Chapter 4

Analysis of Mass Spectrometry Data in Proteomics

Rune Matthiesen and Ole N. Jensen

Abstract

The systematic study of proteins and protein networks, that is, proteomics, calls for qualitative and quantitative analysis of proteins and peptides. Mass spectrometry (MS) is a key analytical technology in current proteomics and modern mass spectrometers generate large amounts of high-quality data that in turn allow protein identification, annotation of secondary modifications, and determination of the absolute or relative abundance of individual proteins. Advances in mass spectrometry–driven proteomics rely on robust bioinformatics tools that enable large-scale data analysis. This chapter describes some of the basic concepts and current approaches to the analysis of MS and MS/MS data in proteomics.

Key words: Database searching, *de novo* sequencing, peptide mass fingerprinting, peptide fragmentation fingerprinting, quantitation.

1. Introduction

1.1. Protein Identification, Annotation of PTMs, and Quantitation

DNA sequencing and microarray technology have provided large-scale, high throughput methods to quantitate cellular mRNA levels and determine distributions of single nucleotide polymorphisms (SNPs) at a genome-wide scale. However, DNA-based technologies provide little, if any, information about dynamic biomolecular events, such as protein interactions, protein-based regulatory networks, and post-translational modifications of proteins. For example, regulatory events governed at the level of mRNA translation have been reported *(1)*, and the activity of proteins is often controlled by post-translational modifications *(2, 3)*. Proteomics, the systematic study of proteins, grew out of protein chemistry during the 1980s and 1990s. Initially based mainly on two-dimensional gel electrophoresis, proteomics has now embraced a range of

Jonathan M. Keith (ed.), *Bioinformatics, Volume II: Structure, Function and Applications, vol. 453*

Book doi: 10.1007/978-1-60327-429-6 Springerprotocols.com

biochemical, immunological, and computational fields as well as a series of sensitive analytical technologies, including mass spectrometry (3–5). Currently, most large-scale protein identification work in proteomics is based on mass spectrometry. The mass spectrometer is used to determine the accurate molecular mass of proteins and the derived peptides, and tandem mass spectrometry (MS/MS) facilitates amino acid sequencing and mapping of post-translational modifications *(2, 4)*. Modern mass spectrometers generate a wealth of proteomics data in a short time and the computational processing of this data remains a bottleneck in many proteomics projects *(5, 6)*. The aim of this chapter is to introduce the basic concepts for analysis and processing of mass spectrometry data obtained in proteomics experiments.

1.2. Mass Spectrometry and Proteomics Workflows

Protein analysis by mass spectrometry is typically performed in a "bottom-up" fashion, meaning that proteins are digested into peptides, which are in turn analyzed by mass spectrometry. The protein sequence and secondary modifications are subsequently assembled based on peptide MS and MS/MS data *(8)*. Recently, the concept of "top-down" analysis of intact proteins was introduced *(9)*, but it is beyond the scope of this chapter to describe this approach (*see* **Note 1**). Peptides are normally generated by trypsin cleavage of protein. Trypsin specifically and efficiently cleaves the amide bond C-terminal to arginine and lysine residues, unless a proline is the next residue (*see* **Notes 2** and **3**). The peptide fragments are then analyzed using mass spectrometry–based strategies, including peptide mass mapping by MS and peptide sequencing by MS/MS.

Peptide mass mapping strategies are usually used for characterization of simple protein samples containing only one or a few protein species and it is often used in combination with protein separation by 2D gel electrophoresis (**Fig. 4.1**). In the peptide mass mapping approach, the molecular masses of a set of tryptic peptides derived from a protein sample is measured. Since each individual protein has a distinct amino acid sequence, the pattern of tryptic peptides for each protein will be unique and it can be used to identify the protein by database searching combined with scoring algorithms to retrieve the best matches (*see* **Note 4** and **Fig. 4.2**). Protein identification by peptide mass mapping relies on accurate molecular weight determination by mass spectrometry and the assumption that trypsin faithfully cleaves at Arg and Lys residues *(10)*. More detailed analysis of individual peptides is achieved by MS/MS, which allows amino acid sequencing of selected peptides (*see* **Fig. 4.1**). Peptide separation by liquid chromatography (LC) is advantageous when analyzing complex peptide samples as LC equipment is readily interfaced to electrospray ionization tandem mass spectrometers, so called LC-MS/MS systems.

LC-MS/MS strategies are used for detailed analysis of complex protein mixtures and for mapping of post-translational modifications. First, the MS/MS instrument records a mass spectrum

