New Data Base-independent, Sequence Tag-based Scoring of Peptide MS/MS Data Validates Mowse Scores, Recovers Below Threshold Data, Singles Out Modified Peptides, and Assesses the Quality of MS/MS Techniques*

Mikhail M. Savitski‡, Michael L. Nielsen, and Roman A. Zubarev

The Mascot score (M-score) is one of the conventional validity measures in data base identification of peptides and proteins by MS/MS data. Although tremendously useful, M-score has a number of limitations. For the same MS/MS data, M-score may change if the protein data base is expanded. A low M-value may not necessarily mean poor match but rather poor MS/MS quality. In addition M-score does not fully utilize the advantage of combined use of complementary fragmentation techniques collisionally activated dissociation (CAD) and electron capture dissociation (ECD). To address these issues, a new data base-independent scoring method (S-score) was designed that is based on the maximum length of the peptide sequence tag provided by the combined CAD and ECD data. The quality of MS/MS spectra assessed by S-score allows poor data (39% of all MS/MS spectra) to be filtered out before the data base search, speeding up the data analysis and eliminating a major source of false positive identifications. Spectra with below threshold M-scores (poor matches) but high S-scores are validated. Spectra with zero M-score (no data base match) but high S-score are classified as belonging to modified sequences. As an extension of S-score, an extremely reliable sequence tag was developed based on complementary fragments simultaneously appearing in CAD and ECD spectra. Comparison of this tag with the data base-derived sequence gives the most reliable peptide identification validation to date. The combined use of M- and S-scoring provides positive sequence identification from >25% of all MS/MS data, a 40% improvement over traditional M-scoring performed on the same Fourier transform MS instrumentation. The number of proteins reliably identified from *Escherichia coli* cell lysate hereby increased by 29% compared with the traditional M-score approach. Finally S-scoring provides a quantitative measure of the quality of fragmentation techniques such as the minimum abundance of the precursor ion, the MS/MS of which gives the threshold S-score value of 2. ***Molecular & Cellular Proteomics 4:1180–1188, 2005.***

Protein identification using MS/MS (1–3) commonly encounters two problems. The first one is the well known problem of assessing the reliability of identified data ("reliability issue"). The second problem is the fact that a large portion of MS/MS data (often 90% or more) does not produce any useful identification ("efficiency issue"). These problems, as will be shown below, are inter-related.

The reliability problem is effectively addressed by a variety of different search engines, *e.g.* Mascot (4), Sequest (5), etc., through the use of a scoring technique that evaluates the probability of a false positive identification. Although the evaluation methods might be sophisticated, they have some limitations. One limitation is that for the same MS/MS data the score may significantly change if the content of the protein data base is altered (protein addition to and deletion from the data base are everyday phenomena); thus the score is dependent on whether the data base accurately represents all occurring peptides. Another important limitation is that some poor quality data that should not be trusted can give by pure chance a nearly perfect match, and thus corresponding peptides are wrongly identified with a very high score. Thus, even above threshold identifications call for confirmation by search engine-independent techniques (6–8). Conversely a low score may arise because of two very different reasons: poor matching and poor MS/MS data quality. Thus when an extremely high quality, informative in terms of fragmentation, MS/MS spectrum returns zero or a below threshold score, it is most often discarded, whereas the actual reason for the poor score is that the peptide in question is not present in the data base. Alternatively an MS/MS spectrum of a peptide definitely present in the data base will receive a low score and be discarded because of the poor quality resulting from the presence of noise spikes, distorted isotopic distribution, missing fragments, etc. For instance, Mascot

From the Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, S-75123 Uppsala, Sweden

Received, April 19, 2005, and in revised form, May 9, 2005
Published, MCP Papers in Press, May 22, 2005, DOI 10.1074/mcp.T500009-MCP200

search engine uses M-score (9) that takes into account the number of mismatched fragments and their relative abundances. The user does not usually know what caused the poor M-score unless a time-consuming manual inspection of the spectrum is performed. Thus the automatic routine cannot make an intelligent decision, which should be in the first case to search the data in question allowing for modifications/mutations (10, 11). In the second case, the decision should be to repeat the analysis or look for supporting information, *e.g.* other peptides belonging to the same protein, or to invoke the retention time (12). Thus a large fraction, perhaps more than 50%, of potentially useful MS/MS data is currently discarded; this aggravates the efficiency problem.

Complementing the M-score with another data base-independent score (13) that would evaluate primarily the quality of the MS/MS data provides a means for distinguishing between the reasons for the poor M-score and for making the above intelligent decision, rendering the tedious and time-consuming manual data inspection superfluous. This has recently been realized by several groups who have designed data base-independent scoring principles. For instance, Bern *et al.* (13) assessed the quality of tandem MS data obtained on a low resolution instrument and managed to filter out 75% of the unidentifiable spectra while losing only 10% of the identifiable spectra. They found that the number of peaks and their abundances (the guidelines for manual data quality inspection) had in fact little classification power compared with the number of peak pairs separated by an amino acid mass. In this study, we introduce and evaluate a new scoring principle (S-score) that differs from the scoring suggested by Bern *et al.* (13) in several aspects. First, the S-score is based on just one parameter, the maximum length of peptide sequence tag, which simplifies the interpretation of the S-score value. Second, it utilizes high mass accuracy afforded by FTMS (14). Finally S-score uses the MS/MS information that comes not only from the traditional collision activated dissociation (CAD)[1] but also from electron capture dissociation (ECD) (15, 16) performed on the same peptide.

The application of the S-score goes beyond the mere filtering out of the "bad" spectra and includes salvaging some of the below threshold data. Moreover a further verified sequence tag used for the S-score is compared with the search engine sequence assignment, revealing cases of false positive identifications. In some cases, the sequence tag immediately reveals the presence of a modified sequence, removing the need for a separate *de novo* sequencing program.

