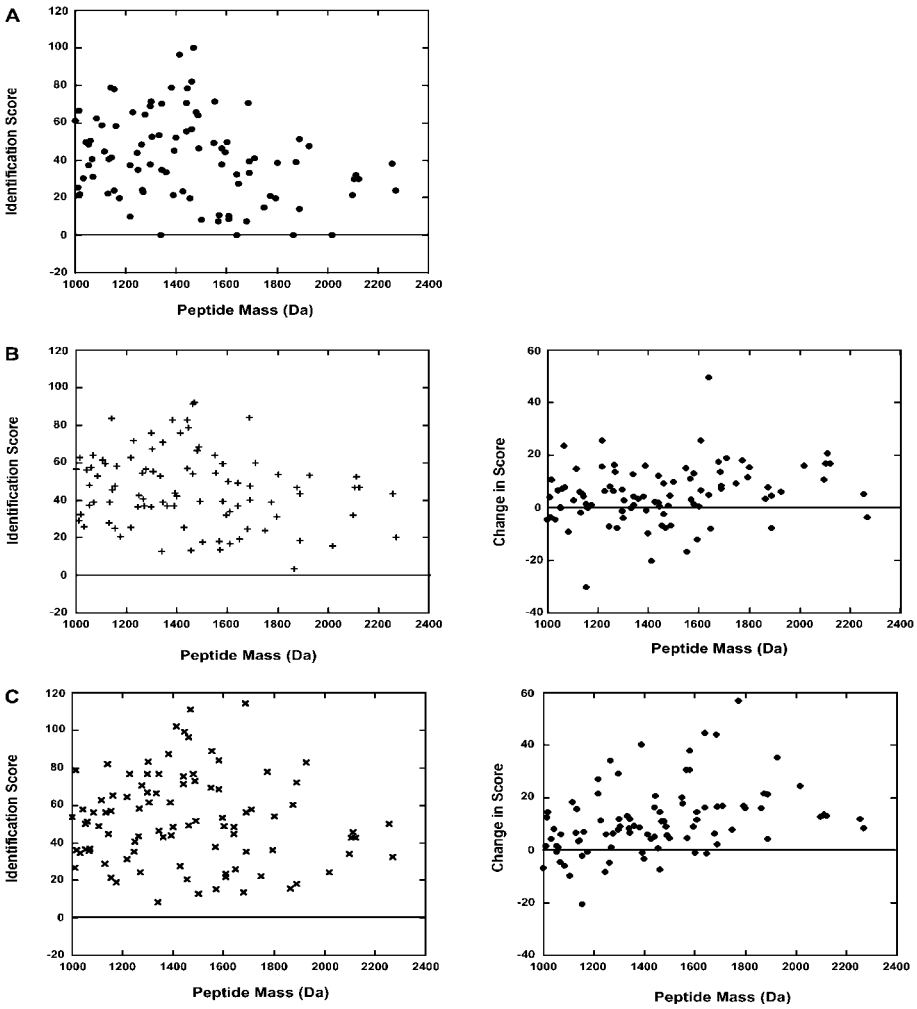

Figure 5. Scores of 91 different peptides from the digested model proteins, trypsin and keratin. Panel A shows the scores achieved when the unprocessed data are submitted to a database search, panel B the scores of the centroided spectra, panel C of the centroided and deisotoped data. Panel B and C show both the peptide scores and the differences from the original scores before processing. A score value of 0 means that the correct sequence was not identified. When comparing panels A, B and C the general trend is visible that sequences can be related better to the processed spectra than to the unprocessed. Panel C shows that some correct sequences lose score upon processing. These peptides are all relatively small which makes them more difficult to identify. No peptide was lost from the set of identified ones but four peptides could be assigned correctly only after their fragment spectrum had been processed.

tool for protein identification and is used by many scientists. This is why we discuss the influence of our preprocessing on the MASCOT identification scores.

Of 91 peptides 15 lose in score, 6 of the 15 by more than 5 scoring points. The average gain *per* peptide is 11.05, the maximum gain is 56.6, the maximum loss 20.7. No peptide was lost by searching the processed data but four peptides are correctly identified only after preprocessing. Considering these numbers and the overall impression gained from Fig. 5, we think that preprocessing increases the correctness in the automatic peptide identification using MASCOT, even though only 15% of the original data are still present.