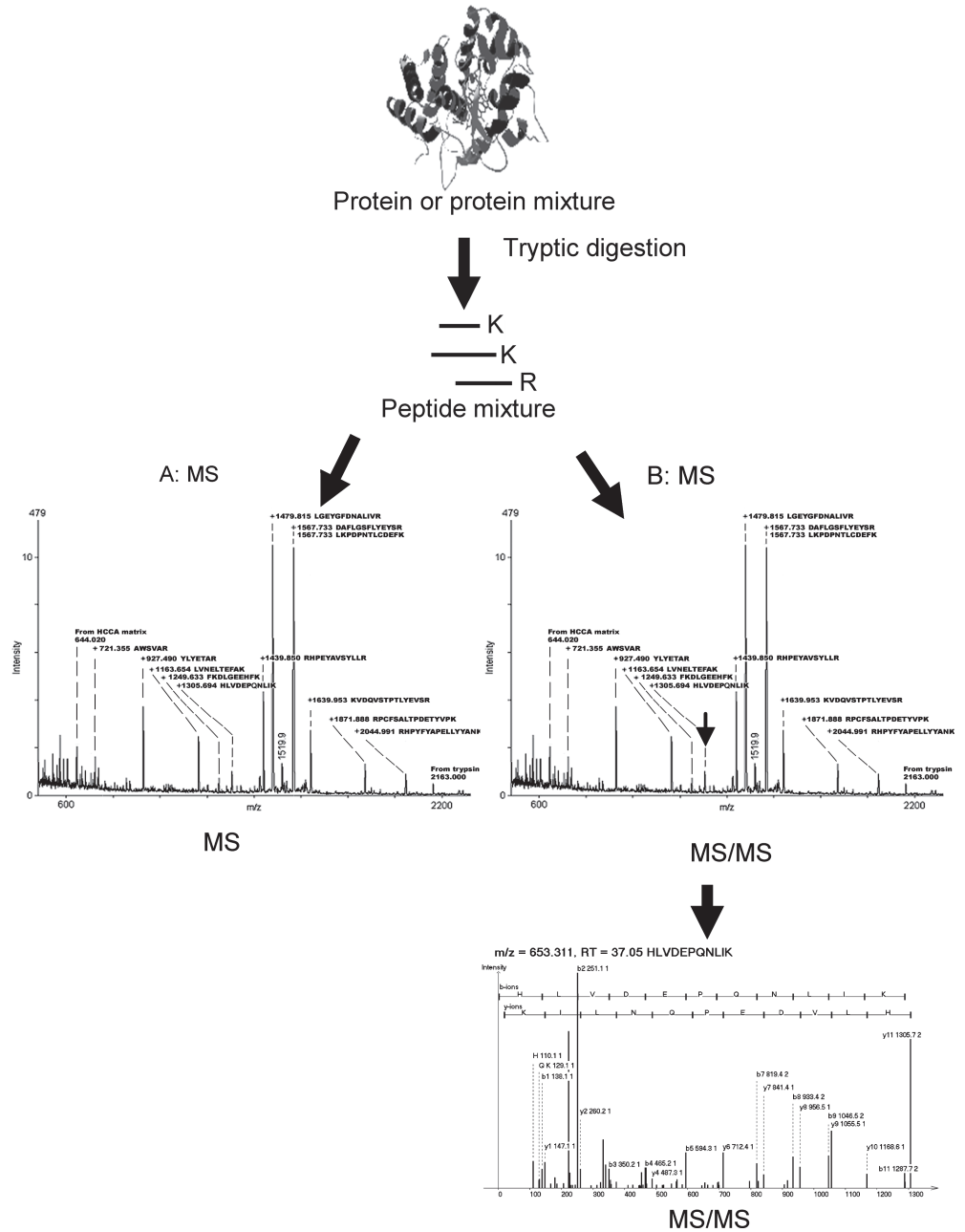

Fig. 4.1. The workflow for the MS based peptide mass mapping (**A**) and MS/MS-based peptide sequencing (**B**) methods. The latter method goes one step further by recording the MS/MS spectrum/spectra of chosen peptides.

to determine the masses of all peptides that are eluting from the LC-column at a given time. Next, the mass spectrometer control software determines the masses (m/z values) of the two to five most intense peptide signals for further analysis by MS/MS, i.e., for sequencing (*see* **Note 5**). Each of these gas-phase peptide ions is, in turn, isolated and fragmented inside the mass spectrometer.

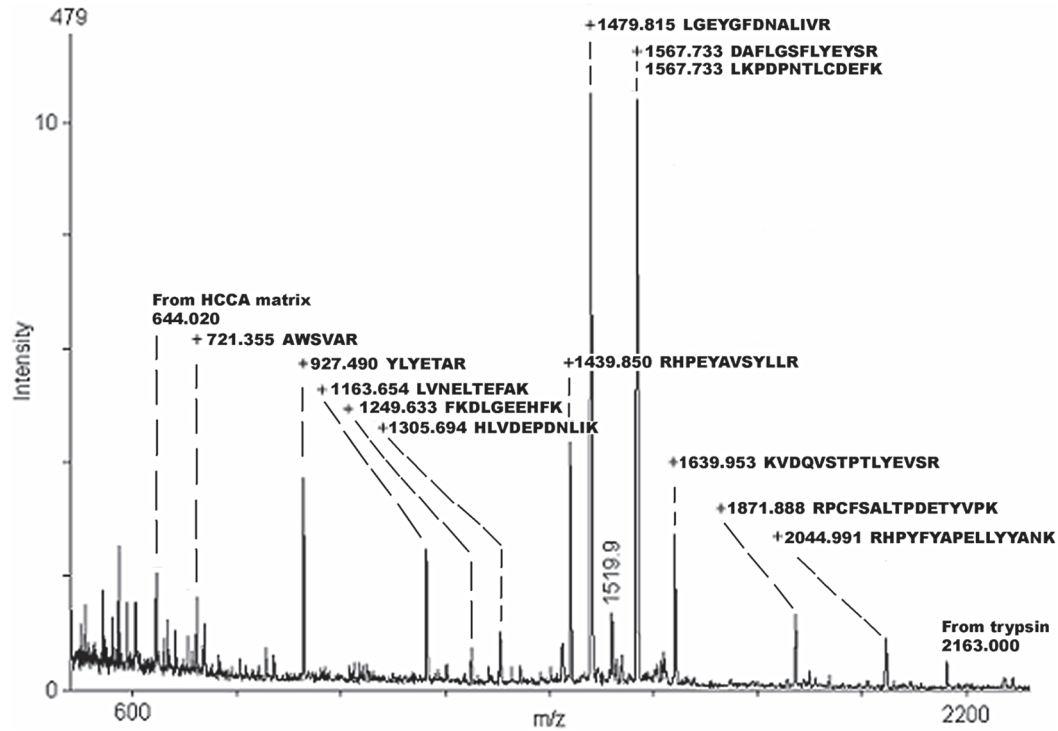

Fig. 4.2. MALDI-TOF MS spectrum (PMF/PMM strategy). (+) indicates mass peaks that matched the masses of theoretical tryptic peptides from BSA.

The set of peptide fragments is then recorded by a second mass analyzer to generate the MS/MS spectrum for that peptide. Notably, the most frequent type of fragmentation occurs by cleavage of the amide bond between amino acid residues to generate y-ions containing the C-terminal parts and b-ions containing the N-terminal parts of the peptide (**Fig. 4.3A**). The mass differences between the peptide fragment ion signals in the MS/MS spectrum can be correlated with amino acid residue masses (*see* **Fig. 4.3B** and **Table 4.1**): In this way amino acid sequence information can be obtained by tandem mass spectrometry (*8, 11*). The precise fragmentation chemistry depends on the fragmentation method and mass analyzer used, and has been reviewed elsewhere *(12, 13)*.