An important issue also pertaining to this discussion is the instrument performance, including also the performance of a given fragmentation technique. Traditionally fragmentation efficiency has been measured as a ratio of the total abundance of fragmentation products and the abundance of the precursor ion before fragmentation (17). However, this is a general approach that is silent about the information quality of the data. For instance, the same efficiency could be assigned to two different MS/MS peptide spectra, one containing a single but very intense peak corresponding to the NH_3 loss from the precursor ion and the other containing low intensity but extensive backbone fragmentation. Clearly the information content of these two MS/MS spectra would be different. We demonstrate the applicability of the S-score for quantitative assessment of the information content in peptide MS/MS spectra.

EXPERIMENTAL PROCEDURES

Mass Spectrometry—All experiments were performed on an Agilent 1100 nanoflow system coupled to a 7-tesla hybrid linear ion trap Fourier transform mass spectrometer (LTQ FT, Thermo Electron Corp., Bremen, Germany) equipped with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark) as described previously (18). Briefly 70 μg of an *Escherichia coli* whole cell lysate was loaded onto a one-dimensional gel, and the protein bands were visualized with colloidal Coomassie Blue. The gel lane was excised into 20 equally sized fractions, and finally the proteins were reduced, alkylated, and in-gel digested with modified sequence grade trypsin (Promega, Madison, WI) as described previously in the literature (19). Samples were vacuum-centrifuged to remove all organic solvents and prior to mass spectrometric analysis were reconstituted into 20 μl of HPLC water containing 0.1% TFA. Mass spectrometric experiments were performed using unattended data-dependent acquisition in which the mass spectrometer automatically switches between a high resolution survey scan (r = 100.000) followed by consecutive ECD and CAD fragmentation (r = 25.000) of the two most abundant multiply charged peptides eluting at this moment from the nano-LC column. A total number of 20 samples of an *E. coli* lysate were analyzed, and prior to data mining all acquired dta files were combined for more thorough analysis.

In the previous analysis (18) the numbers of uniquely identified proteins for each gel band (sample) were calculated and then added together. Here instead the number of unique proteins from the combined gel strips was derived. Thus the number of proteins identified by using the complementary pairs approach was 224 in this case compared with 256 reported earlier.

S-score Description—The S-score was calculated from the so called dta files that contain the mass and the charge state of the precursor as well as the m/z values and intensities of all the fragment ion peaks in the spectrum above a certain cutoff intensity value. For every precursor, two dta files are present, one representing CAD and the other representing ECD fragmentation.

To build a sequence tag that serves as the basis for the S-score, the program takes deisotoped, neutral CAD fragment masses (potentially true fragments (PTFs)), adds the molecular mass and the mass of a water molecule to the peak list, and builds a sequence ladder (tag) between them, fitting masses of amino acid combinations.

To create a PTF list the data in the CAD and ECD dta files were deisotoped and charge-deconvoluted (20) to the neutral state. Ion fragments of monoisotopic mass $<$800 Da appear often without their heavier isotopes due to the noise cutoff. In these cases, the neutral masses were derived assuming that the charge state of the peak cannot exceed the charge state of the precursor ion if the peak originates from a CAD dta file and will be less than the charge state of the precursor if the peak originates from an ECD dta file (due to charge reduction in ECD). Thus a peak without isotopes located at

[1] The abbreviations used are: CAD, collisionally activated dissociation; AGC, automated gain control; ECD, electron capture dissociation; ID, identification; LTQ, linear trap quadrupole; M-score, Mascot score; PTF, potentially true fragment; RST, reliable sequence tag.

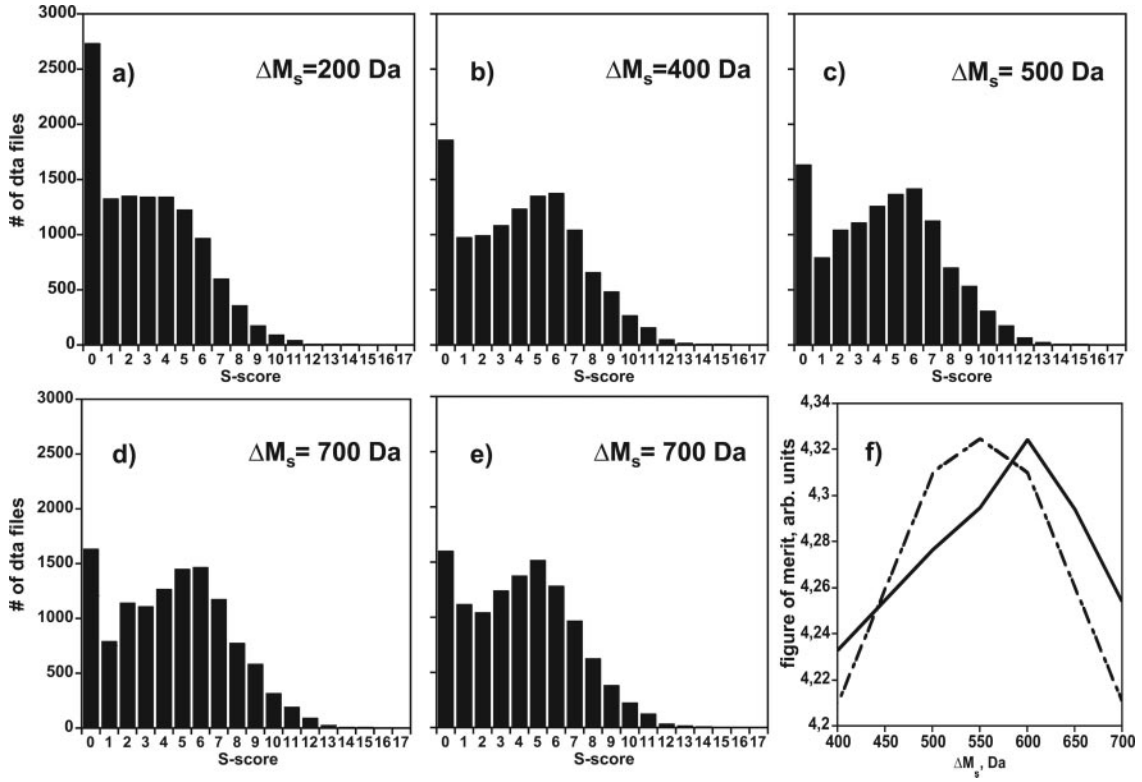

FIG. 1. *a–e*, distributions of S-scores of the acquired data for different ΔM_s values. *e*, ΔM_s = 700 Da, but the peptide mass is not included in the PTF set for each dta file. *f*, the figure-of-merit curves of the two different criteria for ΔM_s estimation. The *solid line* corresponds to criterion 1, and the *dashed line* corresponds to criterion 2. *arb.*, arbitrary.

m/z = 500.2 in a CAD spectrum of a 2+ precursor will be allowed to have a neutral mass of 500.2 and 1000.4 Da. Both these masses are considered for the construction of complementary pairs.

