research articles Journal of proteome research

Rapid Validation of Protein Identifications with the Borderline Statistical Confidence via De Novo Sequencing and MS BLAST Searches

Natalie Wielsch,[†,‡] Henrik Thomas,[†,‡] Vineeth Surendranath,[†] Patrice Waridel,[†] Ari Frank,[§] Pavel Pevzner,[§] and Andrej Shevchenko*[,†]

Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany, Department of Computer Science & Engineering, University of California—San Diego, La Jolla, California 92093

Received April 28, 2006

Protein identifications with the borderline statistical confidence are typically produced by matching a few marginal quality MS/MS spectra to database peptide sequences and represent a significant bottleneck in the reliable and reproducible characterization of proteomes. Here, we present a method for rapid validation of borderline hits that circumvents the need in, often biased, manual inspection of raw MS/MS spectra. The approach takes advantage of the independent interpretation of corresponding MS/MS spectra by PepNovo de novo sequencing software followed by mass spectrometry-driven BLAST (MS BLAST) sequence-similarity database searches that utilize all partially inaccurate, degenerate and redundant candidate peptide sequences. In a case study involving the identification of more than 180 *Caenorhabditis elegans* proteins by nanoLC −MS/MS analysis on a linear ion trap LTQ mass spectrometer, the approach enabled rapid assignment (confirmation or rejection) of more than 70% of Mascot hits of borderline statistical confidence.

Keywords: de novo sequencing • database searching • borderline hits • MS/MS • MS BLAST • PepNovo

Nanoflow liquid chromatography−tandem mass spectrometry (nanoLC−MS/MS) is employed in a variety of bottom-up proteomics projects (reviewed in refs 1−4). Individual proteins or protein mixtures are usually digested in-solution or in-gel with proteolytic enzyme(s) and recovered peptides separated by one-dimensional or multidimensional chromatography, interfaced on-line to a tandem mass spectrometer (reviewed in ref 5). The mass spectrometer is operated under data-dependent (also termed information-dependent) control: a typical acquisition cycle consists of a survey scan that usually acquires a mass spectrum or a precursor ion spectrum followed by real-time selection of precursors for the series of subsequent MS/MS experiments. Linear ion trap instruments,[6] or hybrid instruments, such as quadrupole time-of-flight,[7] or recently introduced linear ion trap−Fourier transform or linear ion trap−orbitrap,[8−10] typically acquire a few thousand tandem mass spectra in a single nanoLC−MS/MS run. The acquired pool of MS/MS spectra is then submitted to unattended database searches by a dedicated software (reviewed in refs 11,12), which matches individual spectra to peptide sequences in a database and scores the statistical confidence of hits in respect to a randomly occurring similarity. If the software fails to match a spectrum to any database sequence, error-tolerant searches could be applied to circumvent possible discrepancies between peptide sequences in a database and the actual sequences of fragmented peptides (reviewed in ref 13).

Regardless of the employed algorithm and its software implementation, database searching is, in essence, a probabilistic process in which the confidence of hits is evaluated by the comparison of some matching quality scores against empirical or semiempirical statistical significance thresholds.[14−20] While the confidence in protein hits matched by a few high-scoring peptide spectra is usually undisputed, the assignment of statistically nonsignificant hits or borderline hits is far more problematic.[21] The conformity between a spectrum and a database sequence depends on the number and relative abundance of matched and mismatched fragments. However, when peptide sequencing is performed at the low-femtomole level, MS/MS spectra of bona fide peptide precursors are often contaminated by co-selected background ions,[22] and peak intensity ratios are affected by poor ion statistics.[23] Thus, it is difficult to tell without, often biased, manual inspection of raw MS/MS spectra, if a low-scoring hit represents a bona fide match to a low-quality spectrum or it is a false positive. The problem is even more severe if searches are performed against a species-restricted, rather than comprehensive, sequence database. It is not uncommon that protein preparations are rich in contaminants originating from exogenous, totally unrelated species, such as human and sheep keratins, fragments of proteolyic enzymes, antibodies, fragments of expression vectors or proteins from host organisms.[24]

* Corresponding author. E-mail: shevchenko@mpi-cbg.de.
[†] Max Planck Institute of Molecular Cell Biology and Genetics.
[‡] equal contribution of these authors.
[§] University of California—San Diego.

10.1021/pr060200v CCC: $33.50
Published on Web 07/22/2006

Here, we present a method for rapid evaluation of hits with the borderline statistical confidence that are produced by conventional database searches. It takes advantage of the independent interpretation of corresponding MS/MS spectra by de novo sequencing software followed by sequence-similarity database searching. De novo sequencing algorithms[25−31] do not rely upon correlating the masses of fragment ions to database sequences. Instead, they deduce stretches of peptide sequences directly from the interpreted spectra. Since de novo sequencing is seldom fully accurate, we combined it with mass spectrometry-driven BLAST (MS BLAST) sequence similarity searching protocol,[32] which utilizes redundant, degenerate and partially accurate peptide sequence candidates and employs an independent scoring scheme for evaluating the confidence of database searching hits.[33] We demonstrated that a combination of Mascot software,[34] de novo sequencing software PepNovo,[27] and MS BLAST, bundled by a simple scripted interface, enabled rapid and efficient validation of a large number of borderline hits, produced by matching of one or two MS/MS spectra with marginal statistical significance.

Materials and Methods

Chemicals. Cleland's reagent (dithiothreitol, DTT) was obtained from Merck (Darmstadt, Germany). Others chemicals were from Sigma-Aldrich (Munich, Germany). Bovine trypsin (sequencing grade) was from Roche (Mannheim, Germany). Solvents for liquid chromatography were of Lichrosolv grade; formic and trifluoroacetic acids were purchased from Merck (Darmstadt, Germany).

Protein Dataset. Proteins were isolated from *Caenorhabditis elegans* worms by affinity chromatography in an ongoing collaboration project with Prof. A. Hyman's laboratory in MPI of Molecular Cell Biology and Genetics, Dresden, Germany. Proteins were separated by one-dimensional SDS-polyacrylamide gel electrophoresis, and protein bands were visualized by Coomassie Brilliant Blue R250 staining. The excised bands were in-gel-digested with trypsin.[35, 36] Tryptic peptides, recovered from the gel pieces by extraction with 5% formic acid and acetonitrile, were dried in a vacuum centrifuge and stored at −20 °C until analyzed.

