DOI 10.1002/pmic.200900473 *Proteomics* 2010, *10*, 2712–2718

TECHNICAL BRIEF

Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets

Kang Ning[1], *Damian Fermin*[1] *and Alexey I. Nesvizhskii*[1,2]

[1] Department of Pathology, University of Michigan, Ann Arbor, MI, USA
[2] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

In a typical shotgun proteomics experiment, a significant number of high-quality MS/MS spectra remain ''unassigned.'' The main focus of this work is to improve our understanding of various sources of unassigned high-quality spectra. To achieve this, we designed an iterative computational approach for more efficient interrogation of MS/MS data. The method involves multiple stages of database searching with different search parameters, spectral library searching, blind searching for modified peptides, and genomic database searching. The method is applied to a large publicly available shotgun proteomic data set.

Received: July 2, 2009
Revised: April 1, 2010
Accepted: April 22, 2010

Keywords:
Bioinformatics / Interactive database search / Novel peptides / Peptide polymorphisms / Spectral quality assessment / Unassigned spectra

A typical shotgun proteomic experiment involves generation of thousands of tandem mass spectra. The development of computational tools has made automatic identification of peptides from these spectra a routine approach. Continuous efforts are made to improve the sensitivity and specificity of peptide identification methods, and methods to estimate the error rates in the resulting data sets [1, 2]. However, despite improvements in MS instrumentation and peptide identification algorithms, a significant number of MS/MS spectra in any large-scale study remain ''unassigned'' (*i.e.* no high confidence peptide identification) [3]. A significant fraction of these spectra are of high quality, as measured using various spectral features [4–6], and additional studies are necessary to gain a better understanding of their nature and significance.

Peptides are most often identified from MS/MS spectra using sequence database searching, either in a direct fashion [7–9] or with the aid of sequence tags [10–12]. High-quality spectra may remain unidentified in a typical data analysis pipeline due to several reasons: inaccurate charge state or precursor ion m/z measurement, constrained database search parameters (*e.g.* search for tryptic peptides only), the presence of chemical modifications or PTMs, and incompleteness of the searched protein sequence database [3, 13, 14]. The last two categories, peptides containing PTMs and novel peptides, are particularly interesting. Such peptides can be identified using *de novo* sequencing, error-tolerant (or ''blind'') database search [15], and by searching against genomic databases such as translated EST databases [16, 17]. However, these methods are not commonly applied as primary peptide identification methods because they are more time consuming and error prone than conventional database searching.

This study extends our previous analysis of the unassigned high-quality spectra [3]. We implement and apply a more comprehensive and efficient computational strategy based on iterative analysis of MS/MS data using a combination of several existing computational tools. The analysis, outlined in Fig. 1, involves the use of spectral library searching [18, 19], blind searching for PTM analysis, and genomic database searching. The iterative nature in which these different approaches are applied to data allows increasing the number of assigned MS/MS spectra without a substantial increase in the computational time.

Correspondence: Dr. Alexey Nesvizhskii, Department of Pathology, University of Michigan, 4237 Medical Science I, Ann Arbor, MI 48109, USA
E-mail: nesvi@umich.edu
Fax:+1-734-936-7361

Abbreviations: FDR, false discovery rate; **NCBI NR,** NCBI non-redundant; **SNP,** single nucleotide polymorphism; **SQS,** spectral quality score; **WCL,** whole-cell lysates