A fragment ion M_c, I_c with neutral mass M_c and intensity I_c in a CAD dta file of a peptide with neutral mass M will qualify for the PTF set if it passes at least one of the following tests. (*a*) It is an element of a "golden pair" (21). That means there is a corresponding fragment M_e, I_e in the ECD dta file that together with the CAD fragment ion satisfies one of the following equations.

$$M_c - M_e = -17.03 \quad \text{(Eq. 1)}$$

$$M_c - M_e = -16.02 \quad \text{(Eq. 2)}$$

$$M_c + M_e - 17.03 = M \quad \text{(Eq. 3)}$$

$$M_c + M_e + 16.02 = M \quad \text{(Eq. 4)}$$

To reduce the computational redundancy and uncertainty, the algorithm primarily targets *y*-ion sequence tags. Thus if M_c is identified as a *b*-ion (satisfies Equation 1 or Equation 4), the mass of the complementary *y*-ion is calculated using the peptide mass and added to the list instead of M_c. Note that the total amount of information in the fragment list remains unaltered. (*b*) It is an element of a "complementary pair." That means there is a corresponding fragment ion $M_{c'}$, $I_{c'}$ in the same CAD dta file satisfying the following equation.

$$M_c + M_{c'} = M \quad \text{(Eq. 5)}$$

Here of course *y*-ions are indistinguishable from *b*-ions. (*c*) It is the first peak in an isotopic cluster. Furthermore all masses of fragment ions have to pass through the peptide mass window (18) to qualify.

After adding to the PTF set of qualified masses the mass of H_2O and the neutral mass of the peptide M (the two additional masses are added to improve the length of the *y*-ion sequence tag), the algorithm proceeds to find the longest possible amino acid sequence tag that can be constructed from these masses. The number of masses used for the construction of this tag minus one is the S-score value. The maximum allowed mass difference between adjacent masses was ΔM_s. The choice of 575 Da for the ΔM_s value will be explained in the next section.

The masses in the PTF set are sorted in an increasing order $\{m_1, \ldots, m_N\}$. The algorithm assigns to each value m_i, $1 \le i \le N$, a value of a running tag length TL_i, which is initially zero for all i. Following that, the masses m_i are selected in an increasing order (i = 1, 2, . . . , N), and the corresponding values TL_i are determined for each mass m_i according to the following recipe. First the differences $\Delta_j = m_i - m_j$ are calculated for all $j < i$. Then the program goes through all Δ_j values and, if the current Δ_j is equal to any combination of one or several amino acids and does not exceed the maximum allowed mass difference ΔM_s, then TL_i accepts the value of $TL_j + 1$ unless its current value is higher. After determining all TL_i values, the largest TL_i is selected, and the value of this TL_i is the S-score for this dta file.

Reliable Tag—Besides the sequence tag that gives rise to the S-score, a "reliable" sequence tag (RST) was constructed for each spectrum. The objective was to design the most reliable sequence tag available in mass spectrometry to date and to use this tag to produce the most reliable sequence identification. The RST is constructed as follows. Only fragment ions satisfying conditions *a* and *b simultaneously* are selected for the construction of this tag. The PTF masses of the tag are thus *doubly* confirmed. The maximum step difference for the reliable tag was chosen to be 398 Da. The justification for this will be given in the following section.

RESULTS

To evaluate the significance of the S-score for characterizing the quality of MS/MS data, histograms of S-scores obtained for 2+ precursor ions were built (Fig. 1). The expectation was that the S-score could provide a means of differentiation between good peptide spectra on one hand and poor or non-peptide spectra on the other hand. The 2+ charge state was selected because it is the predominant charge state in the precursor population (75%) and also because we wanted to obtain the clearest possible picture by untangling the effect of different charge states.

Selection of ΔM_s—The optimum ΔM_s value was chosen by using two independent criteria. Criterion 1 measured the extent of bimodality (*B*) of the part of the analyzed data for 2+ peptides (Fig. 1, *a–d*). The following simple formula was used for the figure of merit,

$$B = \mathrm{Max}_1 * \mathrm{Max}_2/\mathrm{Min}^2 \qquad \text{(Eq. 6)}$$

where Max_1 and Max_2 were the two local maxima of the histogram, and Min was the height of the lowest point between them. The step difference that yielded the highest *B* was 600 Da (Fig. 1*f*, *solid line*). Criterion 2 measured the difference between the mean values of the S-score distributions of assumed "good" and bad spectra. The criterion for good spectra was the presence of complementary pairs or golden pairs; the spectra that did not meet this criterion were deemed bad. We should mention that this was a rather crude separation but one thought to be sufficiently good for optimizing the ΔM_s value. This approach gave a maximum at 550 Da (Fig. 1*f*, *dashed line*). A compromise between the two criteria was chosen, $\Delta M_s = 575$ Da. In Fig. 1, *a–d*, the evolution of bimodality is shown along with an abnormality in the abundance of the S = 2 peak that originates at $\Delta M_s = 500$ Da and increases with higher masses. This behavior is explained by the fact that, given a sufficiently small peptide mass, *e.g.* 900 Da, a single cleavage site located in the middle of the peptide can give S = 2, an increment of two from the previous value S = 0. When the peptide mass is removed from consideration, this abnormality disappears as seen in Fig. 1*e* (compare with Fig. 1*d*).