Analysis by NanoLC−MS/MS. Tryptic digests were redissolved in 10 μL of 0.05% TFA, and 4 μL were injected into a nanoLC−MS/MS Ultimate system (Dionex, Amsterdam, The Netherlands) interfaced on-line to a linear ion trap LTQ (ThermoElectron Corp., San Jose, CA). Peptides were first loaded onto a 1 mm × 300 μm i.d. trapping microcolumn packed with C18 PepMAP100 5 μm particles (Dionex) in 0.05% TFA at the flow rate of 20 μL/min. After a 4 min wash, they were back-flush-eluted and separated on a 15 cm × 75 μm i.d. nanocolumn packed with C18 PepMAP100 3 μm particles (Dionex) at the flow rate of 200 nL/min using the following mobile phase gradient: from 5 to 20% of solvent B in 20 min, 20−50% B in 16 min, 50−100% B in 5 min, 100% B during 10 min, and back to 5% B in 5 min. Solvent A was 95:5 H_2O/acetonitrile (v/v) with 0.1% formic acid; solvent B was 20:80 H_2O/acetonitrile (v/v) with 0.1% formic acid. Peptides were eluted into the mass spectrometer via a dynamic nanospray probe (Thermo Electron Corp.). A silicatip uncoated needle (20 μm i.d., 10 μm tip ID) (New Objective, Woburn, MA) was used with a spray voltage of 1.8 kV, and the transfer capillary temperature was set at 200 °C. Data-dependent acquisition was controlled by Xcalibur 1.4 software (ThermoElectron Corp.). The acquisition cycle consisted of a survey scan covering the range of m/z 350−1500 followed by MS/MS fragmentation of the three most intense precursor ions under the relative collision energy of 35%, triggered by a minimum signal threshold of 500 counts with the isolation width of 4.0 amu. Spectra were acquired under automated gain control (AGC) in one microscan for survey spectra and in three microscans for MS/MS spectra, with maximal ion injection time of 100 ms. The m/z of fragmented precursor ions were dynamically excluded for a further 60 s, but otherwise no pre-defined exclusion lists were applied. Spectra were exported as *dta* files using BioWorks 3.1 software (Thermo Electron Corp.) under the following settings: peptide mass range, 500−3500; minimum total ion intensity threshold, 1000; minimum number of fragment ions, 15; precursor mass tolerance, 1.4 amu; group scan, 1; minimum group count, 1.

Protein Identification by Mascot Database Searches. Tandem mass spectra were searched against an MSDB database (updated May 15, 2005; contains 2 011 425 protein sequence entries) by Mascot v. 2.1 software (Matrix Science Ltd., London, U.K.) installed on a local 2 CPU server. Mass tolerance for precursor and fragment ions was 2.0 and 0.5 Da, respectively; instrument profile, ESI-Trap; fixed modification, carbamidomethyl (cysteine); variable modification, oxidation (methionine). Where specified, searches were performed against a subset of *C. elegans* proteins that comprised 30 304 protein sequence entries.

De Novo Peptide Sequencing and MS BLAST Searches. Where specified, files in *dta* format were converted into Mascot generic format (*mgf*) and sequenced de novo by a modified version of PepNovo software[27] installed on a desktop (Pentium IV) PC. A single MS/MS spectrum was typically interpreted de novo in less than 0.5 s, and up to seven partially redundant candidate sequences were produced. To each interpreted spectrum, PepNovo assigned a quality score, which stands for the expected number of confidently determined amino acid residues in the most accurate sequence proposal. This score was derived from the sum of the probabilities of the individual amino acids being correct, which were computed using a logistic regression model.[37] Candidate sequences were then edited according to MS BLAST conventions and merged into a single search string in arbitrary order.[33,38] MS BLAST searches were performed against nr database at http://genetics.bwh.harvard.edu/msblast/ under the following settings: Scoring Table, 99; Filter, none; Expect, 1000. Statistical significance of hits was evaluated according to MS BLAST scoring scheme.[33] A typical search with a query of seven candidate sequences required less than 15 s to complete.

Computational Evaluation of the PepNovo/MS BLAST Validation Performance. A simulation dataset was built out of 100 high-quality peptide spectra, each represented by a single *dta* file. Upon Mascot database search, each spectrum unequivocally hit a single peptide sequence of, on average, 12 amino acid residues with the ions scores above 70. In each spectrum, peaks with relative intensities below 1% of the base peak intensity were declared noise, and their absolute intensity was left unchanged, whereas a dedicated script reduced the absolute intensity of other peaks with the step of 1% and, hence, produced the series of 100 spectra with the gradually altered signal-to-noise ratios. Their *dta* files were merged into a single *mgf* file and submitted to Mascot search, and ions scores of spectra matched to the correct database peptide sequence were registered. In parallel, the same *mgf* file was sequenced de novo by the PepNovo program in a batch mode,

recording up to seven sequence candidates for each interpreted spectrum. PepNovo scores of predicted sequences were registered, and sequences were merged into query strings and submitted to MS BLAST searches. The outcome of MS BLAST searches was sorted into three groups as follows: where MS BLAST produced a hit that was also confident according to MS BLAST scoring scheme (first group); where the target peptide was listed in the output of the MS BLAST search as a borderline or nonconfident hit (second group); or where the target protein was not hit by MS BLAST at all (third group). In each series, we aimed to identify (if possible) the two spectra with the lowest signal-to-noise ratios that belonged to the first and second groups and registered their ions scores (Mascot), sequence quality scores (PepNovo), and MS BLAST scores (solely for the reference).

The same simulation routine was applied to all 100 high-quality spectra from the initial dataset.

Results and Discussion

Could De Novo Sequencing Validate Borderline Hits Produced by Conventional Database Searches? Conventional database search algorithms (reviewed in refs 11,12) rely upon matching peaks of fragment ions (considering their m/z or both m/z and intensities) to peptide sequences from database entries, whereas de novo sequencing algorithms deduce and score stretches of peptide sequence independently of available sequence resources (reviewed in ref 39). Both algorithms benefit from better representation of bona fide fragment peaks and lower chemical noise[40] in analyzed spectra. We reasoned, however, that, even if de novo interpretation of the spectrum is ambiguous, it still could be employed as an independent means to cross-verify the Mascot hits if combined with a sequence-similarity searching tool that tolerates redundancy and partial inaccuracy of candidate peptide sequences.

To assess if a combination of PepNovo and MS BLAST could validate Mascot hits with marginal ions scores, we composed a dataset comprising 100 high-quality tandem mass spectra that unequivocally matched sequences of full tryptic peptides in a database. In each spectrum we altered in silico the actual signal-to-noise level by gradually decreasing the intensities of matching peaks, while the abundance of peaks of chemical noise was fixed. Effectively, we simulated the extreme scenario, in which the protein identification solely relied on matching a single spectrum of marginal quality.

Each series of spectra with perturbed signal-to-noise ratios was subjected, in parallel, to Mascot searches and de novo interpretation by PepNovo software. Up to seven sequence candidates per each interpreted spectrum were merged into a query string, which was then submitted to MS BLAST search.