3.2.5 Influence on the specificity of the identification

The specificity of the identification is expressed by the difference in score between the correct and the best wrong sequence. Figure 6 gives a graphical summary. Figure 6A shows the scoring difference before processing. Fig. 6B after processing, and Fig. 6C illustrates the change in the difference between the correct identification and the best wrong sequence realized by data preprocessing. Negative values in Fig. 6A and 6B mean that the correct peptide was not ranked first. MASCOT still reports these peptides because it realizes that they belong to proteins which have already been identified by other peptides. The effect is not only that four peptides are not listed at all before processing, but an additional four peptides were not ranked first even though their sequences are correct. All four peptides are moved up to first place after processing. The only correct sequence which is not ranked first is one of the four peptides which was not listed when the unprocessed data were searched. This peptide can be identified as one of three candidates if a sequence tag algorithm with a semiautomatically generated tag is used. Figure 6C shows that – as with the scores – not all identifications gain in specificity but in general the specificity does increase. Of 91 peptides 16

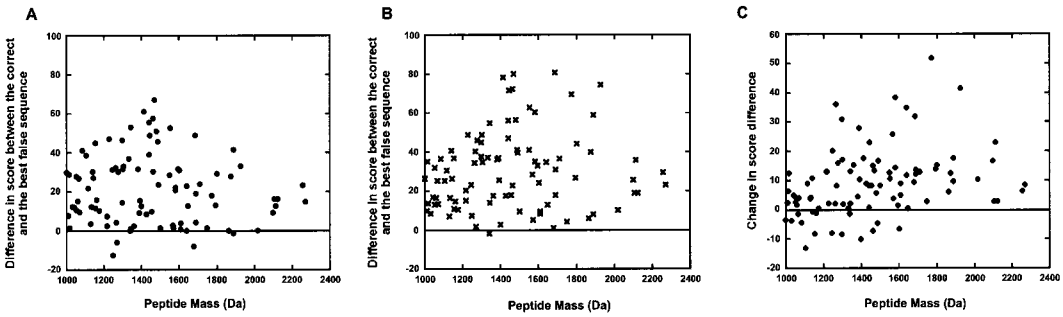

Figure 6. The difference in score between the correct sequences and the best false sequences. Panel A shows the difference for the unprocessed spectra, panel B for the processed ones. Panel C shows graphically the change in scoring difference between the correct sequences and the best false sequence realized by processing the spectra. Negative values in panel A and B mean that the correct sequence was not ranked first. As can be seen by comparing panels A and B, processing generally improves the specificity of the identification. Panel C shows that some peptides lose specificity of identification. However, no peptide which was ranked first lost this position to a false sequence but four correct sequences moved up to rank one only after their spectra had been processed.

lose in specificity, 7 of the 16 by more than 5 scoring points. The average gain *per* peptide is 9.0, the maximum gain is 51.6, the maximum loss 13.2.

The numerical similarity between the average gain in absolute score and in specificity means that nearly the entire positive effect is realized by improving the specificity of the identification. This is shown in Fig. 7, which demonstrates the change in scoring of the best false sequence. The values are distributed around zero. There is an average increase *per* peptide of 2.0. This is to be expected because the sequences of the best wrong peptides are often quite similar to the correct sequences. That the scores of correct identifications are raised and that this effect is almost entirely due to an improvement in specificity is the intended result of preprocessing. Obviously, the data are successfully reduced to the information which specifies the fragmented peptides.

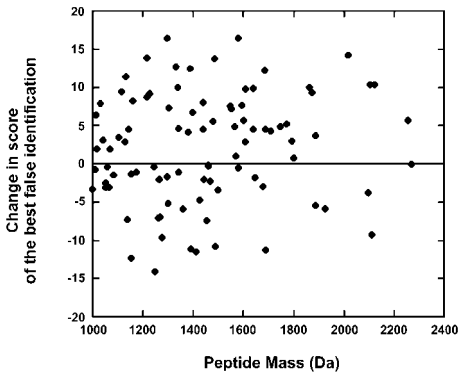

Figure 7. Change in scoring of the best false sequence realized by processing the spectra. In contrast to the scoring of the correct sequences in Fig. 5C and the difference in score from the best false sequence in Fig. 6C, the scores of the best false sequences show no clear positive trend, underlining that data processing improves the assignment of correct sequences to their fragment spectra.

3.3 Application to a larger experimental dataset

We have discussed the effect of our preprocessing steps at the level of individual peptides since database searches are done with individual fragment spectra. When submitting a complete LC-MS/MS run MASCOT sorts the retrieved peptides under common proteins. They represent the final result of the experiment. In this paragraph we discuss the effect of preprocessing on the identification of proteins. We will not mention all details but concentrate instead on the final outcome of proteins retrieved from the database.

For this purpose we took a protein fraction from a total cell lysate of *Drosophila* eluted from an affinity column to YPS, a major cytosolic RNA binding protein, digested it with trypsin and submitted the peptide mixture to an LC-MS/MS experiment. We produced two datasets from the exported spectra, one “original” and one processed as described in Section 2.4. The exported dataset contains 384 fragment spectra with 519 500 data points, the processed dataset 338 spectra with 80 100 data points (15.5% of the total). The processing took 11 min on a G3 Macintosh (350 MHz).