Classification of MS/MS Spectra—The prime purpose of introducing the S-score was to partition the acquired MS/MS data into three classes: A, B, and C. Class A was reserved for data that has been identified by Mascot and whose credibility was proven either by the significance of the Mowse score, by the RST, or by both. Elements of class B have not fulfilled the criteria for qualifying as class A data, and in some cases Mascot has not even suggested a sequence for them, but according to the S-score they are most likely peptides with decent MS/MS spectra and should be worth pursuing further identification. Finally class C consists of MS/MS data that according to the S-score either belongs to non-peptides or peptides with such poor MS/MS spectra that reliable identification is impossible, and thus an attempt of identification would be counterproductive.

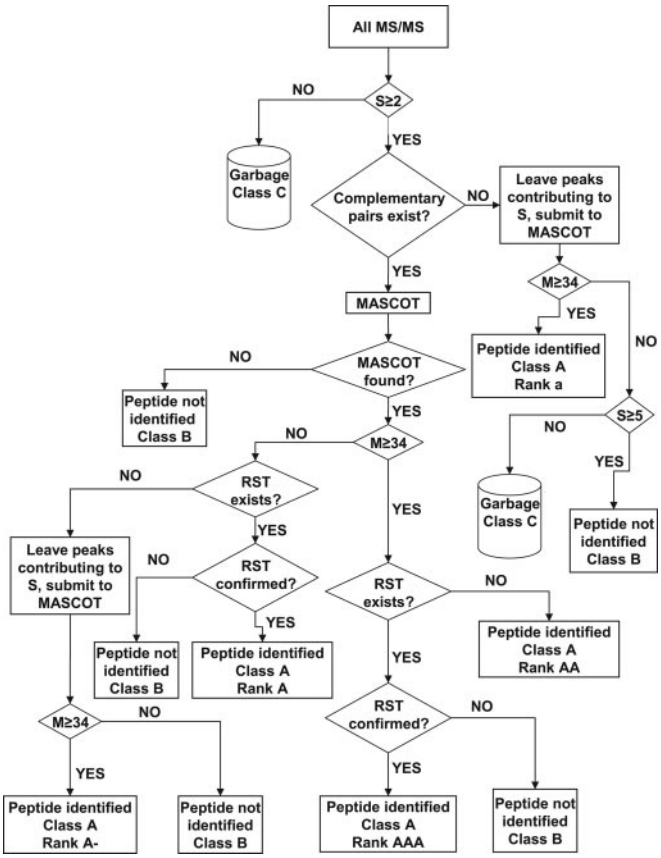

FIG. 2. **Flowchart for the data processing procedure.** The MS/MS data are sorted into three classes: A (identified by Mascot), B (good spectra, not identified by Mascot), and C (bad spectra, no identification). The class A data is ranked according to its validity.

Ranking of A-data—In the conventional approach to proteomics, the peptide identification quality is synonymous to the M-score value. With the introduction of parallel S-scoring, this is no longer so. With this in mind, the following ranking procedure was devised. The acquired MS/MS data were processed according to the scheme depicted in Fig. 2.

The elements of class A were ranked in the following order of increasing validity: *a*, *A*−, *A*, *AA*, *AAA*. For peptides with the rank *AAA* (the most reliable), Mascot has identified sequences with an above threshold score. Furthermore by using the complementary and golden pairs extracted from the MS/MS data, RSTs were derived and compared with the Mascot data. An agreement with the Mascot-suggested sequence ensured that no misidentification due to the mix-up of *b*- and *y*-ions has occurred. Peptides with rank *AA* were also identified by using the complementary and golden pairs, and Mascot had assigned a score above the suggested significance level, but an RST did not exist, and the sequence direction could not be confirmed. Peptides with rank *A* were identified with a below threshold score, but they had an RST that complied with the

TABLE I
Results of MS/MS data classification and ranking according to the procedure shown in Fig. 1

Class	Rank	Percentage of total MS/MS	Sum of identified proteins/ increase/percentage	N_{pep}/increase/ percentage
A	*AAA*	10.4	157/157/100	5.06/5.06/100
A	*AA*	7.5	224/64/28.2	6.12/1.06/17.3
A	*A*	4.5	262/38/18.3	6.39/0.20/3.1
A	*A*−	2.6	286/24/8.4	6.78/0.39/5.8
A	*a*	0.3	288/2/0.7	6.86/0.08/0.12
B		39.2		
C		35.5		

sequence. Peptides with rank *A*− were fished out from the set of peptides with a Mascot score below the threshold and without an RST. This was done by researching the data for these peptides, but instead of using only the complementary/golden pair data, the pairs of masses from the PTF list with differences in mass that corresponded to amino acid combinations of mass <575 Da were searched. The peptides that by using this approach got an above-the-threshold Mascot score were ranked *A*−. Finally the peptides ranked *a* were those that did not have complementary/golden pairs but had S > 2. These spectra were searched with Mascot using the fragments from the PTF set as for *A*− data and qualified if they got an above-the-threshold Mascot score. Table I shows the resulting statistics.

Validation of A-class Ranking—Ideally all found peptides should belong to the most valid rank *AAA*. The worst case is when the majority is of low validity, *a* or *A*−. In reality, there is a distribution of ranks. As can be seen in Table I, the amount of peptides in each rank decreases with decreasing rank. This is consistent with most identifications being true and valid. Indeed correctly identified peptides would fail to pass one of the filters only because of a mishap, such as unfavorable statistics in the isotopic distribution. Such a mishap should be a low probability event, so two mishaps for the same peptide would be even less probable and so on. Consequently if the bulk of the peptides are correctly identified, the number of peptides should decrease for ranks of lower confidence as in Table I.

The average number of peptides per protein N_{pep} increases each time when the peptide set is extended by adding the peptides of the lower rank. This is an indication of the validity of peptide IDs in the lower ranks because if the added peptides were false positives, they would likely be distributed among unrelated proteins, and N_{pep} would decrease.