Within each series, we aimed to determine the Mascot ions score and PepNovo quality score for the two spectra having the lowest signal-to-noise ratios, whose PepNovo sequencing and MS BLAST searching either confidently identified the correct peptide in a comprehensive database or listed the correct peptide among the top 50 nonconfident hits in the MS BLAST output (Figure 1). However, in several cases, such spectra were not identified. On several occasions, PepNovo/MS BLAST failed to match the expected sequence by interpreting even the initial high-quality spectrum, or the expected peptide was missing among nonconfident hits in the MS BLAST output. Therefore, the actual number of data points in Figures 2 and 3 specified in the corresponding figure legends was less than the expected 100.

We first checked if Mascot ions scores and PepNovo quality scores correlated when both interpretations of the same marginal quality spectrum pointed to the same correct peptide sequence (Figure 2). Although weak correlation was observed, we noticed that PepNovo scores corresponding to spectra with a given Mascot score (or vice versa) varied within a broad range of values. This indicated that the two interpretations were, indeed, complementary and in many instances could independently cross-validate each other.

Figure 3 presents cumulative distributions of PepNovo scores (panel a) and Mascot ions scores (panel b) obtained for the same dataset of in silico modified peptide spectra (Figure 1). They provide a complementary view on the ability of MS BLAST (Figure 3a) and PepNovo/MS BLAST combination (Figure 3b) to positively validate the assignment of spectra, depending on their PepNovo scores and Mascot ions scores, respectively.

More than 60% of spectra, in which candidate peptide sequences were produced with PepNovo scores above 8, were confidently matched to the correct protein entries by MS BLAST (Figure 3a), and for almost 80% of these spectra, correct peptide sequences were listed in search outputs. Once PepNovo scores exceeded 10, more than 90% of these spectra were confidently matched. This provided us with a qualitative estimate of the de novo interpretation reliability, irrespective of the actual Mascot ions scores of examined spectra.

Using the same spectra dataset, we plotted the cumulated proportion of positive PepNovo/MS BLAST assignments of spectra against their Mascot ions scores (Figure 3b). Note that ions scores do not depend on the database size, whereas thresholds of statistical confidence of Mascot searches do (Figure 3b).

To positively identify a protein in a comprehensive (all species) database, the ions score of a one peptide hit should exceed a relatively high threshold (>53), even at the moderate $p < 0.05$. Therefore, positive protein identifications with one or two matched peptides would require exceptional quality of corresponding MS/MS spectra, and therefore, false negatives are common. For searches in smaller, species-restricted databases, threshold scores are lower (Figure 3b). These searches, however, often produce false positives by matching the spectra of peptides from exogenous protein contaminants to sequences of the assumed organism.

Figure 3b suggests that approximately 80% of borderline (potentially, false negative) one-peptide hits produced by searches against a comprehensive database should be directly verifiable via de novo sequencing and MS BLAST. Although the expected success rate also remains substantial for smaller species-restricted databases, de novo verification would be most helpful in discriminating against false positive, rather than validating false negative hits. Ions scores of false-positive hits are often marginal, since they are falsely matched to wrong database entries, although rich patterns of fragment ions together with low chemical noise enable confident readout of long stretches of their sequences.

The Protein Identification Workflow. A protein identification routine employed in this work (Figure 4) started with the stringent database search against a species-restricted database, which typically resulted in a few confident hits. We note that different proteomics laboratories adopted varying confidence criteria, even if the same software was used for database mining.[20,41] The independent validation step allowed us to use conserved criteria of positive protein identifications together with relatively loose selection of nonconfident hits, although

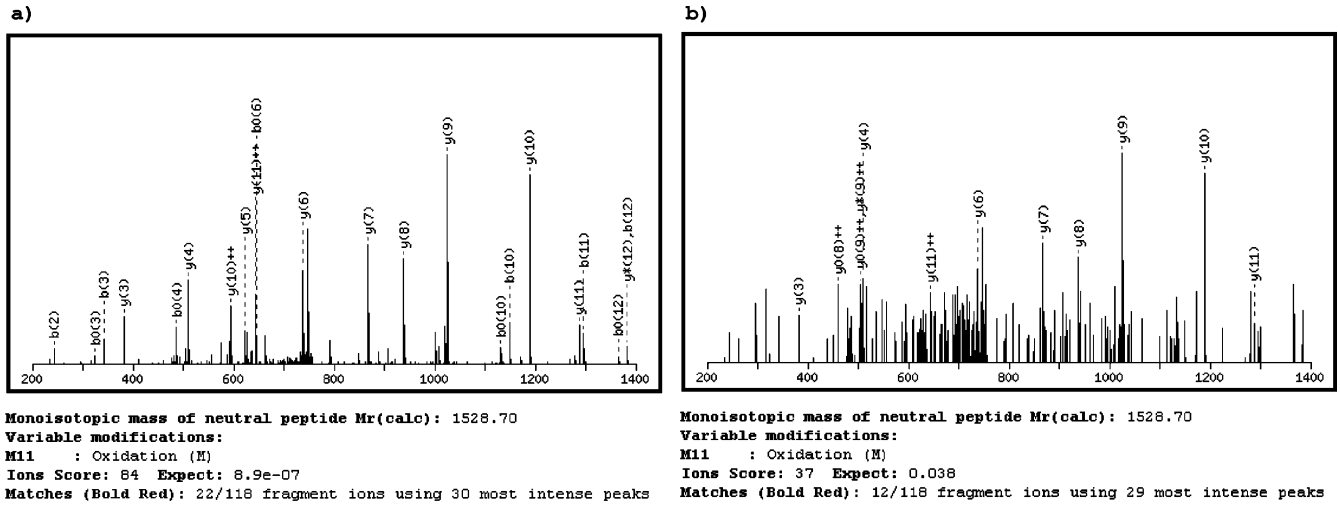


Figure 1. Altering MS/MS spectra for in silico simulation experiments. (a) The spectrum hit upon Mascot-search of the peptide (K)-ELVYSAEDLEMSK from *C. elegans* protein Snf5 with ions score of 84; (b) a spectrum with altered signal-to-noise ratio, produced from the spectrum in panel a by reducing the intensity of fragment ions by 95%, while maintaining the same intensity of noise peaks. Mascot search identified the same peptide, albeit the ions score was 37. (c) De novo interpretation of the spectrum in panel a by PepNovo software produced seven partially redundant candidate sequences, with the top candidate having a quality score of 11.3. According to MS BLAST conventions, (M.) stands for mono-oxidized methionine residues, and B stands for a generic trypsin cleavage site (arginine or lysine residues) preceding the peptide sequence. Since isobaric oxidized methionine and phenylalanine residues were not distinguished in ion trap spectra, both candidate sequences were included into the query string for MS BLAST search, which also produced a confident hit (MS BLAST confidence threshold score for a single reported HSP was 64). (d) The same procedure was applied to the modified spectrum from panel b. Both ions score and PepNovo score decreased, yet MS BLAST search was still able to produce a confident hit.

this strategy yielded a large number of borderline hits. Many of them were produced by matching one or two spectra, and therefore, we could use their ions scores as direct selection criteria (Figure 4).