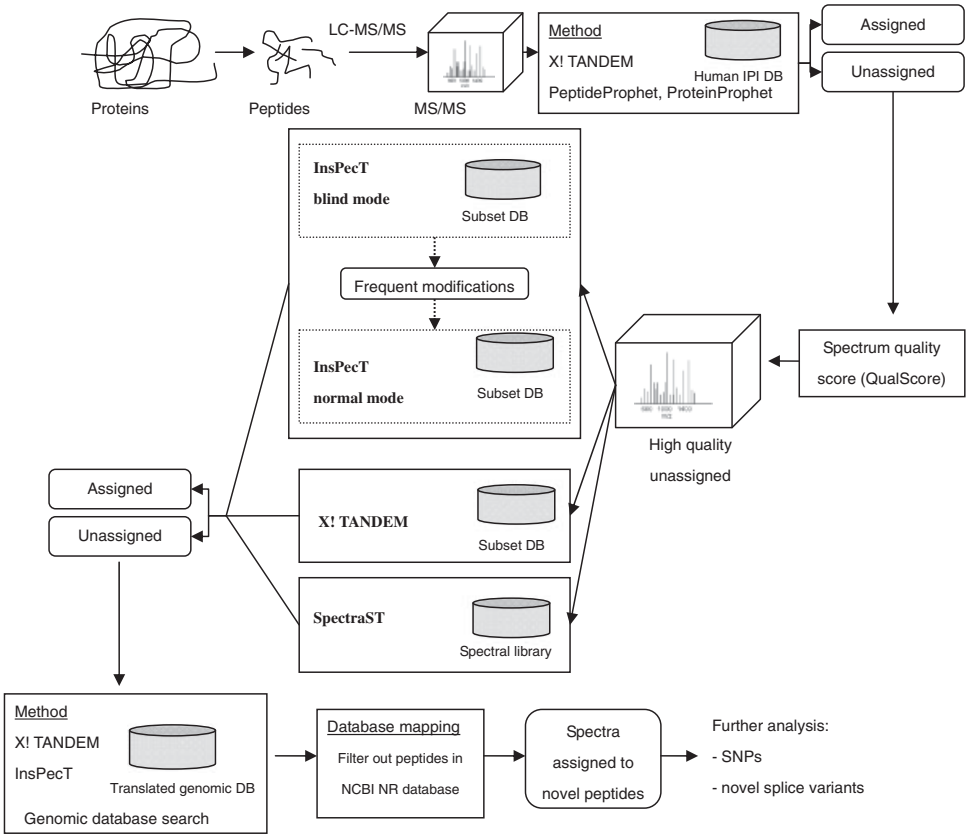

Figure 1. Overview of the iterative peptide identification strategy. Proteins are digested into peptides, and peptides are sequenced using MS/MS. Acquired spectra are analyzed using conventional database searching. Peptide identifications are processed using PeptideProphet and ProteinProphet. A spectral quality assessment tool is used to select unassigned high-quality spectra. These spectra are reanalyzed using X! TANDEM and InsPecT (normal and blind mode) against the subset protein database, and using SpectraST spectral library search tool. The remaining unassigned spectra are searched against the translated genomic database to identify novel peptides and peptide polymorphisms.

The method is applied to a large publicly available data set of MS/MS spectra from the Human T leukemic cells [20]. Briefly, the whole-cell lysates (WCL) were separated by one-dimensional gel electrophoresis and the gel lanes were cut into 18 gel slices. The proteins contained in the gel slices were digested with trypsin, and the peptides were extracted and analyzed by LC-MS/MS using an LTQ ion trap mass spectrometer. The data set contains 14 replicate analyses of the WCL, which was used here as the primary data set. Additional analysis was performed using two subcellular fractions: the plasma membrane and the lipid raft.

The protein sequence databases used in this work included the Human International Protein Index (IPI) database v3.32 containing 67 575 entries [21], the NCBI non-redundant (NR) Human database containing 383 745 entries (downloaded on 02/15/2008), and the translated genomic database, compiled from multiple sources and compressed for computational efficiency as described in [17] (downloaded on 01/02/2008 from ftp://ftp.umiacs.umd.edu/pub/nedwards/PepSeqDB).

In the initial analysis, the spectra were searched with X! TANDEM/k-score [22] against the Human IPI database described above appended with an equal number of reversed protein sequences as decoys [21]. The search parameters were as follows: parent ion mass tolerance window of −2.0 to 2.0 Da, 0.8 Da monoisotopic fragment ion mass tolerance, tryptic peptides only. Two variable modifications were considered: methionine oxidation and N-terminal acetylation. The refinement mode was not used. PeptideProphet [23] was then used to calculate the probability for each of the spectrum assignments. The spectra with QualScore [3] spectral quality score (SQS) above 1.0 were considered high-quality spectra. The spectra with PeptideProphet probability below 0.1 and SQS score above 1.0 were considered unassigned high-quality spectra. These spectra, representing approximately 10% of the full MS/MS data set, were the main focus of this work. We also note that among the unassigned spectra of lower quality, which were not further interrogated here, many are likely to represent valid peptides. Peptides that fall into the non-mobile proton model category, or contain extra liable bonds, are known to fragment poorly in conventional MS strategies [24], and their analysis requires the use of more sophisticated peptide fragmentation models [25, 26] than what is implemented in most currently available database search tools.

