Limitations and Pitfalls in Protein Identification by Mass Spectrometry

Gert Lubec* and Leila Afjehi-Sadat

Medical University of Vienna, Department of Pediatrics, Waehringer Guertel 18, A-1090 Vienna, Austria

Received October 31, 2006

Contents

1. Introduction 3568
2. Limitations of Peptide Identification (PI) by Sample Preparation 3569
2.1. Spot Picking 3569
2.2. In-Gel Protein Digestion 3569
2.2.1. Selection of the Protease(s) or Chemical Agents 3569
2.2.2. Nonspecific Cleavage and Missed Cleavages 3570
2.3. Matrix as Limiting Factor 3570
2.4. The Target (Sample Support) as a Factor for PI 3570
2.5. Contaminants Hampering the PI Process 3571
3. Limitations of PI by Instrumentation 3571
3.1. The Role of Calibration for PI 3571
3.2. The Factor “LASER” for PI 3572
3.3. Importance of Maintenance 3572
3.4. Selecting the Mass Spectrometry Method 3572
3.4.1. Limitations of MALDI Technology 3572
3.4.2. Limitations of ESI Technology 3572
4. Limitations of PI Due to Data Processing and Data Mining 3573
4.1. Spectra Quality 3573
4.2. Peptide Identification 3573
4.2.1. PI by Databases 3573
4.2.2. PI by “Database-Independent” Strategies 3574
4.3. Database Errors—A Major Unsolved Problem for PI and PrI 3574
5. Validation of Peptide and Protein Identification 3575
5.1. Scoring for MS/MS Peptide Identification 3575
5.2. Use of Randomized Databases for the Validation of PI 3575
6. Limitations of PI by Protein Properties 3576
6.1. Short and Very Short Proteins 3576
6.2. Hydrophobic Proteins 3576
6.3. The Isobaric Amino Acids Problem 3576
6.4. The Problem of Low Complexity Regions of Proteins 3577
6.5. Hypothetical Proteins, Proteins with Unknown Function, and Unknown Proteins 3578
6.6. Protein Modifications: Artifacts and Post-translational Modifications 3578
6.6.1. Chemical Modifications 3579
6.6.2. Post-translational Modifications 3579
6.7. Splice Variants/Isoforms 3579
7. Protein Identification 3580
8. Conclusion 3580
9. Acknowledgments 3581
10. References 3581

1. Introduction

The advent of proteomics techniques for protein identification (PrI) represented a major step forward in protein chemistry,[1,2] and indeed, a legion of laboratories are now using different methods of mass spectrometry for peptide identification (PI). Many thousands of proteomics articles are published annually, but not all proteins from these reports have been unambiguously identified. Criteria for reliable PrI became gradually more and more stringent, and a significant part of PrIs from the past—and unfortunately also some present PrIs—are not reliable.[3] In this review, limitations and pitfalls are seen from two different viewpoints: that of a user and that of an editorial board member and reviewer of a top proteomics journal. Herein, problems, limitations, and shortcomings are addressed, and most mistakes were made also in our laboratory in the beginning of the proteomics area at different levels, instrumental and data mining. By and by, knowledge exponentially increased, and literature about PrI is abundant.

While a host of publications praises and proposes the use and applications of mass spectrometry for PI, there is not too much information on the limitations and pitfalls. The wide use of mass spectrometry techniques is hampered by several factors: one factor is the limited experience of some investigators in mass spectrometry per se;[4] a second factor is poor bioinformatic know-how, i.e., data mining (this a major factor for non- or misidentifications); and a third factor is low abundance proteins.[5,6] A multitude of problems linked to PrI is being addressed herein, and these range from selecting a MS method to selecting an appropriate database. This review is not designed to address all open questions of MS technologies or for troubleshooting but to indicate some potential weaknesses of PrI. It is written to provide information on the reliability of PrI and finally on validation of the identification process for users of MS. It may serve to enable scientists to critically read publications in the field of proteomics and to probably avoid some mistakes and pitfalls.

The article may be useful for the peer reviewing process to test the validity and confidence of identification methods. At this point, a controversy can be mentioned: that is, the fact that results from automated methods are very often and correctly not considered appropriate in favor of human experience. This may enable new discoveries and disprove erroneous existing data. On the other hand, human ap-

* Telephone: +43-1-40400-3215. Fax: +43-1-40400-3200. E-mail: gert.lubec@meduniwien.ac.at.

10.1021/cr068213f CCC: $65.00 © 2007 American Chemical Society
Published on Web 07/24/2007

Gert Lubec is a Full Professor at the Medical University of Vienna and has been working in the field of protein chemistry for two decades. With the advent of proteomics, he focused on neuroproteomics and carried out basic analytical methodology as well as applied work in animal models of several central nervous system disorders, such as Down's Syndrome. He is currently involved in the investigation of proteins linked to cognitive function in rodents.

Leila Afjehi-Sadat is a Junior Scientist in the Neuroproteomic Laboratory of G.L. and is working on basic proteomic methodology including protein renaturation and the determination of the biological activity of proteins from gel spots.

proaches may sometimes fail, in particular in the high-throughput applications. Therefore, a combination of both automated and manual approaches may lead to optimal PrI, despite possible limitations due to time restrictions. Last but not least, the goal of the present review is to prevent further accumulation of false PrI in literature and databases. It may complement guidelines for publication of peptide and protein identification data.[3,7,8] It must be stated, however, that only a selection of PrI problems is identified and that, as the review is also based upon our own experience, priority is given to issues in MALDI-TOF/TOF and nano-LC-ESI-MS/MS technologies.

2. Limitations of Peptide Identification (PI) by Sample Preparation

We are not addressing the more than complex sample preparation procedures but highlight some common problems limiting PI at this level. Based upon our own experience, we start describing the problems from spot picking from a 2-DE gel as an example.