To validate the selected borderline hits, corresponding *dta* files were fetched by Windows Shell Scripts developed in-house and re-submitted to another round of Mascot searches, now against a full database with unrestricted species specificity. The second search typically identified and removed good quality spectra of full tryptic peptides, originating from trypsin, keratin, GST, and other background proteins, which produced statistically confident hits in searches against a full database. The remaining spectra were interpreted de novo, and sequence candidates obtained by the interpretation of all spectra pertinent to the validated hit were merged into a single query[33,38] and searched against a comprehensive database by MS BLAST. The results of MS BLAST searches were interpreted as follows: if the same peptide as in the Mascot search was either confidently matched by MS BLAST, or was present in the output of the MS BLAST search, and the reported high-scoring segment pair[42] (HSP), which corresponds to the alignment of the database peptide sequence and the sequence deduced from MS/MS spectrum by its de novo interpretation,[33] covered at least 50% of the verified peptide sequence, then these hits were considered confirmed. Note that, while interpreting MS/MS spectra, PepNovo was set to produce complete sequence proposals. However, because of the poor quality of the target spectra, they were not necessarily fully accurate. They might also contain correct sequence stretches that, however, did not produce statistically significant alignments and therefore were not reported within an HSP. In most cases, the length of the aligned noninterrupted peptide sequences exceeded six amino acid residues. The Mascot hits were considered as false positives and rejected if MS BLAST searches either confidently hit another protein, or hit a common background protein and more than 50% of the peptide sequence (and, at least, 6 amino acid residues) were covered by the aligned HSP. The third criterion came from the consideration of the expected de novo interpretation accuracy, which is related to the PepNovo quality score.

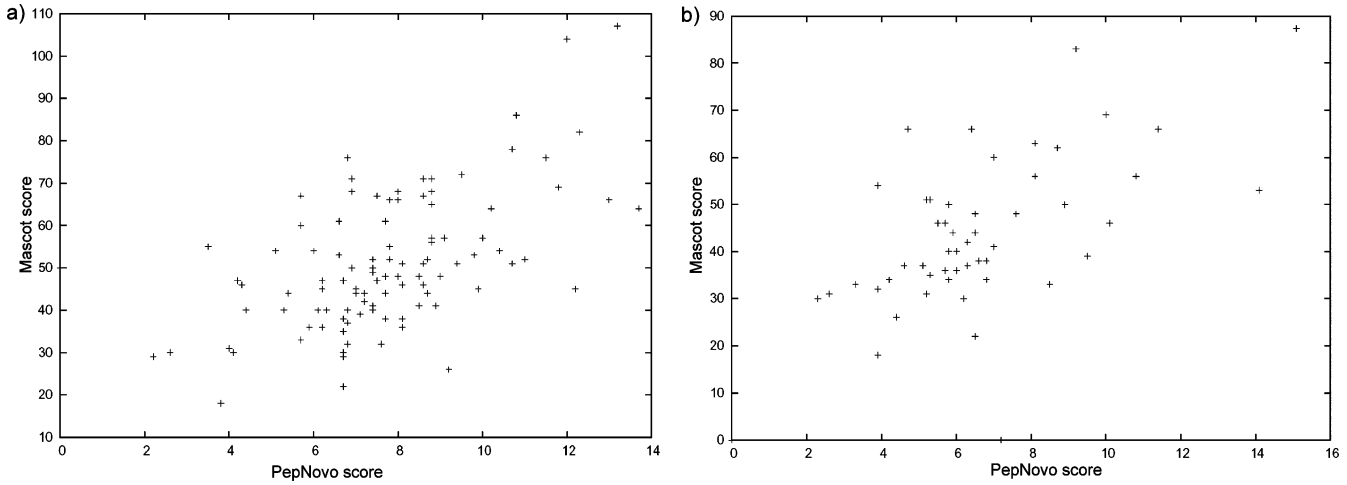


Figure 2. Plotted diagram of Mascot ions scores versus PepNovo sequence quality scores built using the series of simulated spectra (Figure 1, Figure 1S in Supporting Information) that enabled their confident (panel a, data for 98 spectra) and nonconfident (panel b, data for 48 spectra) assignment to the correct database sequences by MS BLAST.

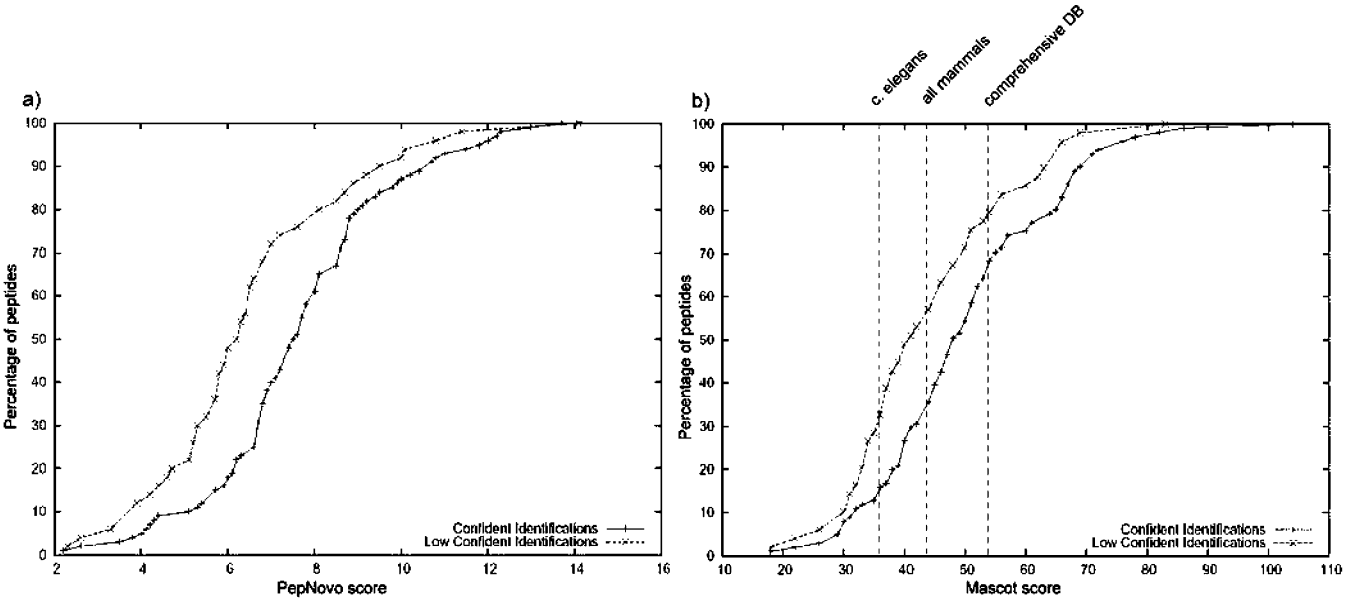


Figure 3. Cumulative distributions of confident and low confident MS BLAST hits obtained by searches with de novo sequences produced from tandem mass spectra with altered signal-to-noise ratio are plotted against their PepNovo scores (panel a) and Mascot ions scores (panel b). The dataset was the same as in Figure 2. Vertical bars in panel b stand for Mascot thresholds of statistically confident protein identifications supported by matching a single peptide ($p < 0.05$) in the organism-specific databases: *C. elegans* (30 304 proteins entries), threshold score of 36; all mammals (287 223 protein entries), threshold score of 43; a comprehensive (all species) database (2 011 425 protein entries), threshold score of 53.