The analysis of peptide mass spectra requires several steps that are described in the following (**Fig. 4.4**): *(1)* conversion of the continuous mass spectral data to a list of peptide masses (MS) or peptide fragment masses (MS/MS) (**Fig. 4.5**); *(2)* spectral interpretation and sequence annotation to generate a list of peptides and proteins; *(3)* quantification of peptides and proteins and comparative analysis; *(4)* bibliographic and computational sequence analysis; *(5)* data storage in a relational database.

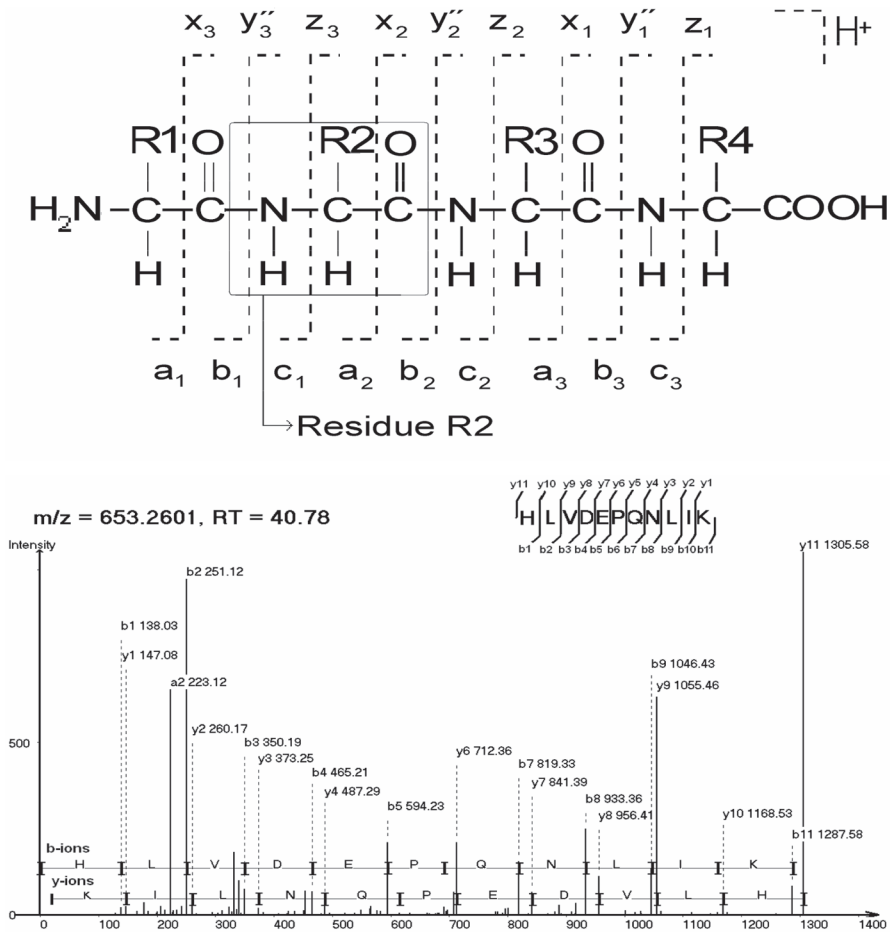

Fig. 4.3. (**A**) The nomenclature used for observed MS/MS fragments of peptides. (**B**) MS/MS spectrum obtained from the double charged peptide HLVDEPQNLIK. For simplicity only the intense y-, b-, and a-ions are annotated. Notice that the mass difference between the peaks from ions in the b-ion series gives the amino acid sequence directly. The mass difference between the peaks from ions in the y-ion series gives the reverse amino acid sequence.

2. Data Analysis

2.1. Data Preprocessing: The Need for Standardization of Data Formats

Proteomics is still a rather young scientific field, and it is only recently that proposals for common data standards and data exchange formats have emerged: mzXML *(14)* and mzDATA *(15)*. There is now an ongoing effort to merge these two standard formats *(16)*.

The mzXML and mzDATA converters either keep the continuous data obtained from the mass spectrometers or convert them into a peak list by using the centroid method (*see* **Note 6**). These converters do not use any advanced spectra processing techniques. More advanced spectra processing algorithms can reduce the "background signals" in MS and MS/MS data

Table 4.1
Delta masses that can be observed between peaks in a MS/MS spectrum and their corresponding interpretations

Identity	Delta mass (Da)
G	57.0215
A	71.0371
S	87.0320
P	97.0528
V	99.0684
T	101.0477
L	113.0841
I	113.0841
N	114.0429
D	115.0269
Q	128.0586
K	128.0950
E	129.0426
M	131.0405
H	137.0589
F	147.0684
R	156.1011
Y	163.0633
W	186.0793
C unmodified	103.0092
C carbamidomethylated	160.0307
M oxidation	147.0354

(*see* **Note 7**) (*see* **Fig. 4.5A,B**), isotope-deconvolute the peptide ion signals, and charge-deconvolute the protonated peptide ion signals (*see* **Fig. 4.5**). The final result from the advanced processing algorithms is a "peak list" of peptide masses (MS) or peptide fragment masses (MS/MS) that correspond to singly charged

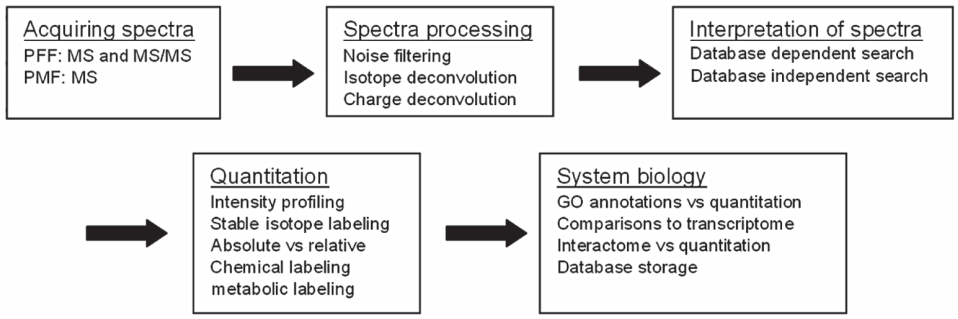

Fig. 4.4. Steps involved in interpretation of mass spectrometry data.