Unassigned high-quality spectra were reanalyzed using several additional steps: X! TANDEM database searching against the subset database containing sequences of proteins identified with high ProteinProphet probabilities (greater than or equal to 0.9) [27] in the initial search (to identify additional tryptic peptides by searching against a smaller database compared to the original search, as well as semi-tryptic peptides, and peptides with inaccurately measured precursor ion mass [3], see step i below);

blind InsPecT searching (extensive PTM analysis to identify the most common modifications, step ii); normal InsPecT with an extended set of modifications (for better identification of most frequent modifications discovered using the blind search, step iii); spectral library searching using SpectraST [18] (more sensitive scoring method, as compared to conventional database searching, for assigning MS/MS spectra produced by previously identified peptides [28, 29], step iv); genomic database searching (for identification of peptides not present in the protein sequence database used in the initial search and in steps i–iv, see step v). These steps are described below in more detail (see Fig. 1):

(i) X! TANDEM search (without refinement), subset database, larger (than initial search) parent ion mass tolerance of 4.0 Da, allowing semi-tryptic peptides. The same modifications were considered as in the initial search, *i.e.* methionine oxidation and N-terminal acetylation.

(ii) InsPecT blind mode search, subset database, 2.5 Da parent ion mass tolerance, and allowing tryptic peptides only. In the blind mode, InsPecT attempts to identify peptides with unexpected PTMs or chemical modifications by allowing unrestricted (any mass shift) modification of any one residue in the peptide sequence.

(iii) InsPecT normal mode, subset database, 2.5 Da parent ion mass tolerance, and allowing semi-tryptic peptides. The most frequent modifications based on the InsPecT blind mode analysis (step ii) were specified as variable modifications (see below).

(iv) SpectraST search with default settings against a spectral library generated by combining the NIST Human MS/MS library (v. 2006-12-13) and the experiment-specific library generated from the spectra identified in the initial X! TANDEM search [30].

(v) The high-quality spectra that remained unassigned after steps i–iv above were searched against the translated Human genomic database with X! TANDEM (3.0 Da parent ion mass tolerance, tryptic peptide only, without refinement) and InsPecT (2.5 Da parent ion mass tolerance, tryptic peptides only) for novel peptide identifications. Only methionine oxidation was allowed as a variable modification in both searches.

To estimate the false discovery rate (FDR) [1] at each step in the process, an equal number of decoy protein sequences (reversed sequences) were appended to the searched database. In the case of the translated genomic database search, due to its large size, the number of appended decoy sequences was a fourth of the database size. A non-parametric probability mixture model [31] was applied to X! TANDEM and InsPecT genomic database search results. This model does not require that the target and decoy database to be of equal size. The filtering thresholds were selected to achieve the FDR of less than 0.05. It should be noted that the validity of the FDR estimates in the case of iterative database searches has yet to be carefully investigated. However, because the focus of this study is on exploring general trends and understanding the sources of unassigned high-quality spectra, the results presented below should not be significantly affected by the details of FDR analysis and data filtering performed at each step.

The distributions of SQS for all of the spectra, and separately for assigned and unassigned spectra (after the initial search), were analyzed (Fig. 2A). The quality score of 1.0 was found to separate assigned and unassigned spectra fairly well. Of all spectra in the data set, over 65% were