Validation of S-score Threshold—As can be seen from the flowchart (Fig. 2), a total of four different filters were applied at various stages. The first one is simply a requirement for the S-score to have an above threshold value of 2. The justification for this value is deduced from Fig. 3. In Fig. 3, the S-score distribution is presented for MS/MS files of charge state 2+. The distribution (*black columns*) is clearly bimodal and has a distinct valley at S = 1. The second distribution (*light columns*) in Fig. 3 is that of MS/MS files for which complementary pairs exist. Note the relatively low abundances of S = 0 and S = 1

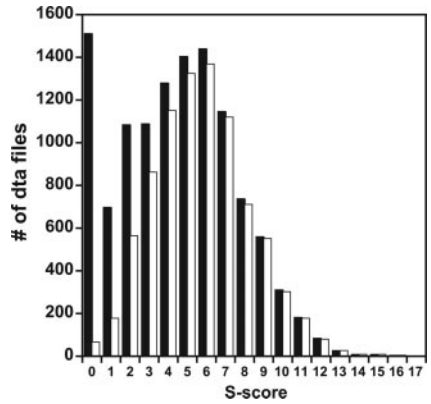

FIG. 3. **Distribution of S-scores of all dta files (*black columns*) *versus* the distribution of S-scores for dta files from which complementary pairs were derived according to Equations 1–5 (*light columns*).**

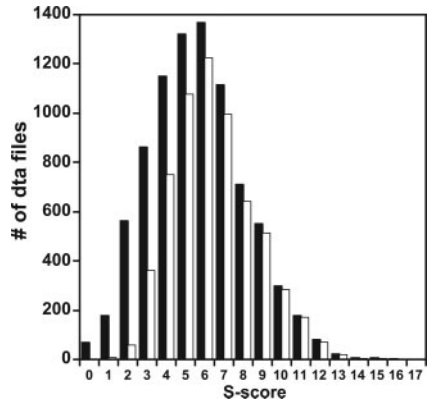

FIG. 4. **Distribution of S-scores for dta files from which complementary pairs were derived (*black columns*) *versus* the distribution of S-scores for dta files for which Mascot found a peptide sequence (*light columns*).**

that support the choice of S = 2 as a threshold value. This distribution has a mean of 6.7. In Fig. 4, the distribution of Mascot-identified (meaning Mascot suggested a sequence with *any* score) peptides (*light columns*) is plotted against the complementary pair distribution (*black columns*). The mean of the Mascot-identified distribution is clearly shifted toward higher values (mean of 7.4) and also supports the choice of S = 2 as a threshold. No peptides with scores above M = 34

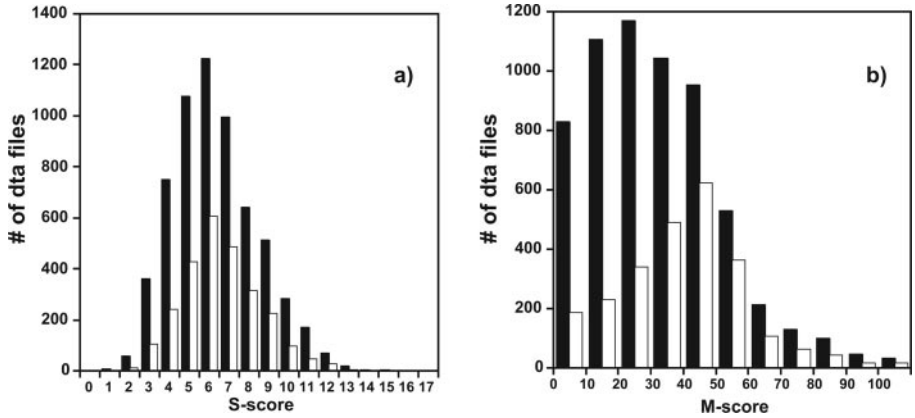

FIG. 5. *a*, distribution of S-scores for dta files for which Mascot found a peptide sequence *versus* the distribution of S-scores for dta files from which an RST was derived. *b*, distribution of Mascot scores for dta files for which Mascot found a peptide sequence *versus* the distribution of Mascot scores for dta files from which an RST was derived.

were lost using the S = 2 cutoff. The highest scoring Mascot identification that was discarded had an M-score of 21. We should note here that Mascot searches were made against the non-redundant data base and with oxidation of methionine chosen as the only viable modification, and so peptides with other modifications have no chance of being correctly identified. These peptides could account for the difference between the "complementary pairs exist" and the "Mascot found" distributions. The fact that this difference (data not shown) has a significantly lower mean (4.7) could be due to the generally inferior fragmentation of modified peptides.

Reliable Sequence Tag—The RST filter is in our opinion the most powerful filter in terms of verification. The maximum step length of the tag, 398 Da, is below the value of 575 Da selected for the S-score. It was chosen to reduce the number of different amino acid combinations that can fit into a given interval and thus increase the S-score applicability to smaller peptides. The value 398 Da is 1 Da below 399 Da, which is the nominal mass of seven glycines ensuring that the number of amino acids will not exceed six.

Fig. 5, *a* and *b*, shows the same two distributions plotted differently. The first is the total distribution of peptides for which Mascot suggested a sequence (*black columns*), and the second is the distribution of peptides for which a sequence was suggested and an RST existed (*light columns*). Here the RST existed in 40% of the cases. Note that although the average S-scores were similar for both distributions (Fig. 5*a*), the average M-score for the RST-backed distribution was much higher (38 *versus* 27). Only 41% of the total distribution is above the Mascot-suggested threshold of 34, whereas 69% of RST-containing spectra gave hits above this threshold. Thus the mere presence of RST increased the probability of positive ID by Mascot by more than 50%.