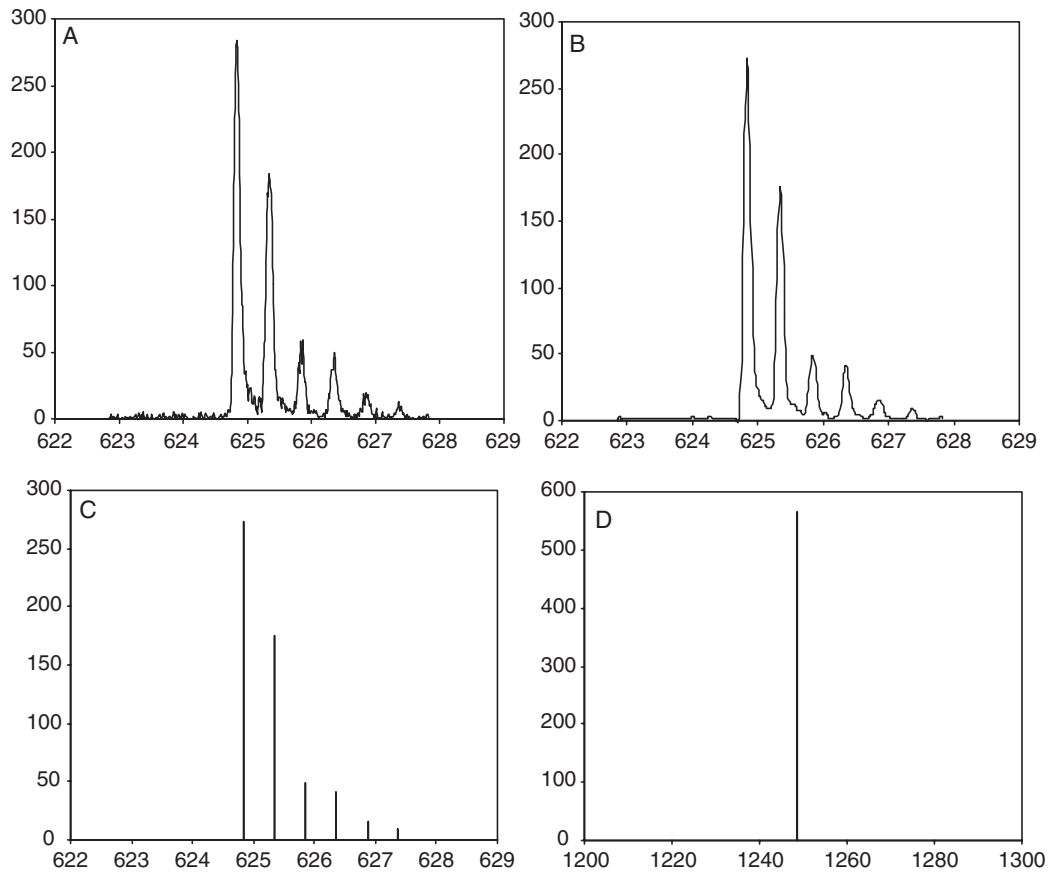

Fig. 4.5. The result of noise filtering, centroiding, charge, and isotope deconvolution. Mass over charge is plotted on the x-axis and intensity on the y-axis. (**A**) The raw data MS spectrum of the double charged peptide LGGQTYNVALGR. (**B**) The data after three iterations of a nine-point Savitsky-Golay filtering algorithm. (**C**) The peak list obtained by centroiding the peaks in (**B**). (**D**) Charge and isotope deconvolution of the peak list in (**C**).

monoisotopic peaks. Such algorithms and data reduction make the subsequent data analysis much simpler. To our knowledge, there is no publicly available application that can process raw data from any of the standard MS data formats into singly charged

monoisotopic peak lists. However, this can be done by the MaxEnt algorithm, which is available in the commercial programs MassLynx v4.0 and PLGS v2.05 from Waters.

In the VEMS *(17)* platform, the program ExRaw interfaces to the mzXML converters *(14)* and the Maxent algorithm from PLGS v2.05 for convenient automated data processing of several LC-MS/MS runs. The spectra mass and intensity values are stored by Base64 encoding in both the mzXML and mzDATA formats.

2.2. Database Searching

Peak lists of peptide masses (MS) and/or peptide fragment masses (MS/MS) are searched against protein sequence databases *(18)*. It is also possible to search DNA and EST databases *(19, 20)*, if needed. The techniques of comparing mass spectrometry data against a sequence database are also called database-dependent search algorithms *(21)* since these algorithms are dependent on a sequence database for interpreting the data. A large number of algorithms and programs have been proposed; many of them are publicly available and others are commercial (**Table 4.2**). It is often useful to compare search results from different search algorithms since current software, e.g., VEMS, Mascot *(22)*, and X!Tandem *(23)* generate slightly different results.

Table 4.2
Useful programs for interpreting MS and MS/MS spectra of peptides

Name	Input Data	Interfaced from VEMS	Public
VEMS v3.0	MS, MS/MS	No	Yes
Mascot	MS, MS/MS	Yes	Semi
X!Tandem	MS/MS	Yes	Yes
P3	MS/MS	No	Yes
Inspect	MS/MS	No	Yes
Phenyx	MS/MS	No	Semi
PepNovo	MS/MS	No	Yes
Lutefisk	MS/MS	Yes	Yes
OpenSea	MS/MS	No	Yes
De Novo peaks	MS/MS	No	Yes
PepHMM	MS/MS	No	Yes
ProteinProphet	MS/MS	No	Yes

The user-defined search parameter settings of the different search engines are rather similar and are briefly discussed in the following. The preprocessing of raw data into peak lists, which can be done with the mass spectrometry instrument vendor software, is necessary before MS and MS/MS data can be submitted to database search engines (*see* **Section 2.1**).