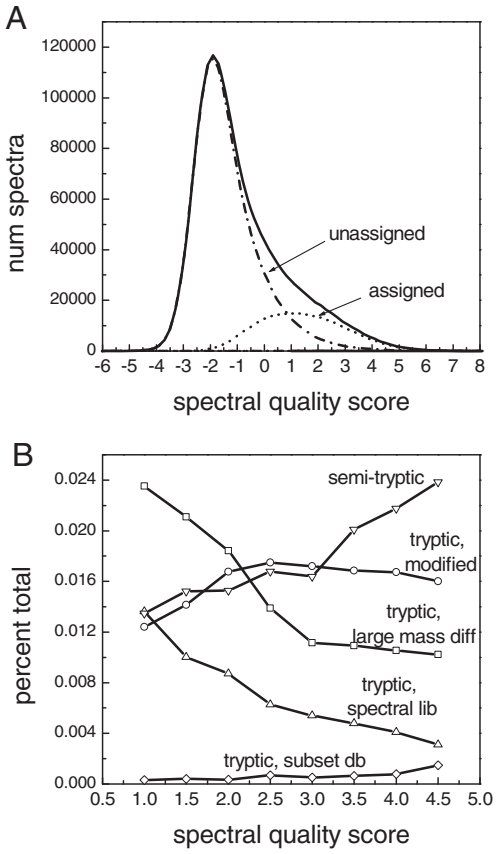

Figure 2. Prevalence and categories of unassigned high-quality spectra. (A) The distribution of spectral quality scores plotted for all spectra (solid line), and separately for unassigned (dash dot line) and assigned (short dash) spectra after the initial database search. (B) The ratio of spectra assigned to peptides of different types (''percent total'' refers to the proportion of spectra assigned to peptides of different type among the total number of initially unassigned spectra) during reanalysis, plotted as a function of the spectral quality score. The category ''tryptic, subset db'' refers to spectra corresponding to unmodified tryptic peptides that were identified due to reduced search space. The category ''tryptic, spectral lib'' refers to spectra corresponding to unmodified tryptic peptides identified using spectral library searching, and it includes some spectra that were also identified by other methods. WCL fraction data.

unassigned low-quality spectra (SQS < 1, probability below 0.1), and close to 10% of spectra were of high quality (SQS > 1) and unassigned. A significant proportion of these unassigned high-quality spectra (more than 30%) could be assigned at one of the reanalysis steps described above. Figure 2B shows the distribution of spectra as a function of SQS identified at steps i–iv of the reanalysis pipeline and grouped into different categories. Several trends are apparent. The ratio of spectrum assignments (fraction of assigned spectra among all initially unassigned spectra) obtained *via* spectrum library searching decreases with increasing SQS scores. This indicates that spectral library searching is most advantageous (for gaining additional identifications) when applied to spectra of lower quality. Thus, in those applications where the primary goal is to increase the number of spectral assignments, spectral library searching should be applied on the entire data set, *i.e.* without spectral quality filtering. It was also found that a significant number of spectra were due to semi-tryptic peptides (more than 5% of the total number of assigned spectra, including peptides identified in the initial search). The other two main categories were modified tryptic peptides and tryptic peptides with incorrectly measured parent ion m/z value. A small number of tryptic peptides were identified due to reduced database size (subset database). These peptides were masked in the original search by other peptides (''distraction effect'' [32]).

The initial search included two modifications only: methionine oxidation and N-terminal acetylation (or carbamylation, as it cannot be distinguished from acetylation in low mass accuracy data sets). A much larger space of PTMs and chemical modifications was explored using the blind mode of InsPecT, which allows any mass shift on any one residue in the peptide sequence. The blind search revealed a large number of frequent modifications (the most frequent ones are listed in Fig. 3). However, we also found that while the blind mode of InsPecT was successful at identifying the most frequent types of modifications in this data set, it was not as sensitive at detecting any particular type of modification as the normal mode InsPecT with that modification explicitly specified in the input file as a variable modification. Furthermore, blind mode InsPecT had a difficulty with localizing the site of the modification (*e.g.* in the case of phosphorylation, in some instances the +80 Da shift was placed on a residue other than S, T, or Y). Due to these concerns, and acknowledging the difficulty of accurate FDR control in the case of blind searches, we have not counted spectral assignments identified by the blind InsPecT search only. Instead, the blind mode was used for identifying the most frequent modifications in the data set, which was followed by the normal InsPecT search allowing these most frequent modifications only (see Fig. 1).