For the unprocessed data 180 peptides identified 60 proteins with a score larger than 50, a score MASCOT sets as a threshold for significance. One of the 60 proteins is most likely an obvious false positive identification. The protein is identified with a single peptide sequence that is assigned to a fragment spectrum as second best hit. The sequence representing the best hit is a tryptic peptide of another protein amongst the identified 59 proteins. The two peptide sequences are similar (DIPGLTDTTIPR, DIPGLTDNTVPR with one deamidation). Since it is very unlikely that these two peptides elute at the same time from the LC column we decided to consider the protein identified with this single peptide as a wrong identification and eliminated it from further consideration. After processing, 180 peptides identified 61 proteins with significant scores. Two proteins from the formerly significant proteins had been dropped from this list and four additional proteins were elevated to a significant score. Therefore, six proteins were qualitatively affected when 50 respective 61 were identified altogether.

Figure 8 gives a summary of how the processing affects the score of individual proteins. Of the 63 proteins which were identified either before or after processing with a significant score, 18 have a reduced score after processing, one remains unchanged and 44 have an increased score. On average the scores increase by 24.5, the biggest loss in score is 25, the most important gain 162. Relative to the scores before processing, the scores increase on average by 22%, the biggest loss is 26%, the largest gain 133%. These numbers express the scale of the effect of preprocessing on the scores. They do not have the same meaning as in the case of peptide identifications discussed earlier. Here, the reduction of score can be a positive result if it occurs to a false positive identification. This is the case for the protein with the biggest absolute loss in score. Its score before processing was 120 based on four peptides, after processing 95 based on two peptides. The two peptides eliminated were probably false positive identifications. One spectrum was mapped after processing to a different peptide that was already identified from another fragment spectrum corresponding to a different charge state of the precursor and the other one was a very short peptide (IANQIVFK) which could not be confirmed by a sequence tag-based approach or by comparing the sequence to the spectrum.

Six proteins were affected in a qualitative way by spectrum processing. Two were removed from the list of significantly identified proteins and four were added. One of the two proteins whose score fell below 50 was probably a false positive identification. It was identified by a single peptide. This peptide was replaced by another peptide as first hit after the spectra had been processed, which related the spectrum to another protein already identified by other peptides (NEIIIPKIDK was replaced by RIEAIPQIDK from glutathione S-transferase). Due to the similarity of the two peptides it is not possible to make a decision when comparing the sequences with the fragment spectrum manually. The other protein, identified by one peptide (VVGQLGQVLGPR), seems to have been a correct identification. The sequence corresponds in an excellent way to the fragment spectrum. Before processing its score was 61, after processing 45. This is the protein with the biggest relative loss in score. After the spectrum had been processed the y6 ion was not assigned. The centroiding generated a larger number of peaks with intensities comparable to that of the y6 ion.

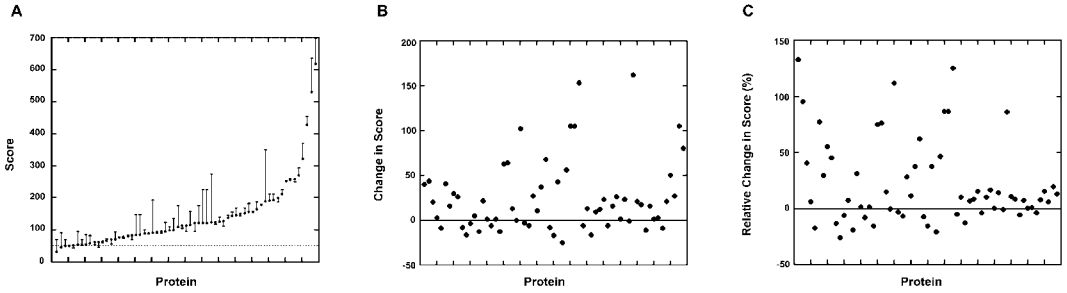

Figure 8. Impact of data preprocessing on the score of proteins in a larger experimental dataset. The proteins are sorted along the x-axis according to their score before data processing. Panel A shows the scores before (dots) and after (bars) processing, panel B the change in score realized by preprocessing, and panel C this change in percent of the score before processing. The line in panel A is drawn for the threshold value of 50, which MASCOT uses to define significant protein identifications. After spectrum preprocessing 61 proteins had scores above 50. Four proteins rose above this level after fragment spectra of their peptides were processed, two dropped below it. One of these two was a false positive identification, the other one had already been correctly identified. Panel C shows that the biggest relative gains occur to proteins of relatively low score.

It is possible that its assignment would have brought so many more peaks into the list of considered peaks that the overall score would have been even smaller.