1. Database. The FASTA format used for the sequence databases can be obtained from resources such as NCBI *(24)*, EBI *(25)*, or Swiss-Prot *(26)*. The choice of database is an important issue. It is essential to understand that the search engines work by finding the optimum match in the database. This means that if an incomplete database is searched then some spectra might be matched to a wrong peptide sequence. Another issue is that data sets often have cross-contamination from organisms other than the one used in the study (e.g., keratin from *Homo sapiens* and peptides from porcine trypsin may contaminate the search results obtained from an *A. thaliana* [plant] protein sample). The generally accepted procedure in the mass spectrometry field is to search the international protein indexed database IPI and then include the most likely contaminants from other organisms. The IPI database is the top-level database for a number of other databases (**Fig. 4.6**). IPI is a non-redundant database composed of sequences from Uniprot knowledgebase, Ensembl, and RefSeq. Ensembl and RefSeq are composed of translated predicted genes. Uniprot knowledgebase is composed of sequences from Swiss-Prot, PIR (protein information resource), and TrEMBL. Swiss-Prot and PIR are manually curated databases with many cross-references to other resources, including literature references. TrEMBL contains translated nucleotide sequences from EMBL. A great effort has been made to make IPI non-redundant without removing isoforms. Uniprot knowledge base is further divided into UniREF 100, 90, and 50. In the UniREF databases, all sequences with a length longer than 11 amino acids and sequence identity of 100, 90 or 50 percent are represented as one sequence entry.

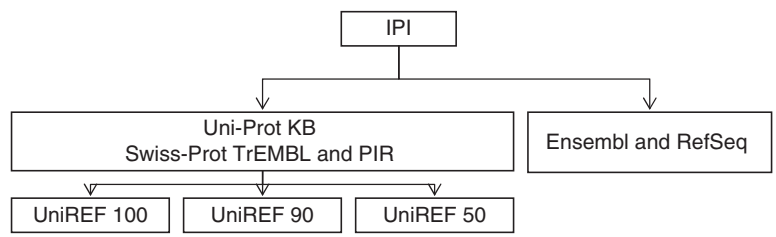

Fig. 4.6. The organization of a selected set of databases frequently used for searching mass spectrometry data.

2. Cleavage method. There are both enzymatic and chemical cleavage methods that can be used to generate peptide fragments. However, trypsin digestion is used most often (*see* **Notes 2** and **3**). Some database search programs allow specification of enzyme specificity. In such cases it is useful to know that trypsin does not cleave all sites followed by Arg and Lys equally well. If there are some charged amino acids in the neighborhood of the cleavage site or if the Lys or Arg is followed by a Pro, then the cleavage is generally less efficient. If the sample contains proteases other than trypsin, then one can expect non-tryptic cleavages as well.
3. Fixed and variable modifications. Most search algorithms allow specification of both fixed and variable modifications. Fixed modifications are considered to be present on all of the specified amino acids. Specifying a fixed modification is equivalent to changing the mass of an amino acid and will not have an effect on the computational search time. A frequently used modification is alkylation of Cys residues by, e.g., iodoacetamide which should be specified as a fixed modification. A variable modification may or may not occur on the specified amino acid, for example partial phosphorylation of Ser, Thr, or Tyr residues. Introducing variable modifications has an exponential effect on the computational search time. In practice this means that it is not always possible to test for all the possible variable modifications in one search. It is therefore essential to mine the literature to narrow down the most relevant variable modification for the sample in question. Mascot has a threshold of nine variable modifications. VEMS has searched the human IPI database with 16 variable modifications *(17)*.
4. Mass accuracy. The mass accuracy should be set to include the largest possible mass deviation in the data set. This will depend on the mass spectrometer used. The mass accuracy can be specified as deviation in Da (absolute mass error) or in ppm (parts per million, relative mass error). In mass spectrometry there is often a linear systematic error in the data which is largest for high mass values. Specifying the mass accuracy in ppm has the advantage that it allows higher mass deviation for larger peptides, which is often observed. If the maximum allowed deviation in ppm is set to 10 ppm, then it would correspond to 0.01 Da for a 1 kDa peptide and 0.02 Da for 2 kDa peptide. In Mascot the scores and the significance values are dependent on the specified mass accuracy. The VEMS program gives the same score and significance for the same data as long as the mass deviation is set appropriately high to include the worst mass deviation.

2.3. Search Specificity

An important issue when performing database dependent searches is an evaluation of the rate of false-positives versus the rate of true positives. These rates can be visualized using receiver operating characteristics curves (ROC curves) *(27)*. The ROC curves can be made by searching a data set of known proteins against various databases using different search algorithms. Each search algorithm will give one ROC curve.

The search specificity can also be studied by evaluating the false discovery rate. This is often done by searching the protein sequence database first in the forward direction and then in the reverse direction *(28)*. The search in the reverse direction can then be used to evaluate the false discovery rate at different score thresholds. A more realistic approach would be to reverse all the tryptic peptide sequences, leaving all the arginines and lysines at the C-terminal position. The false discovery rate can also be evaluated by searching a large set of random peptides with the same masses as the identified peptides.

2.4. Database Independent Interpretation

Database independent sequencing a.k.a. *de novo* sequencing, is often used if no complete protein sequence database is available for the organism under study. It is also useful to validate results obtained from database-dependent searches. Database-independent interpretation only uses the molecular mass information available in the peptide MS/MS spectra. There are a large number of algorithms available for database independent sequencing (*see* **Table 4.1**). Some database independent algorithms work by generating all possible peptide sequences with a theoretical peptide mass corresponding to the measured mass of the intact peptide, allowing for an instrument-specific mass inaccuracy. Most algorithms use graph theory to find the best amino acid sequence path through the MS/MS spectrum, by "jumping" from peak to peak and finding a matching amino acid residue mass (*see* **Fig. 4.3B**) *(29)*. There are some variations in the methods used to trace the amino acid sequence through the spectrum. For example, some algorithms continuously calculate and exclude fragment masses from other ion series every time a new fragment ion is included in an ion series.

The main differences between various algorithms are that they use different scoring functions, which are often optimized for a specific instrument. PepNovo *(30)*, Lutefisk *(31)*, and VEMS are examples of programs that offer database independent *(de novo)* sequencing algorithms.

When one or several peptide sequences are obtained from a protein by database-independent assignment, they can be merged to search for homology or similarity by using the very useful online tool "MS-BLAST" (http://dove.embl-heidelberg.de/Blast2/msblast.html) *(32)*.