A more detailed analysis of modified peptides revealed several interesting trends. The higher the protein abundance (measured using spectral counts [1]), the more likely it was to observe a modified peptide from that protein

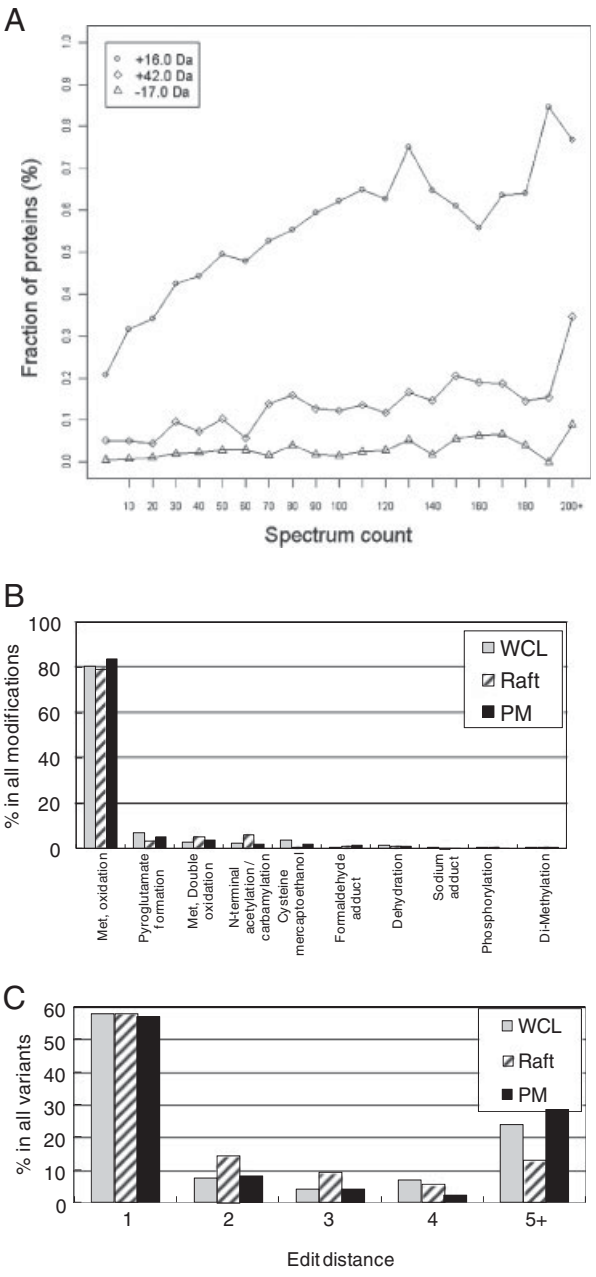

Figure 3. Additional analysis of peptide categories. (A) The ratio of proteins (among proteins of similar abundance as measured using spectral counts) containing at least one modified peptide of a particular type (WCL fraction data). Shown are methionine oxidation (+16), N-terminal acetylation/carbamylation (+42), and pyroglutamic acid formation from N-terminal glutamic acid (−17.0). (B) Most frequent modifications and their normalized frequencies in WCL, plasma membrane (PM), and raft fractions. (C) Novel peptides (according to NCBI NR database) identified by the genomic database search and categorized by edit distance (WCL, plasma membrane, and raft fractions).

(Fig. 3A), in agreement with previous observations [3, 14]. The rate of modifications across different samples was investigated as well. Figure 3B shows the distribution

of the most frequent modifications in WCL, as well as in plasma membrane and raft fractions. While the overall trend is the same (*e.g.* oxidation was the most common modification), there are noticeable differences, likely reflecting variations in sample handling. It is also apparent that the dominant majority of identified modifications are chemical modifications likely due to sample handling, and not biologically relevant modifications.

Many high-quality spectra (more than 50% of all initially unassigned high-quality spectra) remained unassigned after all reanalysis steps involving searches against protein sequence databases. A small fraction (<10%) of these still unassigned spectra were identified by performing X! TANDEM and InsPecT searches against the translated genomic database, followed by a non-parametric target-decoy based FDR control [31] to achieve less than 5% FDR. The peptides found by genomic database searching were mapped to protein sequences in the NCBI NR database. For each peptide, the edit distance was computed between the peptide and its closest match (smallest edit distance) in the NCBI NR databases. Edit distance is defined here as the number of amino acid differences between the two peptides. Peptides with non-zero edit distance were referred to as ''novel'' peptides. Figure 3C shows the distribution of novel peptides in terms of the edit distance. While many of the novel peptides differed from the best matching NCBI NR peptide by edit distance of 1 (putative single nucleotide polymorphisms, SNPs), a substantial proportion had high edit distances (putative novel splice variants). A more detailed analysis was then performed by searching the sequences of novel peptides against the human dbSNP database [33]. Results show that only about 5% of peptide polymorphisms found in this data set corresponded to known SNPs in dbSNP. Furthermore, a number of possible alternative splice variants were discovered from alignment of novel peptides against the gene models in UCSC genome browser by BLAT [34]. As a part of this technical report, however, no attempt was made to further validate any of the specific identifications.