If de novo interpretation of validated MS/MS spectra produced peptide sequence candidates with PepNovo scores above 10, then, according to Figure 3, it was expected that subsequent MS BLAST searches would confirm more than 90% of the corresponding Mascot hits. Otherwise, these hits were considered false positives, even if MS BLAST searches produced no significant alignments to other proteins. However, low PepNovo scores (practically, less than 5) indicated that, for any reason, PepNovo failed to produce a reliable sequence of sufficient length. In these cases, negative outcomes of MS BLAST searches were inconclusive and the hits remained unassigned.

How the Validation Method Works? To demonstrate the practical applicability of the proposed workflow, we present here as a case study the validation of the two nanoLC−MS/MS identifications of gel-separated *C. elegans* proteins that both relied upon matching a single tandem mass spectrum with the marginal ions score.

The protein Y6B3B.8 was identified by Mascot in a *C. elegans* protein database under the fixed trypsin cleavage specificity settings. The protein was hit by a single MS/MS spectrum with the ions score of 31 (Figure 5), while the proposed confidence threshold for *C. elegans* database was 36. Manual inspection of the spectrum suggested that almost all abundant peaks matched m/z of expected fragment ions. To further validate this hit, the corresponding spectrum was first searched against a comprehensive database. The search pointed to the same protein, albeit, because of the increased database size, the matching confidence dropped. Therefore, the hit was further

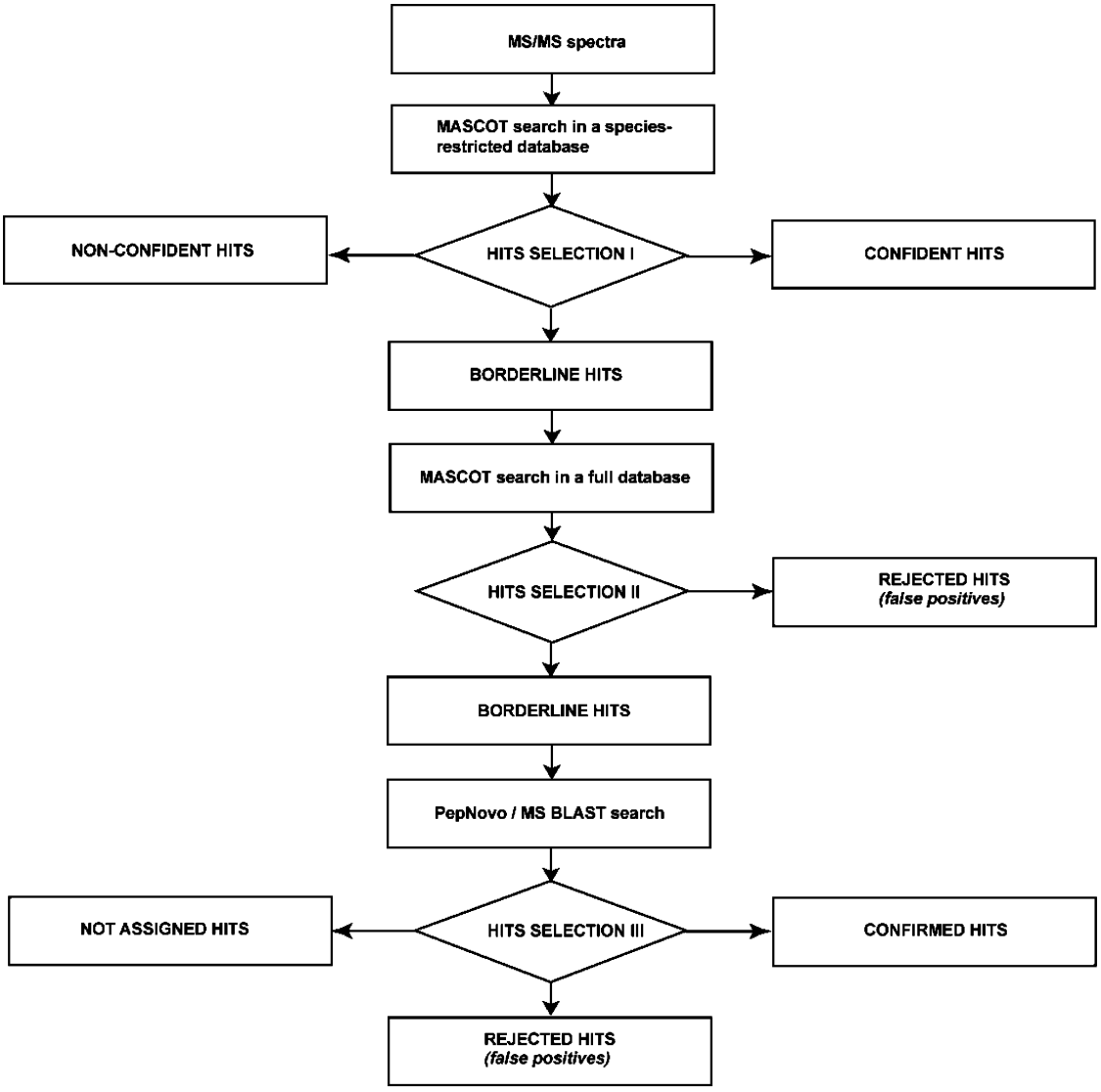


Figure 4. The protein identification workflow. Diamonds stand for the workflow junctions, where the following selection criteria were applied. Hit selection I: (1) confident hits, more than three peptides matched by Mascot with ions scores above the confidence threshold for a species-specific database (36 for *C. elegans* protein database), or at least one score was above the threshold for a comprehensive database (53 for MSDB). (2) Borderline hits: Mascot matched less than four peptides and the ions score of at least one peptide was within the range of ±30% of the threshold score (from 26 to 46 for *C. elegans*). (3) Nonconfident hits: the rest. Hit selection II: (1) rejected hits, the searched peptide confidently hit other than expected protein in a comprehensive database (ions score should exceed 53). (2) Borderline hits, the rest. Hit selection III: (1) confirmed hits, hits either confidently matching the expected protein by MS BLAST or in which the aligned HSP covered more than 50% of the expected peptide sequence spanning over more than six amino acid residues. (2) Rejected hits, common background proteins (trypsin, keratins, GST) matching the same criteria; or other proteins confidently matched by MS BLAST; or hits that did not match the expected peptide albeit their PepNovo scores were above 10. (3) Not assigned hits, the rest.