Note in Fig. 5*b* that the RST distribution seems to have abnormally high values at low M-scores. This phenomenon is explained by Fig. 6, which shows how the RST distribution partitions into two different distributions, one that confirms the Mascot suggested sequence and one that conflicts with it. If we assume that the part of the conflicting distribution, which stretches past the 95% significance level of M = 34, is entirely

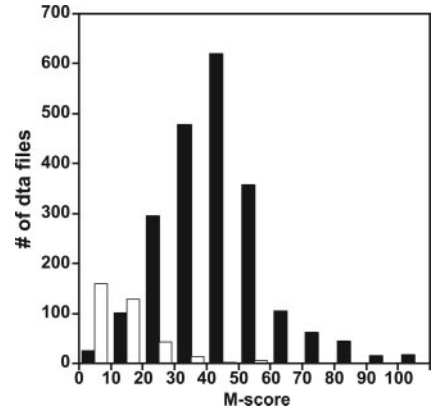

FIG. 6. **Distribution of Mascot scores for dta files for which the Mascot-suggested peptide sequence was confirmed by the RST (*black columns*) *versus* the distribution of Mascot scores for which the Mascot-suggested peptide sequence was rejected by the RST (*light columns*).**

due to the shortcoming (incorrectness) of the sequence tag and evaluate the reliability of the RST as a quotient between the number of "wrong" RSTs and "correct" RSTs (those that complied with the Mascot-suggested sequence), the reliability of the RST would be 98.6%. However, because the above assumption that the conflict is solely due to the shortcoming of the RST is certainly wrong (the reliability of Mascot-suggested sequences is only 95%), the actual reliability of RST is surely higher than 98.6%.

The benefits of the RST are 2-fold. It confirms the direction of the Mascot-suggested sequence, which is important because a common cause for false positives is the mix-up of *y*- and *b*-ions (this becomes obvious when conducting reverse data base searches (22, 23)). The RST also ensures that the ions that are its building blocks have not been omitted by Mascot when selecting the sequence from the data base. Because the probability of these ions being noise peaks or peaks not related to the peptide is so small (≈1.4%), their absence in the Mascot-suggested sequence is an almost sure sign of a false positive. In summary, we found an RST for 40% of the Mascot-identified MS/MS data (17% of all MS/MS data), and the reliability of the identifications supported by these tags is estimated to ~98.6%.

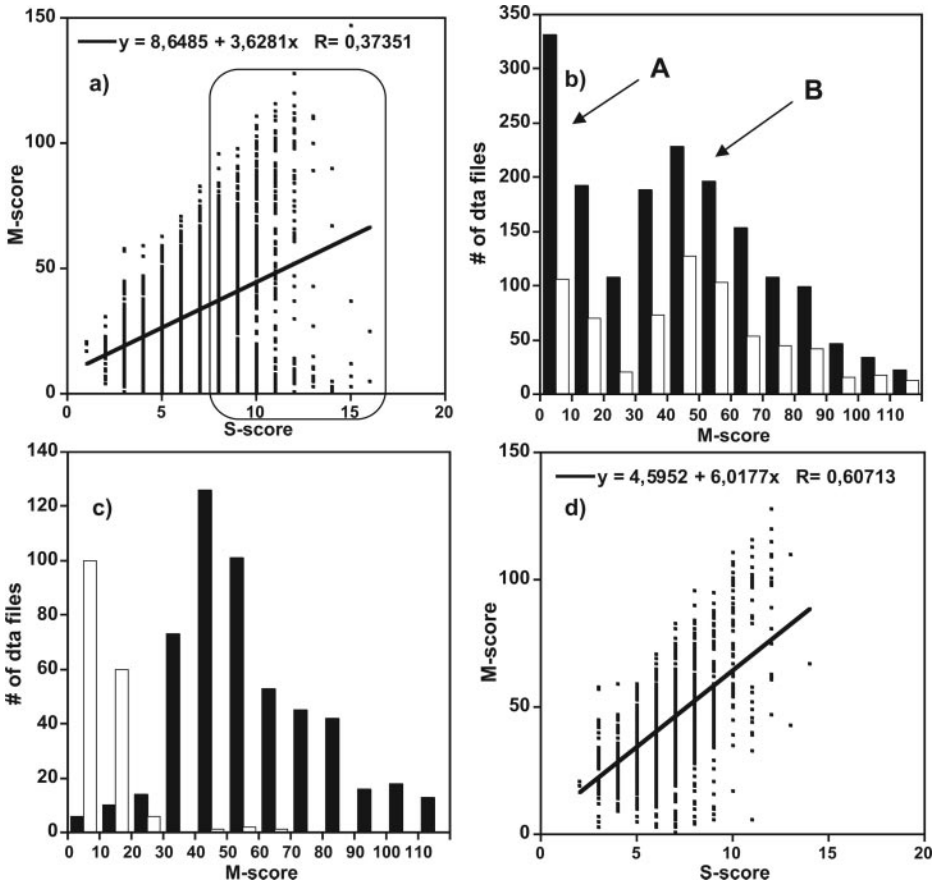

FIG. 7. *a*, a scatter plot showing the correlation between the S-score values and Mascot score values of all dta files for which Mascot found a peptide sequence. *b*, the Mascot score distribution of all dta files with S > 7 and for which Mascot found a sequence (*black columns*) *versus* the subset of this distribution for which an RST existed (*light columns*). *c*, the part of the *light columns* distribution in *b* for which the RST confirmed the Mascot-suggested sequence (*black columns*) *versus* the complementary part to the same distribution for which the RST rejected the Mascot-suggested sequence (*light columns*). *d*, a scatter plot showing the correlation between the S-score values and Mascot score values of all dta files for which Mascot found a peptide sequence and from which an RST was derived and confirmed the Mascot sequence.

Revealing Modifications with RST—As an example, the peptide LFSVVADDR was identified by Mascot with a below threshold M-score of 13. The S-score, however, was rather high (S = 7), and the spectra contained 12 complementary pairs of fragments. Moreover RST existed and gave possible sequences ([AY]‖[HP]‖[FS]‖[SMo])(I)(V)(A)(D)(D)([RV]‖[AAI]‖[GVV]) where Mo stands for oxidized methionine. In this notation, the amino acids in square brackets do not have a defined order, *e.g.* [AY] could in reality be both AY and YA. The ‖ sign means "or," *i.e.* ([RV]‖[AAI]‖[GVV]) means that at the C-terminal sequence could be [RV], [AAI], or [GVV] with I being either leucine or isoleucine. Thus this RST is consistent with many possible sequences, for instance PHLVADDAIA or YAIVADDVR. However, the Mascot-suggested sequence was not among them, thus conflicting with the RST.