Efficiency of the computational analysis is an important practical consideration. All work was done on a Linux server with a 2.2 GHz CPU and 16 GB of memory. An average mzXML file took 1 h (*per* CPU) for the initial search using X! TANDEM. The blind InsPecT search was performed against the subset database (a fraction of the original database size), and the searches against the genomic database were manageable due to the database compression [17]. Limiting the reanalysis to unassigned high-quality spectra only was also important since these spectra represented a small fraction of the original data set. The time to build and search the spectral library was not significant compared to various sequence database search steps. Overall, the reanalysis of spectra took less than 1 h (*per* CPU/mzXML file) for X! TANDEM, InsPecT (normal mode) and SpectraST searches combined, less than 2 h for InsPecT blind mode, and 2–3 h for EST database search. The movement of data and integration of different search results was carried out using several in-house developed programs.

The iterative database search strategy described here is flexible, and different combinations of methods for reanalysis of unassigned high-quality spectra can be applied. It is worth noting, however, that despite all efforts, a substantial fraction of the high-quality spectra in the data set used in this work remain unassigned. The success rate of peptide identification can be further improved by using a combination of different search engines (in addition to X! TANDEM and InsPecT used here) [35, 36], as well as by implementing more accurate peptide fragmentation models [37]. It has been suggested that a significant number of unidentified spectra are chimera spectra resulting from co-fragmentation of two or more different peptides [25, 32, 38]. Additional work is necessary to develop methods to analyze MS/MS data allowing for the possibility of chimera spectra, as well as to get a better understanding of the practical importance of identifying such spectra for increasing the total number of identified peptides and proteins. Other computational strategies not relying on database searching may also be required for further improving the sensitivity of peptide identifications, *e.g. de novo* sequencing [39, 40]. It should also be noted that given continuous improvements in MS instrumentation, the strategy presented here will need to be revised in the future. For example, in the case of MS/MS spectra generated on high mass accuracy instruments, substantial improvements can be achieved *via* more accurate determination of the precursor ion charge state and m/z [41].

The iterative approach utilized in this work could be of general interest beyond the primary focus of this technical brief on understanding the sources of unassigned high-quality spectra. First, it can be used to more effectively search for novel peptides (SNPs, novel splice variants) as way to improve genome annotation [42, 43]. Second, the method can assist in obtaining a more complete picture of how the rates of various modifications (post-translational and chemical), as well as numbers of semi-tryptic peptides and peptides with missed cleavages, vary from sample to sample and change as a function of experimental or sample handling conditions. Such an analysis is particularly important in the context of targeted proteomic studies using multiple reaction monitoring assays, where accurate peptide quantification requires normalization to account for peptide modifications and changes in the efficiency of trypsin digestion [44]. Finally, one may envision that iterative/multistep data analysis strategies will play a more prominent role in future proteomic studies. We note, however, that routine application of iterative strategies such as the one utilized in this work, especially in a high-throughput environment, will require further substantial work on the development of statistical FDR estimation methods applicable to a wide range of peptide identification approaches, including subset database searching, blind PTM analysis, and genomic searches.

This work was supported in part by NIH/NCL grant R01 CA-126239 and NIH/NCRR grant P41-18627. The authors thank Hyungwon Choi and Xia Cao for helpful discussions.

The authors have declared no conflict of interest.

References

[1] Nesvizhskii, A. I., Vitek, O., Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 2007, *4*, 787–797.

[2] Hernandez, P., Muller, M., Appel, R. D., Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrom. Rev.* 2006, *25*, 235–254.

[3] Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M. *et al.*, Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* 2006, *5*, 652–670.

[4] Flikka, K., Martens, L., Vandekerckhoe, J., Gevaert, K., Eidhammer, I., Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* 2006, *6*, 2086–2094.

[5] Moore, R. E., Young, M. K., Lee, T. D., Method for screening peptide fragment ion mass spectra prior to database searching. *J. Am. Soc. Mass Spectrom.* 2000, *11*, 422–426.