validated by PepNovo/MS BLAST (Figure 5b), which confidently hit a half-tryptic peptide VVEGNEQFISASK that originated from bovine trypsin, presumably via orifice fragmentation of the abundant autodigestion product LDEDNINVVEGNEQFISASK. Note that approximately the same number of peaks matched the expected fragment ions in panels a and c of Figure 5, which illustrates why manual inspection of hits is often biased and prone to interpretation flaws. To further check the MS BLAST identification, another Mascot search was performed without restricting the enzyme cleavage specificity. Despite higher ions score (67 for trypsin peptide vs 31 for *C. elegans* peptide), the hit was still nonconfident since the threshold score under the assumed settings was 74. The *Expect* value (the expected number of false-positive hits produced by searching a database with the given spectrum) was not improved and stayed well within the nonconfident range.

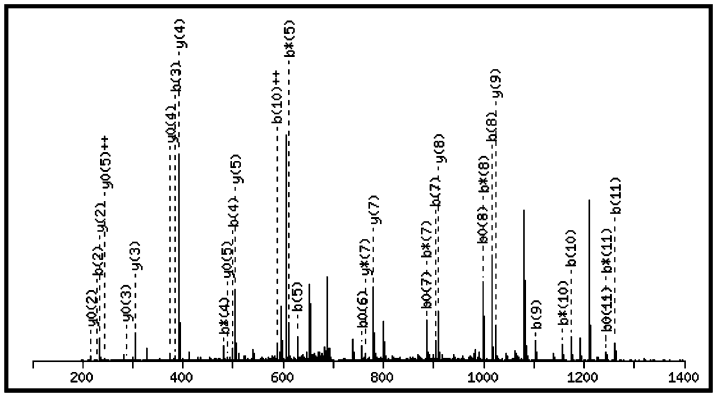
In another case, *C. elegans* protein C56G2.1 was identified by matching one peptide with below the threshold ions score of 31 (Figure 6). Mascot search against a comprehensive database with and without trypsin cleavage specificity restrictions also pointed to the same protein, although both ions scores were statistically insignificant. At the same time, de novo interpretation of the spectrum followed by MS BLAST search confidently hit the expected peptide sequence from C56G2.1 protein, thus, rescuing this, otherwise false negative, hit (Figure 6b).

Validating Borderline Hits at the Larger Scale. We then applied the PepNovo/MS BLAST validation routine in nanoLC−

a) F87991 **Mass:** 38297 **Score:** 31 **Queries matched:** 1

protein Y6B3B.8 [imported] - Caenorhabditis elegans

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
1628	704.38	1406.75	1406.72	0.04	1	31	0.2	1	R.DIRNEFQLSASK.R



Monoisotopic mass of neutral peptide Mr(calc): 1406.72
Ions Score: 31 **Expect:** 0.2
Matches (Bold Red): 33/126 fragment ions using 100 most intense peaks

b) de novo sequences (PepNovo)

PepNovo score 11.4

BVVEGNEQM.LSASK
BVVEGNEQFLSASK
BVVEGNEGAM.LSASK
BVVEGNEGAFLSASK
BVVEGGGEQM.LSASK
BVVEGGGEQFLSASK
BVVEGGGEGAM.LSASK

MS BLAST search trypsin, bovine

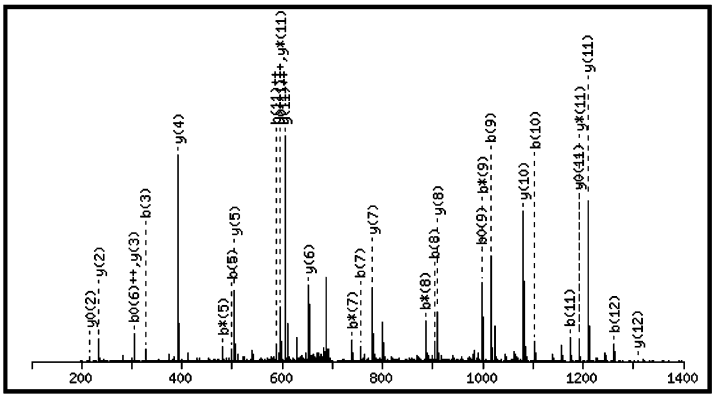
Score = 91 (47.6 bits)
Identities = 12/13 (92%), Positives = 13/13 (100%)

```
Query:    17 VVEGNEQFLSASK 29
             VVEGNEQF+SASK
Sbjct:    77 VVEGNEQFISASK 89
```

c) TRBOTR **Mass:** 24662 **Score:** 61 **Queries matched:** 1

trypsin (EC 3.4.21.4) precursor - bovine

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
1	704.38	1406.75	1406.70	0.05	0	67	0.3	1	N.VVEGNEQFISASK.S



Monoisotopic mass of neutral peptide Mr(calc): 1406.70
Fixed modifications: Carbamidomethyl (C)
Ions Score: 67 **Expect:** 0.3
Matches (Bold Red): 31/130 fragment ions using 54 most intense peaks

Figure 5. PepNovo/MS BLAST discarded a false-positive identification. (a) Mascot search performed against the *C. elegans* database hit the protein Y6B3B.8. Search against a full database also confirmed this hit. Trypsin was specified as proteolytic enzyme in both searches. (b) The same spectrum was interpreted de novo, and candidate sequences were submitted to MS BLAST search, which retrieved the half-tryptic peptide VVEGNEQFISASK from bovine trypsin as a single confident hit. (c) The same spectrum as in panel a with fragment ions matching the sequence of VVEGNEQFISASK.