2.5. Quantitation by Mass Spectrometry

Biological systems are always studied by comparing different states, for example, a control state vs. a perturbed state, or temporal protein expression profiles. Quantitative methods for proteomic analysis are required in these situations. Quantitative proteomic analysis by MS is achieved by peptide intensity profiling or by stable isotope labeling as described in more detail below. The publicly available programs VEMS *(17)*, MSquant *(33)*, RelEx *(34)*, and ASAPratio *(35)* are useful programs in quantitative proteomics.

2.5.1. Quantitation by Intensity Profiling

Quantitative proteomics by intensity profiling is done by comparing the measured intensity signals in the different samples *(36)*. Since there is some degree of instrument-introduced variation in the intensity of the signals, it is important to have replicate measurements. In a recent publication, the dose-response of a histone deacetylase inhibitor (PXD101) on human cell cultures was studied by peptide intensity profiling *(37)*. Histones were purified from six different states, corresponding to human cell cultures exposed to 6 different concentrations of histone deacetylase inhibitor. Triplicate LC-MS/MS analysis was performed on the tryptic digest obtained from each of the six states. Quantitation of the peptides that were covalently modified by acetylation on lysine residues required several analytical and computational steps (**Fig. 4.7**). Protein identifications and PTM assignments were performed using the Mascot and VEMS

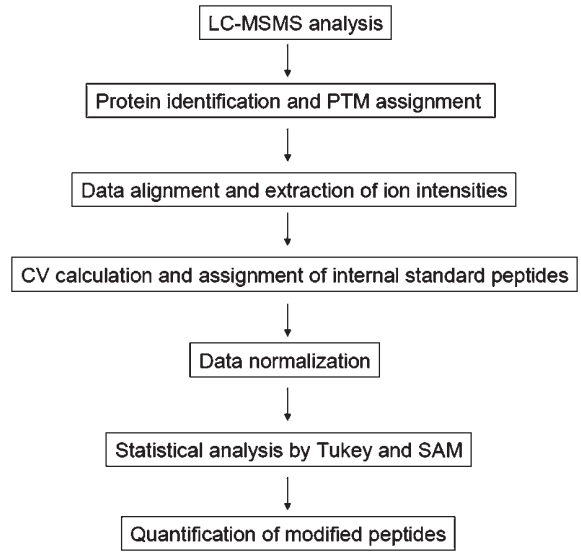

Fig. 4.7. Flow diagram showing the steps in quantitation of post-translational modification by intensity profiling.

search programs. Peptide retention times, peptide masses, and intensities were extracted from the LC-MS raw data. The peptide mass and retention times from the different runs were aligned using the VEMS program. The coefficient of variance (CV) of the ion intensities of all detected and sequenced peptides was calculated and used to identify unmodified peptides with CVs <30% that could be used to normalize the intensity values in the 18 LC-MS/MS runs. Finally, the significance of difference between the intensity values obtained for the different peptide from the six samples was evaluated by a Tukey Q-test *(38)* and SAM analysis *(39)*. This method helped us identify and quantify a variety of acetylated and methylated peptides derived from human histones, thereby revealing the molecular action of the histone deacetylase inhibitor PDX101 *(37)*.

2.5.2. Stable Isotope Labeling

Stable isotope labeling can be used to perform absolute or relative quantitation. In absolute quantitation, the measured peptide mass intensity is compared with the intensity for that peptide labeled with stable isotopes (and added to the sample in known concentrations) *(40)*. In relative quantitation, one or more stable isotope labels are used to label peptides from different samples (**Fig. 4.8**) *(35)*. The different intensities for the peptides with the same sequence but labeled with different stable isotopes is used to calculate the relative quantitations for the samples. The labels can either be introduced chemically *(41)* or by metabolic labeling *(42, 43)*. A broad variety of amino acids labeled with stable isotopes can be used for quantitation. An example of peaks used for relative quantitation for two samples is shown in **Fig. 4.9**. Stable isotope labeling is a very accurate way to make relative comparisons between biological samples. The literature gives many examples of its use, such as studying mitogen activations *(44, 45)*, comparing different cell lines *(46)*, studying the differentiation of cells *(47)*, and differentiating between unspecific and specific binding *(48)*.

2.6. Data Storage

A number of database systems are available for storing proteomics mass spectrometry data. The most difficult part is often to get the data parsed into a database system. Unfortunately, there are currently no good publicly available tools that are able to parse search results and all the experimental settings of importance from the different search engines into one of the publicly available databases such as YASSdb *(49)*, CPAS *(16)*, GPMdb *(50)*, PRIDE *(51)*, Trans-Proteome Pipeline *(6)*, and Proteios *(52)*. However, we are aware of groups that have initiated the development of such parsers. Such functionality is useful to compare search results obtained from different instruments using different instrument settings and/or search engines.