[6] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[7] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[8] Eng, J. K., McCormack, A. L., John, R., Yates, I., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[9] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[10] Mann, M., Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 1994, *66*, 4390–4399.

[11] Tanner, S., Shu, H., Frank, A., Wang, L. C. *et al.*, InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* 2005, *77*, 4626–4639.

[12] Han, Y., Ma, B., Zhang, K., SPIDER: software for protein identification from sequence tags with *de novo* sequencing error. *J Bioinform. Comput. Biol.* 2005, *3*, 697–716.

[13] Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C. *et al.*, Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer - II. New developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell. Proteomics* 2005, *4*, 1194–1204.

[14] Nielsen, M. L., Savitski, M. M., Zubarev, R. A., Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol. Cell. Proteomics* 2006, *5*, 2384–2391.

[15] Tsur, D., Tanner, S., Zandi, E., Bafna, V., Pevzner, P. A., Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* 2005, *23*, 1562–1567.

[16] Choudhary, J. S., Blackstock, W. P., Creasy, D. M., Cottrell, J. S., Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* 2001, *1*, 651–667.

[17] Edwards, N. J., Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.* 2007, *3*, 102.

[18] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. *et al.*, Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007, *7*, 655–667.

[19] Craig, R., Cortens, J. C., Fenyo, D., Beavis, R. C., Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* 2006, *5*, 1843–1849.

[20] Wu, L., Hwang, S. I., Rezaul, K., Lu, L. J. *et al.*, Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling. *Mol. Cell. Proteomics* 2007, *6*, 1343–1353.

[21] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 2004, *4*, 1985–1988.

[22] MacLean, B., Eng, J. K., Beavis, R. C., McIntosh, M., General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* 2006, *22*, 2830–2832.

[23] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383–5392.

[24] Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S. *et al.*, Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* 2003, *75*, 6251–6264.

[25] Sun, S. J., Meyer-Arendt, K., Eichelberger, B., Brown, R. *et al.*, Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Mol. Cell. Proteomics* 2007, *6*, 1–17.

[26] Zhang, Z. Q., Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* 2004, *76*, 3908–3922.

[27] Kolker, E., Purvine, S., Galperin, M. Y., Stolyar, S. *et al.*, Initial proteome analysis of model microorganism *Haemophilus influenzae* strain Rd KW20. *J. Bacteriol.* 2003, *185*, 4593–4602.

[28] Craig, R., Cortens, J. C., Fenyo, D., Beavis, R. C., Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* 2006, *5*, 1843–1849.

[29] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. *et al.*, Development and validation of a spectral library searching

method for peptide identification from MS/MS. *Proteomics* 2007, *7*, 655–667.

[30] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. *et al.*, Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* 2008, *5*, 873–875.

[31] Choi, H., Ghosh, D., Nesvizhskii, A. I., Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* 2008, *7*, 286–292.

[32] Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D. *et al.*, Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* 2004, *76*, 3556–3568.

[33] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J. *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001, *29*, 308–311.

[34] Kent, W. J., BLAT--the BLAST-like alignment tool. *Genome Res.* 2002, *12*, 656–664.

[35] Searle, B. C., Turner, M., Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* 2008, *7*, 245–253.

[36] Edwards, N., Wu, X., Tseng, C. W., An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. *Clin. Proteomics* 2009, *5*, 23–36.

[37] Yen, C. Y., Meyer-Arendt, K., Eichelberger, B., Sun, S. J. *et al.*, A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. *Mol. Cell. Proteomics* 2009, *8*, 857–869.

[38] Zhang, N., Li, X. J., Ye, M. L., Pan, S. *et al.*, ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* 2005, *5*, 4096–4106.

[39] Frank, A., Pevzner, P., PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, *77*, 964–973.

[40] Ma, B., Zhang, K., Hendrie, C., Liang, C. *et al.*, PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2337–2342.

[41] Mayampurath, A. M., Jaitly, N., Purvine, S. O., Monroe, M. E. *et al.*, DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics* 2008, *24*, 1021–1023.

[42] Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R. *et al.*, Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* 2006, *7*.

[43] Tanner, S., Shen, Z. X., Ng, J., Florea, L. *et al.*, Improving gene annotation using peptide mass spectrometry. *Genome Res.* 2007, *17*, 231–239.

[44] Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J. *et al.*, Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* 2009, *7*, 633–641.