MS/MS analysis of *C. elegans* proteins that were purified by immunnoaffinity chromatography and separated by one-dimensional SDS-polyacrylamide gel electrophoresis. In the analysis of 10 Coomassie-stained bands, 127 proteins were confidently identified and another 164 hits were regarded borderline, according to the criteria outlined above (Figure 4). The full list of borderline protein hits and their identification details are provided in Table 1S in Supporting Information. We note here that many of these hits were, potentially, of substantial biological interest as plausible substoichiometric subunits linking the isolated protein complex to a network of physical interactions.[24] However, searching MS/MS spectra against a comprehensive database revealed that the preparation was heavily contaminated with exogenous proteins, such as fragments of the GST construct, proteins from *Escherichia coli* (host organism in which the GST-fused bait protein was

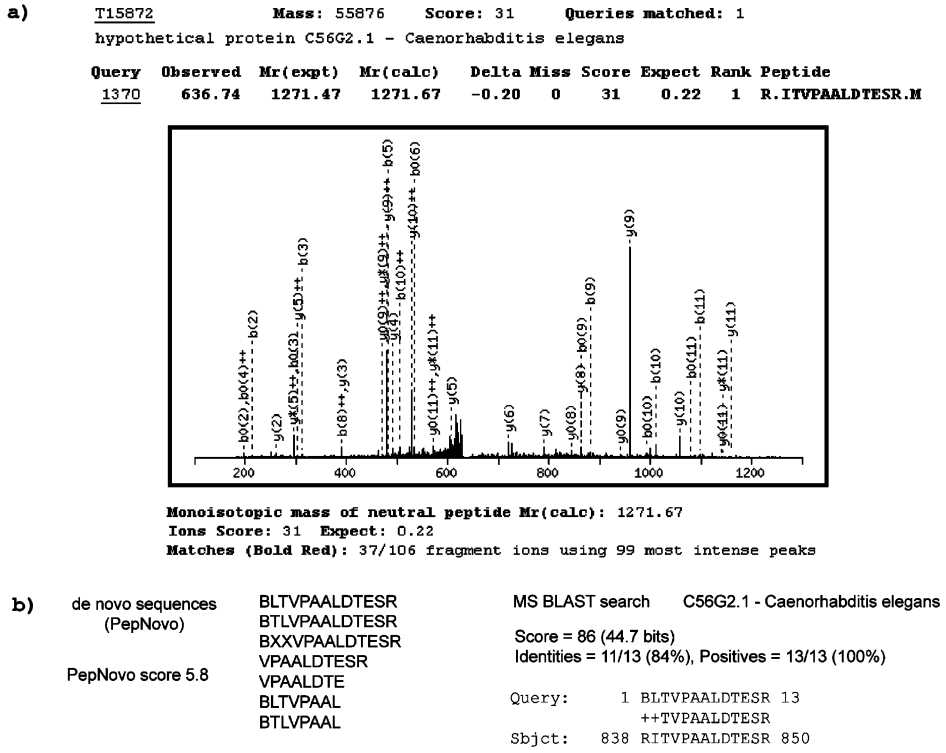


Figure 6. PepNovo/MS BLAST rescued a false-negative identification. (a) Mascot search against *C. elegans* database hit C56G2.1 protein with insignificant ions score. Search against a comprehensive database pointed to the same protein. (b) The same spectrum was interpreted de novo, and candidate sequences were searched by MS BLAST that confidently hit the same *C. elegans* protein.

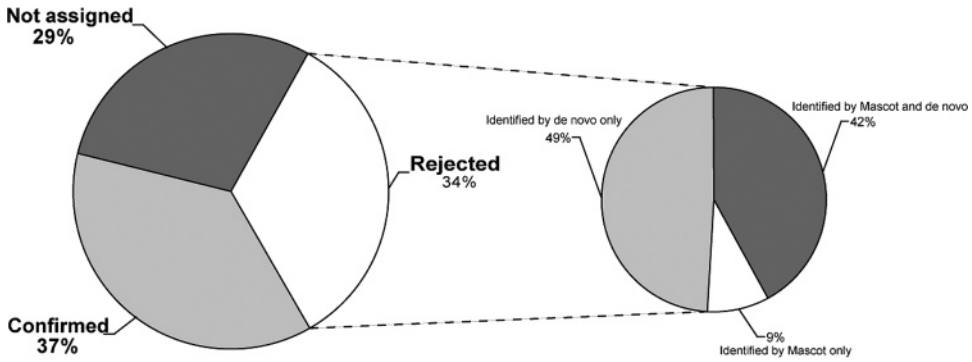


Figure 7. Validation of 164 borderline hits produced by Mascot searches against a database of *C. elegans* proteins. Confirmed: hits confirmed by PepNovo/MS BLAST method. "Rejected": hits were rejected if either Mascot confidently identified another protein in a full (all species) database (designated as "Identified by Mascot" at the inset), or by PepNovo/MS BLAST probing according to the workflow in Figure 5 (designated as "Identified by de novo"). "Not assigned": borderline hits for which both methods did not produce any conclusive identity evidence.

expressed), and human keratins. Figure 7 presents a distribution of 164 validated borderline hits: 37% of them were confirmed by PepNovo/MS BLAST, whereas another 34% were discarded as false positives (either by PepNovo/MS BLAST or by Mascot searches against a nonrestricted database), so that the percentage of borderline identifications that still remained ambiguous was reduced down to 29%. Among recognized false positives, 49% were identified only by PepNovo/MS BLAST, 42% were verifiable by Mascot searches against a full database as well as by PepNovo/MS BLAST, and 9% were identified only by Mascot searches, while PepNovo/MS BLAST failed to produce conclusive assignments.

Taken together, 116 borderline hits (71%) were confirmed or rejected, and the total number of ambiguous identifications was considerably reduced without any recourse to manual inspection of spectra. Thus, we concluded that a combination of de novo sequencing by PepNovo and MS BLAST searches efficiently complemented the conventional (Mascot) protein identification routine by providing an independent means of automated validation of hits with borderline statistical confidence, which substantially reduced the rates of both false-positive and false-negative identification.

Conclusion and Perspectives

Unrecognized false positives and borderline hits plague today's proteomics, and considerable efforts have been directed toward improving the statistical apparatus of database searching engines (reviewed in refs 20, 41, 43, 44) and deeper

understanding of peptide fragmentation pathways and their impact on the accuracy of spectrum-to-sequence matching.[45−49] However, probabilistic scoring only suggests the threshold of statistically reliable peptide assignments. Even if there is a fair chance that the particular hit can be a false positive, the statistics does not ensure it is necessary a false positive since insignificant scoring of the genuine peptide-to-spectrum match might occur. At the edge of the sensitivity and dynamic range of analytical instruments, the identification of proteins usually relies on matching of one or two marginal quality mass spectra.[5] It is therefore important to independently validate individual hits of potential biological interest irrespectively of statistical properties of both the spectra dataset and sequence database.