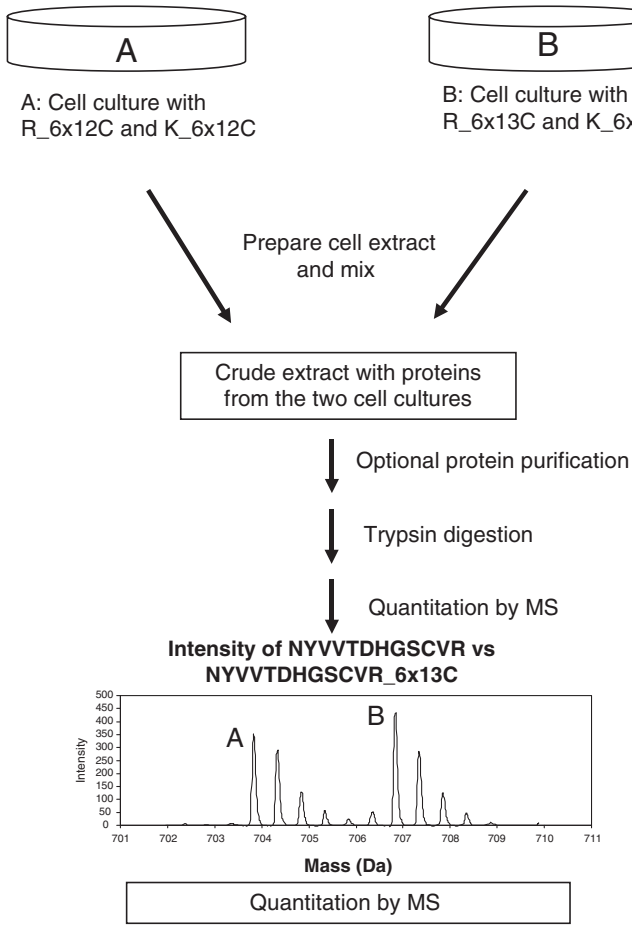

Fig. 4.8. Flow schema of the SILAC method. The result is MS spectra in which peptides from cell culture (**A**) and (**B**) are separated by n*~6 Da where n is the number of stable isotope labeled arginines and lysines in the corresponding peptide. R_6x^{13}C is abbreviation for arginine where six ^{12}C are substituted with six ^{13}C and similar symbols (K instead of R) are used for lysine.

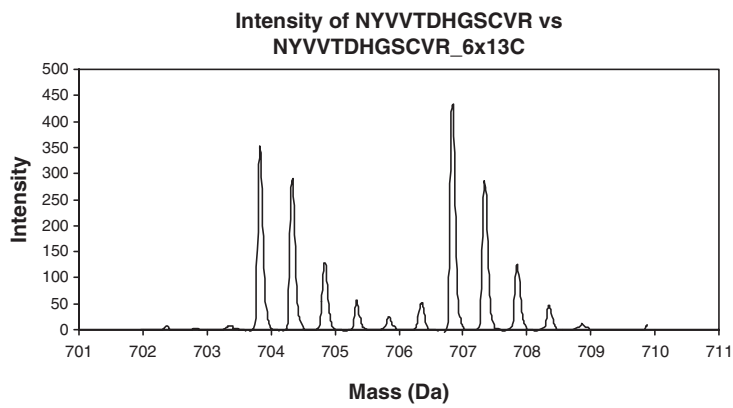

Fig. 4.9. Relative quantification between the light and heavy version of the peptide NYVVTDHGSCVR originating from two different samples. R_6x^{13}C is abbreviation for arginine where six 12C are substituted with six ^{13}C.

3. Conclusion

Mass spectrometry is a very effective proteomics tool for identification and quantitation of proteins and for mapping and quantitation of their post-translational modifications. The diversity of experiments in the proteomics field makes the data analysis challenging and exciting. Computational tools for interpretation of mass spectra, sequence database searching, protein quantitation, and data storage are continuously developed. From the computational point of view, there are still many challenges and problems to solve. With the continuous emergence of novel experimental strategies and analytical instruments, the demand for advanced computational algorithms for data analysis and data integration in proteomics will grow in the coming years.

4. Notes

1. In top-down sequencing, intact proteins are fragmented directly in the mass spectrometer. This is a technique that is under development; therefore, it is not a technique that is widespread in different laboratories.
2. The cleavage method used should be compatible with the mass spectrometer employed. For example, the peptide fragments should be in a mass range that can be measured by the mass spectrometer. Trypsin generates suitable peptide masses in the m/z range covered by most mass spectrometers, i.e., m/z 400 to 3,500. The cleavage method should also generate peptides with a low number of basic amino acids, so that peptides of low charge states are formed in the gas phase. Low charge states simplify the MS and MS/MS spectra and ensure that the m/z values for peptides are within the range of the mass spectrometer. Most tryptic peptides have one basic amino acid as the C-terminal residue.
3. Good efficiency and specificity of the cleavage methods makes it easier to identify peptides in a database since it can be used as an extra constraint during the search. Trypsin has a high specificity; however, the cleavage efficiency is known to be lower for cleavage sites that have neighboring charged residues. For example, the peptide ..ESTVKKT.. or ..ESTVDKT.. will not be cleaved with the same efficiency by trypsin as the peptide ..ESTVAKT..
4. Leucine and isoleucine have identical mass. Therefore, standard MS and MS/MS strategies cannot distinguish proteins that

only differ by having leucine substituted for isoleucine. Further complications arise when two distinct, but near-identical peptides (e.g., EVESTK and VEESTK) have the same mass.

5. The mass spectrometry software normally selects the most intense peaks with a charge state of +2 or higher. The reason for this is that peptides with a charge state above +2 are more likely to produce good fragmentation spectra.
6. The centroid mass m_c and the corresponding intensity I_c can be calculated by the following expressions:

$$m_c = \frac{\sum_{y_i > y_{i,\max}{}^x} m_i I_i}{I_c}.$$

$$I_c = \sum_{y_i > y_{i,\max}{}^x} I_i$$

where m_i is the mass at a certain mass bin and I_i is the corresponding intensity. x is a specified percentage of the maximum intensity.

7. Convolution is a process in which a function g convolves (transforms) another function:

$$f^*g = h$$

f can, for example, be a function that describes the physical quantity of interest. The function g has convolved the function f to the measured function h. Deconvolution is the reverse of convolution; therefore, it can be used to obtain the function f, which describes the physical quantity of interest.

Acknowledgments

R.M was supported by the EU TEMBLOR (IntAct) project and by a Carlsberg Foundation Fellowship. O.N.J. is a Lundbeck Foundation Research Professor and the recipient of a Young Investigator Award from the Danish Natural Science Research Council.

References

1. Kozak, M. (2006) Rethinking some mechanisms invoked to explain translational regulation in eukaryotes. *Gene* Available online 22 June.
2. Seet, B. T., Dikic, I., Zhou, M. M., et al. (2006) Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 7, 473–483.
3. Jensen, O. N. (2006) Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol* 7, 391–403.
4. Aebersold, R., Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207.
5. Patterson, S. D., Aebersold, R. (1995) Mass spectrometric approaches for the

identification of gel-separated proteins. *Electrophoresis* 16, 1791–1814.