In the conflict between RST and Mascot, RST is likely to be more trustworthy, and thus such spectra must be searched in an extended data base allowing for modifications/mutations. Because the RST suggests the presence of at least two aspartic acids, it is logical to assume the possibility of deamidation (24). Indeed a Mascot search allowing for this modification gave a positive identification with an M = 72, and the sequence YAIVANDVR successfully fitted all 12 complementary fragment masses. This peptide is present in an *E. coli* protein that was additionally identified by at least two unmodified peptides. As additional evidence of modification/mutation, the unmodified sequence was also identified from a different dta file with M = 62. This example shows the potential of using the RST to detect modifications and mutations.

Connection between S-score and M-score—The S-score is data base-independent and reflects how successful the peptide backbone fragmentation was. Mascot tries to fit the fragmentation pattern to a tryptic peptide in the data base, and the score reflects the probability of achieving a match of a given quality at random when searching the data base.

The S-score and M-score, however, are not completely independent because both increase when the number of true backbone fragments in an MS/MS spectrum increases. Thus one can expect a statistical correlation between M- and S-scores, which is interesting to derive as it could be used for *a priori* prediction of the M-score for individual spectra before a data base search. A straightforward correlation between the S-score and M-score is low with $r = 0.37$ (25) (Fig. 7*a*). However, this correlation is significant given the large amount of data points, >14000 (26). There appears to be a bifurcation of the data that is visibly detectable for high S-values. In Fig. 7*b*, the dta files with S-score above 7 were selected, and their M-score distribution was plotted (*black columns*). Clearly, the distribution is bimodal. The distribution with the lower mean (*A*) could arise if Mascot fits data to the wrong peptide. This

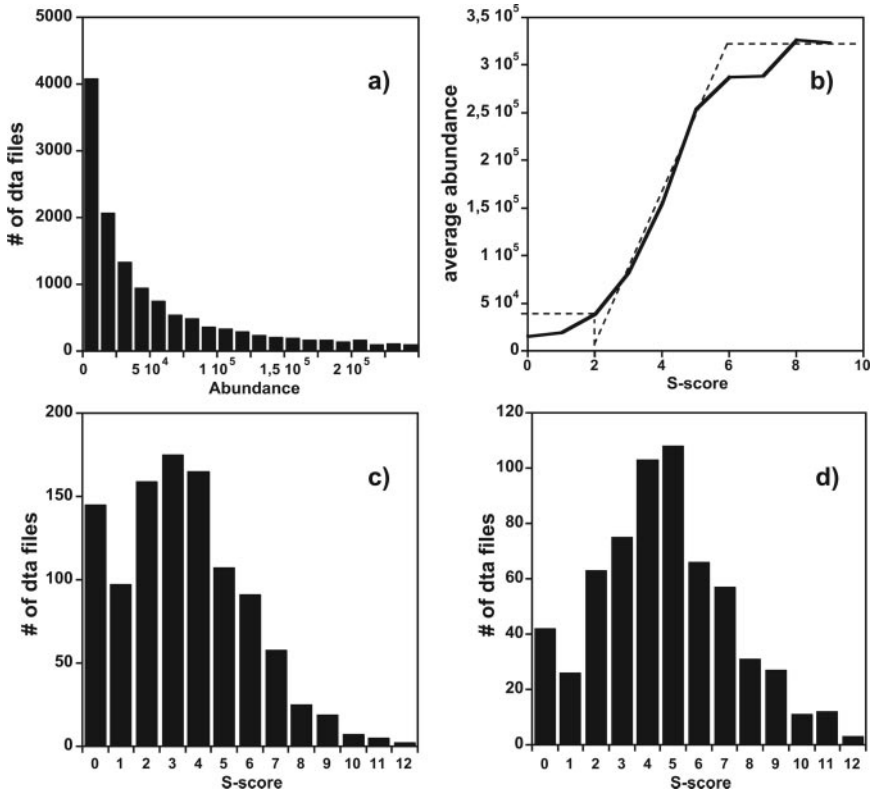

FIG. 8. *a*, the distribution of precursor abundances for all dta files. *b*, relation between the S-score of peptides and their average precursor abundances. *c*, the distribution of S-scores for dta files with precursor abundances between 20,000 and 30,000. *d*, the distribution of S-scores for dta files with precursor abundances between 30,000 and 40,000.

can happen if the true peptide is either not present in the data base or modified; in either case the fit Mascot makes would most likely be a poor one utilizing only part of the available data and thus resulting in a low M-score. To test this hypothesis we looked at the distribution of dta files for which an RST was found (Fig. 7*b*, *light columns*) and then identified the files (Fig. 7*c*) for which the RST affirmed the Mascot-suggested sequence (*black columns*) and for which it conflicted with the suggested sequence (*light columns*). As predicted, the low M-score distribution consisted mostly of false positives, whereas the high M-score distribution consisted mostly of correct identifications. Here false positive IDs were likely due to mutated and/or modified peptides. The ratio between the two distributions in Fig. 7*c* is 1:3. We believe that this ratio reflects the relative content of mutated and/or modified peptides in the test mixture. The relatively large content (~24.7%) is not surprising given the high rate of mutations in *E. coli*.

The following picture now takes shape. In Fig. 7*a*, there is at least one "true" and one "false" M-score distribution. For low S-scores, the means of these distributions are close and are not visibly resolved. However, for higher S-scores the mean of the true distribution grows, and the distance between the two distributions increases. The shape of the distribution of data for which an RST exists and confirms the Mascot-suggested sequence should be representative of the form of the true distribution. Thus the data with an RST that confirms the Mascot prediction should give a better correlation between the S-score and M-score. Indeed Fig. 7*d* shows that for such data the correlation increased dramatically from 0.37 to 0.61. A linear function that passes through the origin and fits the data gives a correlation of 0.60 and incline of 6.7. This implies that for correct Mascot identifications the average Mascot score is around $M = 6.7 \times S$. Thus the threshold value of $S = 2$ corresponds to $M = 13.4$.