The idea of using de novo sequencing as a golden standard for database searching is not new,[25,28,50,51] yet it was seldom used in practical work since it almost never delivers the required accuracy and confidence of produced sequences.[39,52] Yet, when combined with sequence similarity searching tools, it provided the independent interpretation of MS/MS spectra that could validate the identifications relying upon matching of fragment ion patterns. In a case study, nanoLC−MS/MS sequencing of 10 in-gel digests of Coomassie-stained *C. elegans* protein bands identified, in total, more than 180 proteins of varying abundance. Using a combination of Mascot and PepNovo/MS BLAST searches, we were able to independently confirm or reject the assignment of 70% of borderline hits without manual inspection of raw MS/MS spectra. However, the method performance was inherently limited by the ability of de novo sequencing software to produce meaningful sequence candidates from tandem mass spectra with either insufficient fragment representation, or having too complex fragment patterns. Thus, it seems promising to employ simultaneously several independent peptide fragmentation methods within the same nanoLC−MS/MS experiment, which might increase the accuracy of de novo sequencing without compromising the analysis throughput and, presumably, sensitivity.[25]

Acknowledgment. We are grateful to Dr. Nurhan Ozlu and Prof. Tony Hyman (MPI CBG, Dresden) for providing a sample of *C. elegans* proteins and long-standing collaboration; to members of the Shevchenko laboratory for their expert support; and to Ms Judith Nicholls for critical reading of the manuscript. The work in the Shevchenko laboratory was supported by BMBF Grant PTJ-BIO/0313130 from BMBF and 1RO1GM070986-01A1 from NIH NIGMS.

Supporting Information Available: The full list of validated borderline hits, MS BLAST queries, and Mascot identification details. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.
(2) Yates, J. R., III; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 702−714.
(3) Steen, H.; Mann, M. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 699−711.
(4) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. *Nat. Methods* **2005**, *2*, 667−675.
(5) Peng, J. M.; Gygi, S. P. *J. Mass Spectrom.* **2001**, *36*, 1083−1091.
(6) Mayya, V.; Rezaul, K.; Cong, Y. S.; Han, D. *Mol. Cell. Proteomics* **2005**, *4*, 214−223.
(7) Chernushevich, I.; Loboda, A.; Thomson, B. *J. Mass Spectrom.* **2001**, *36*, 849−865.
(8) Hu, Q. Z.; Noll, R. J.; Li, H. Y.; Makarov, A.; Hardman, M.; Cooks, R. G. *J. Mass Spectrom.* **2005**, *40*, 430−443.
(9) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. *Mol. Cell. Proteomics* **2005**, *4*.
(10) Yates, J. R.; Cociorva, D.; Liao, L.; Zabrouskov, V. *Anal. Chem.* **2006**, *78*, 493−500.
(11) Sadygov, R. G.; Cociorva, D.; Yates, J. R., III, *Nat. Methods* **2004**, *1*, 195−202.
(12) Fenyo, D. *Curr. Opin. Biotechnol.* **2000**, *11*, 391−395.
(13) Liska, A. J.; Shevchenko, A. *Proteomics* **2003**, *3*, 19−28.
(14) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.
(15) MacCoss, M. J.; Wu, C. C.; Yates, J. R., III, *Anal. Chem.* **2002**, *74*, 5593−5599.
(16) Liu, H.; Sadygov, R. G.; Yates, J. R., III, *Anal. Chem.* **2004**, *76*, 4193−4201.
(17) Sadygov, R. G.; Liu, H.; Yates, J. R. *Anal. Chem.* **2004**, *76*, 1664−1671.
(18) Sadygov, R. G.; Yates, J. R., III, *Anal. Chem.* **2003**, *75*, 3792−3798.
(19) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
(20) Nesvizhskii, A. I.; Aebersold, R. *Drug Discovery Today* **2004**, *9*, 173−181.
(21) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *Mol. Cell. Proteomics* **2005**, *4*, 1180−1188.
(22) Chalkley, R. J.; Baker, P. R.; Hansen, K. C.; Medzihradszky, K. F.; Allen, N. P.; Rexach, M.; Burlingame, A. L. *Mol. Cell. Proteomics* **2005**, *4*, 1189−1193.
(23) Gentzel, M.; Kocher, T.; Ponnusamy, S.; Wilm, M. *Proteomics* **2003**, *3*, 1597−1610.
(24) Shevchenko, A.; Schaft, D.; Roguev, A.; Pijnappel, W. W. M. P.; Stewart, A. F.; Shevchenko, A. *Mol. Cell. Proteomics* **2002**, *1*, 204−212.
(25) Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Zubarev, R. A. *J. Proteome Res.* **2005**, *4*, 2348−2354.
(26) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327−342.
(27) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964−973.
(28) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594−2604.
(29) Spengler, B. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 703−714.
(30) Zhang, Z. *Anal. Chem.* **2004**, *76*, 6374−6383.
(31) Grossmann, J.; Roos, F. F.; Cieliebak, M.; Liptak, Z.; Mathis, L. K.; Muller, M.; Gruissem, W.; Baginsky, S. *J. Proteome Res.* **2005**, *4*, 1768−1774.
(32) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917−1926.
(33) Habermann, B.; Oegema, J.; Sunyaev, S.; Shevchenko, A. *Mol. Cell. Proteomics* **2004**, *3*, 238−249.
(34) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(35) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.
(36) Shevchenko, A.; Chernushevich, I.; Wilm, M.; Mann, M. *Mol. Biotechnol.* **2002**, *20*, 107−118.
(37) Frank, A.; Tanner, S.; Bafna, V.; Pevzner, P. *J. Proteome Res.* **2005**, *4*, 1287−1295.
(38) Shevchenko, A.; Sunyaev, S.; Liska, A.; Bork, P.; Shevchenko, A. *Methods Mol. Biol.* **2003**, *211*, 221−234.
(39) Standing, K. G. *Curr. Opin. Struct. Biol.* **2003**, *13*, 595−601.
(40) Johnson, R. S.; Davis, M. T.; Taylor, J. A.; Patterson, S. D. *Methods* **2005**, *35*, 223−236.
(41) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531−533.
(42) Altschul, S. F.; Gish, W. *Methods Enzymol.* **1996**, *266*, 460−480.
(43) Chalkley, R. J.; Hansen, K. C.; Baldwin, M. A. *Methods Enzymol.* **2005**, *402*, 289−312.
(44) Baldwin, M. A. *Mol. Cell. Proteomics* **2004**, *3*, 1−9.
(45) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R., III, *Anal. Chem.* **2003**, *75*, 1155−1163.
(46) Venable, J. D.; Yates, J. R., III, *Anal. Chem.* **2004**, *76*, 2928−2937.
(47) Zhang, Z. *Anal. Chem.* **2004**, *76*, 3908−3922.
(48) Gibbons, F. D.; Elias, J. E.; Gygi, S. P.; Roth, F. P. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 910−912.
(49) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat Biotechnol* **2004**, *22*, 214−219.
(50) An Thieu, V.; Kirsch, D.; Flad, T.; Muller, C.; Spengler, B. *Angew. Chem., Int. Ed.* **2006**, *45*, 3317−3319.
(51) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10313−10317.
(52) Shevchenko, A.; Chernushevich, I.; Wilm, M.; Mann, M. *Methods Mol. Biol.* **2000**, *146*, 1−16.

PR060200V