6. Domon, B, Aebersold, R. (2006) Challenges and opportunities in proteomic data analysis. *Mol Cell Proteomics*. Available online 8 August.
7. Patterson S. D. (2003) Data analysis: the Achilles heel of proteomics. *Nat Biotechnol* 21, 221–222.
8. Steen, H., Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5, 699–711.
9. Fridriksson, E. K., Beavil, A., Holowka, D., et al. (2000) Heterogeneous glycosylation of immunoglobulin E constructs characterized by top-down high-resolution 2-D mass spectrometry. *Biochemistry* 39, 3369–3376.
10. Jensen, O. N., Larsen, M. R., Roepstorff, P. (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: strategies and applications. *Proteins* 2, 74–89.
11. Roepstorff, P., Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* 11, 601.
12. Wysocki, V. H., Tsaprailis, G., Smith, L. L., et al. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* 35, 1399–1406.
13. Laskin, J., Futrell, J. H. (2003) Collisional activation of peptide ions in FT-ICR. *Mass Spectrom Rev* 22, 158–181.
14. Pedrioli, P. G., Eng, J. K., Hubley, R., et al. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 22, 1459–1466.
15. Orchard, S., Kersey, P., Hermjakob, H., et al. (2003) The HUPO Proteomics Standards Initiative meeting: towards common standards for exchanging proteomics data. *Comp Funct Genom* 4, 16–19.
16. Cottingham, K. (2006) CPAS: a proteomics data management system for the masses. *J Proteome Res* 5, 14.
17. Matthiesen, R., Trelle, M. B., Højrup, P., et al. (2005) VEMS 3.0: Algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J Proteome Res* 4, 2338–2347.
18. Fenyo, D., Qin, J., Chait, B.T. (1998) Protein identification using mass spectrometric information. *Electrophoresis* 19, 998–1005.
19. Matthiesen, R., Bunkenborg, J., Stensballe, A., et al. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* 4, 2583–2593.
20. Fermin, D., Allen, B. B., Blackwell, T. W., et al. (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* 7, R35.
21. Fenyö, D., 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.
22. Creasy, D. M., Cottrell, J. S. (2002) Error tolerant searching of tandem mass spectrometry data not yet interpreted. *Proteomics* 2, 1426–1434.
23. Craig, R., Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra, *Bioinformatics* 20, 1466–1467.
24. Woodsmall, R. M., Benson, D. A., (1993) Information resources at the National Center for Biotechnology Information. *Bull Med Libr Assoc* 81, 282–284.
25. LinksKersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiment. *Proteomics* 4, 1985–1988.
26. LinksBairoach, A., Apweiler, R. (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res* 26, 38–42.
27. Colinge, J., Masselot, A., Cusin, I., et al. (2004) High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics* 4, 1977–1984.
28. López-Ferrer, D., Martínez-Bartolomé, S., Villar, M., et al. (2004) Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Anal Chem* 76, 6853–6860.
29. Dancik, V., Addona, T., Clauser, K., et al. (1999) De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 6, 327–342.
30. Frank, A., Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 77, 964–973.
31. Johnson, R. S., Taylor, J. A. (2002) Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol Biotechnol* 22, 301–315.
32. Shevchenko, A., Sunyaev, S., Loboba, A., et al. (2001) Charting the proteomes of organisms with unsequenced genomes by

MALDI-Quadrupole time-of flight mass spectrometry and BLAST homologuey searching. *Anal Chem* 73, 1917–1926.

33. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426, 570–574.
34. MacCoss, M. J., Wu, C. C., Liu, H., et al. (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* 75, 6912–6921.
35. Venable, J. D., Dong, M. Q., Wohlschlegel, J., et al. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 1, 39–45.
36. Listgarten, J., Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* 4, 419–434.
37. Beck, H. C., Nielsen, E. C., Matthiesen, R., et al. (2006) Quantitative proteomic analysis of post-translational modifications of human histones. *Mol Cell Proteomics* 5, 1314–1325.
38. Zar, J. H. (1999) *Biostatistical Analysis*. Prentice-Hall, Upper Saddle River, NJ.
39. Tusher, V. G., Tibshirani, R., Chu, G., et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98, 5116–5121.
40. Gerber, S. A., Rush, J., Stemman, O., et al. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* 100, 6940–6945.
41. Turecek, F. (2002) Mass spectrometry in coupling with affinity capture-release and isotope-coded affinity tags for quantitative protein analysis. *J Mass Spectrom* 37, 1–14.
42. Ong, S. E., Blagoev, B., Kratchmarova, I., et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteom* 1, 376–386.
43. Yang, W. C., Mirzaei, H., Liu, X., et al. (2006) Enhancement of amino Acid detection and quantification by electrospray ionization mass spectrometry. *Anal Chem* 78, 4702–4708.
44. Gruhler, A., Schulze, W. X., Matthiesen, R., et al. (2005) Stable isotope labeling of *Arabidopsis thaliana* cells and quantitative proteomics by mass spectrometry. *Mol Cell Proteom* 4, 1697–709.
45. Ballif, B. A., Roux, P. P., Gerber, S. A., et al. (2005) Quantitative phosphorylation profiling of the ERK/p90 ribosomal S6 kinase-signaling cassette and its targets, the tuberous sclerosis tumor suppressors. *Proc Natl Acad Sci U S A* 102, 667–672.
46. Fierro-Monti, I., Mohammed, S., Matthiesen, R., et al. (2005) Quantitative proteomics identifies Gemin5, a scaffolding protein involved in ribonucleoprotein assembly, as a novel partner for eukaryotic initiation factor 4. *J Proteome Res* 5, 1367–1378.
47. Romijn, E. P., Christis, C., Wieffer, M., et al. (2006) Expression clustering reveals detailed co-expression patterns of functionally related proteins during B cell differentiation. *Molecular & Cellular Proteomics* 4, 1297–1310.
48. Blagoev, B., Kratchmarova, I., Ong, S. E., et al. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat Biotechnol* 21, 315–318.
49. http://www.yass.sdu.dk/yassdb/
50. Craig, R.., Cortens, J. P., Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 3, 1234–1242.
51. Jones, P., Cote, R. G., Martens, L., et al. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res* 34, D659–663.
52. Gärdén, P., Alm, R., Häkkinen, J. (2005) Proteios: an open source proteomics initiative. *Bioinformatics* 21, 2085–2087.