All four proteins that were newly identified with significant scores are probably correct identifications. Sequence tag-based database searches and direct comparisons between sequence and spectrum confirm the results. The protein with the largest relative gain in score is amongst them. In summary, for five of the six proteins qualitatively affected by the data preprocessing the resulting effect was correct, for one it was incorrect.

It may be added that these improvements to the automatic proteins identification are achieved for a research tool that has already a very high quality, the MASCOT search engine. For instance, we could hardly detect a dependence of the scores on the calibration to a tolerance of 0.8 Da for the fragments and 800 ppm for the precursor and mass deviations of about 0.3 Da for the current MASCOT implementation.

4 Concluding remarks

LC-MS/MS experiments with subsequent automatic database searches are a major analytical tool for protein identification in biological research. We have developed a series of data preprocessing steps for the fragment data to improve the accuracy and specificity of the database search. The aim of the processing is to transform the mass spectra so that every fragment is represented by exactly one data point, its first single-charged isotope. By doing so the amount of data is reduced to about 15% of its original size. Even though individual peptides can have a reduced MASCOT score, the average gain in score for correct peptides in the investigated example is 11. The specificity of the database search is increased after processing. The average distance in scoring to the best wrong sequence increases by 9 *per* peptide in the MASCOT score. This means that accuracy and specificity improved roughly by a factor of 10 *per* peptide for our control dataset. These improvements are visible when considering only protein identifications. In an experimental dataset with 61 proteins identified with significant scores, six proteins were qualitatively affected by the data processing. Two were removed from the list, four were added. For five of the six proteins the achieved effect was most likely correct, for one it was incorrect. The reduction of dataset size and the improvement in accuracy and specificity are useful for any protein identification but they may be specially relevant for the automatic characterization of secondary protein modifications since a protein modification site can be reflected by only one peptide in the entire LC run.

This research has been supported by a grant from the Bundesminsterium für Bildung und Forschung (BMBF), BioFuture, Project Nr.: 0311862. We gratefully acknowledge the receipt of the affinity purified proteins from Elisa Izaurralde and her group. We thank Dr. David Thomas for bringing the text to the level of a native speaker. The data of our control experiment and the experimental dataset are accessible on the web under the address: WWW browser: ftp://www.narrador.embl-heidelberg.de/%2FServer/WWW/Pub/Outgoing/LCMSMS_Data.zip FTP-client: host IP: 194.94.44.230, user id: anonymous, directory: Outgoing, file: LCMSMS_Data.zip

Received February 21, 2003

5 References

[1] Gavin, A. C., Bosche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[2] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[3] Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M. *et al.*, *Nature* 2002, *419*, 520–526.

[4] Lasonder, E., Ishihama, Y., Andersen, J. S., Vermut, A. M. *et al.*, *Nature* 2002, *419*, 537–542.

[5] Grad, R., Muller, M., *Curr. Opin. Mol. Ther.* 2001, *3*, 526–532.

[6] Yates, J. R., McCormack, A. L., Eng, J., *Anal. Chem.* 1996, *68*, 534A–540A.

[7] McCormack, A. L., Schieltz, D. M., Goode, B., Yang, S. *et al.*, *Anal. Chem.* 1997, *69*, 767–776.

[8] Choudhary, J. S., Blackstock, W. P., Creasy, D. M., Cottrell, J. S., *Proteomics* 2001, *1*, 651–667.

[9] Gras, R., Muller, M., Gasteiger, E., Gay, S. *et al.*, *Electrophoresis* 1999, *20*, 3535–3550.

[10] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392.

[11] Zhang, N., Aebersold, R., Schwikowski, B., *Proteomics* 2002, *2*, 1406–1412.

[12] Washburn, M. P., Wolters, D., Yates, J. R., III, *Nat. Biotechnol.* 2001, *19*, 242–247.

[13] Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S. *et al.*, *Nature* 1996, *379*, 466–469.

[14] Wilm, M., Mann, M., *Anal. Chem.* 1996, *68*, 1–8.

[15] Eng, J. K., McCormack, A. L., Yates, J. R., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[16] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[17] Creasy, D. M., Cottrell, J. S., *Proteomics* 2002, *2*, 1426–1434.

[18] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[19] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[20] Mann, M., *Trends Biol. Sci.* 1996, *21*, 494–495.

[21] Senko, M. W., Beu, S. C., McLafferty, F. W., *J. Am. Soc. Mass. Spectrom.* 1995, *6*, 52–56.

[22] Senko, M. W., Beu, S. C., McLafferty, F. W., *J. Am. Soc. Mass Spectrom.* 1995, *6*, 229–233.

[23] O'Conner, P. B., Little, D. P., McLafferty, F. W., *Anal. Chem.* 1996, *68*, 542–545.