Connection between S-score and Precursor Abundance—Naturally one would expect that the extent of detectable fragmentation of a peptide should relate to its precursor abundance. Indeed for an abundant precursor, more fragment ions will have a chance of making it over the noise level. Thus for higher S-scores, a higher average precursor abundance should be expected. Fig. 8*a* shows the distribution of precursor abundances in our data. The distribution scales to 1/abundance, although the automatic gain control (AGC) mode was used (this was supposed to accumulate the same number of precursors in each scan). Without AGC, the slope would be much steeper (with an ideal AGC, the distribution would be flat). Fig. 8*b* shows how the average precursor abundance increases with the S-score. The increase from $S = 0$ to $S = 2$ is comparatively slow followed by a steep almost linear increase from $S = 2$ to $S = 5$. At $S = 5$ a saturation occurs. The curve again suggests that $S = 2$ is a logical threshold value. At $S = 2$, the average intensity of precursor ions is ~30,000. It is tempting to suggest simplifying the analysis by using this intensity value as a threshold instead of $S = 2$. This, however, is not a good idea. In Fig. 8, *c* and *d*, the distribution of S-values is plotted for the abundance intervals

of precursor ions between 20,000 and 30,000 as well as 30,000 and 40,000. The bimodality of these distributions implies that a simple threshold for intensity values is not enough to discriminate between good and bad spectra. The bimodality is likely due to the presence of non-peptide precursors for which the abundances do not have to be low to produce low S-scores.

DISCUSSION

Two new concepts have been introduced, and their benefits have been made clear. The data base-independent S-scoring enables data prefiltering and classification and improves peptide and consequently protein identification. The combined use of M- and S-scoring provides positive sequence identification from >25% of all MS/MS data, a 40% improvement over traditional M-scoring performed on the same Fourier transform MS instrumentation. The number of proteins reliably identified from *E. coli* cell lysate increased by 29% compared with the traditional M-score approach. S-scoring provides a quantitative measure of the quality of fragmentation techniques such as the minimum abundance of the precursor ion, the MS/MS of which gives the threshold S-score value of 2. The calculation of the S-score is straight forward and fast, which makes it suitable for on-line evaluation of the quality of peptide fragmentation. The RST reveals a significant part of the false positive distribution and improves the reliability of peptide identifications with a below threshold M-score. Furthermore the identification of reliable sequence-tagged peptides that were falsely identified or not identified at all by Mascot can be pursued by alternative means utilizing the highly reliable tag. The high reliability and usefulness of the RST gives yet additional motivation for a broad use of orthogonal fragmentation techniques in proteomics.

Acknowledgments—Christopher M. Adams, Frank Kjeldsen, Thomas Köcher, and Oleg Silivra are acknowledged for fruitful discussions.

* This work was supported by Wallenberg Consortium North Grant WCN2003-UU/SLU-009 (to R. A. Z.). The purchase of the LTQ FT instrument was supported by a Knut och Alice Wallenbergs Stiftelse grant (to R. A. Z. and Carol Nilsson). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Box 583, S-75123 Uppsala, Sweden. Tel.: 46-18-471-5729; Fax: 46-18-471-5729; E-mail: Mikhail.Savitski@bmms.uu.se.

REFERENCES

1. Pandey, A., and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* **405,** 837–846
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
3. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R., III (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1,** 39–45
4. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
5. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
6. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
7. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75,** 768–774
8. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 13417–13422
9. Pappin, D. J. C., Hojrup, P., and Bleasby, A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3,** 327–332
10. Sunyaev, S., Liska, A. J., Golod, A., and Shevchenko, A. (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.* **75,** 1307–1315
11. Mann, M., and Wilm, M. (1994) Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399
12. Kristensen, D. B., Brond, J. C., Nielsen, P. A., Andersen, J. R., Sorensen, O. T., Jorgensen, V., Budin, K., Matthiesen, J., Veno, P., Jespersen, H. M., Ahrens, C. H., Schandorff, S., Ruhoff, P. T., Wisniewski, J. R., Bennett, K. L., and Podtelejnikov, A. V. (2004) Experimental Peptide Identification Repository (EPIR)—an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics* **3,** 1023–1038
13. Bern, M., Goldberg, D., McDonald, W. H., and Yates, J. R., III (2004) Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* **20,** i49–54
14. Marshall, A. G., and Hendrickson, C. L. (2002) Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *Int. J. Mass Spectrom.* **215,** 59–75
15. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **120,** 3265–3266
16. McLafferty, F. W., Fridriksson, E. K., Horn, D. M., Lewis, M. A., and Zubarev, R. A. (1999) Biochemistry. Biomolecule mass spectrometry. *Science* **284,** 1289–1290
17. Zubarev, R. A., Horn, D. M., Fridriksson, E. K., Kelleher, N. L., Kruger, N. A., Lewis, M. A., Carpenter, B. K., and McLafferty, F. W. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **72,** 563–573
18. Nielsen, M. L., Savitski, M. M., and Zubarev, R. A. (March 16, 2005) Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol. Cell Proteomics* **4,** 835–845
19. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379,** 466–469
20. Gentzel, M., Kocher, T., Ponnusamy, S., and Wilm, M. (2003) Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* **3,** 1597–1610
21. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000) Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 10313–10317
22. Peng, J. M., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
23. Shevchenko, A., Sunyaev, S., Loboda, A., Shevehenko, A., Bork, P., Ens, W., and Standing, K. G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73,** 1917–1926
24. Wright, H. T. (1991) Nonenzymatic deamidation of asparaginyl and glutaminyl residues in proteins. *Crit. Rev. Biochem. Mol. Biol.* **26,** 1–52
25. Pearson, K. (1896) Mathematical contributions to the theory of evolution: III. regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lond. A Math. Phys. Sci.* **187,** 253–318
26. Bewick, V., Cheek, L., and Ball, J. (2003) Statistics review 7: correlation and regression. *Crit. Care* **7,** 451–459