2.1. Spot Picking

Spot picking can be carried out manually or automatically, and the earlier the spot is picked following preparation of the gel, the better are the identification results, although proteins from very old—i.e. many year old—gels have been (not unambiguously) identified.

Manual spot picking can be used for individual samples only, due to the time factor. The general disadvantages of manual analyses, including a higher need for personal mix-up of samples and keratin contamination from human skin and hair (in the experience of our laboratory, the increased time for manual spot picking leads to a higher probability of contamination; unpublished observation)[9] that in turn would hamper PI, should be considered. Sample to sample contamination is also higher in manual spot picking (http://www.shimadzu-biotech.net/pages/products/2/xcise.php).

Automated spot picking is a must in high-throughput MALDI-TOF or MALDI-TOF/TOF analysis[10] to make identification reliable. The use of the automated spot picker Proteineer SP (Bruker Daltonics, Bremen, Germany) is a valuable tool in our laboratory when used with disposable tips and gives good results with the exception of the following limitations: due to the picker, we are losing <5% of the protein spots. This is due to technical reasons such as, e.g., problems with cutting tips from the spot picker. Anyway, this represents a significant limitation, as important proteins may be lost by this procedure and cannot undergo the identification process. Some of these nonpicked spots can be recovered by manual spot picking, but this is cumbersome work.

2.2. In-Gel Protein Digestion

2.2.1. Selection of the Protease(s) or Chemical Agents

Protein digestion is the heart of sample preparation, and the selection of the method and the protease is of utmost importance to enable high sequence coverage and subsequent unambiguous PI.[11]

As to the selection of the protease(s) used, trypsin, cleaving exclusively C-terminal to arginine and lysine residues,[12] may be the first method of choice to generate peptides because the masses of generated peptides are compatible with the detection ability of most mass spectrometers (up to 2000 *m/z*), the number and average length of generated peptides, and also the availability of efficient logarithms for the generation of databases of theoretical trypsin-generated peptides. High cleavage specifity, availability, and cost are other advantages of trypsin.[12−14] Further enzymatic cleavage of proteins with low sequence coverage/low number of identified peptides has to be performed according to a secondary strategy (Peptide Cutter; http://www.expasy.org/tools/peptidecutter/).

The use of the wrong protease is a major limiting factor for generation of peptides that can be subsequently used for identification.[15−17] Apart from sequence cutter searches, several reports try to overcome the limitation of digestion of hydrophobic peptides.[15] It may even happen that all proposed enzymes fail to split an individual protein properly, and in this case chemical cleavage has to be employed.[17,18]

Another factor is the choice of a protease that is suitable for use in mass spectrometry, and indeed, there are enormous limitations by individual enzyme preparations and products.

In our laboratory, e.g., one commercially available trypsin product (F. Hoffmann-La Roche Ltd, Basel, Switzerland) leads to significantly lower PI by MALDI-TOF/TOF in contrast to the procrine trypsin (Promega, Madison, WI;

Table 1. Proteases Used in Our Laboratory To Obtain the Highest Sequence Coverage

enzyme	company (cat./part no.)	cleave	N- or C-terminal	buffer	pH	protease specific digestion conditions	
trypsin	Promega (V5113)	KR	C-terminal	10 mM NH_4CO_3	7.8	trypsin conc	40 ng/μL
						digestion time	4 h
						digestion temp	29 °C
Asp-N	Roche (11 420 488 001)	DE	N-terminal	25 mM NH_4CO_3	7.8	Asp-N conc	25 ng/μL
						digestion time	18 h
						digestion temp	37 °C
chymotrypsin	Roche (1418467)	FYWL	C-terminal	30 mM NH_4CO_3	7.8	chymotrypsin conc	40 ng/μL
						digestion time	1.5 h
						digestion temp	29 °C
Lys-C	Roche (11 420 429 001)	K	C-terminal	10 mM NH_4CO_3	7.8	Lys-C conc	30 ng/μL
						digestion time	18 h
						digestion temp	37 °C
subtilisin	Fluka (82490)	unspecific cleavage		6 M urea/1 M Tris (pH 8.5) 50 mM NH_4HCO_3	8.5	subtilisin conc	100 ng/μL
						digestion time	1 h
						digestion temp	37 °C

modified), and this holds also for other commercially available proteases (Table 1).

In many instances, a series of proteases has to be used to produce sufficient peptides for unambiguous PI.

2.2.2. Nonspecific Cleavage and Missed Cleavages

These cleavages represent peptides whose termini do not reflect common cleavage patterns of a protease used. A vast variety of reasons may be responsible for this phenomenon: protease impurities or contamination with other proteases, nonspecific proteolysis that may have taken place in vivo or during sample preparation, or autolytic cleavage of the protease(s) used.[19] Of course, experimental conditions, including solvents, buffers, temperature, and incubation time (longer incubation times increase the likelihood of nonspecific cleavages), are confounding factors. The protein/protease ratio is another factor to explain nonspecific cleavage, and indeed proteins on a gel present within a wide range of levels.[20] The protein's primary structure (i.e., the neighboring amino acids at the cleavage site) is another factor that may lead to nonspecific attack of enzymes.[21] Thiede and co-workers have shown that proline at a certain position accounted for 90% of missed tryptic cleavage sites after Arg and Lys.[22] Miscleavage is considered a major factor for failure or ambiguous results of PI and may not be avoided: [23] The presence of post-translational modifications (PTMs) is a major contributor to the problem of miscleavages[24,25] as well.

Sumoylation is, e.g., a good example of proteolytic miscleavages, as shown by Chung and co-workers.[26] Likewise, phosphorylation has been shown to lead to miscleavages,[27,28] and so does ubiquitination;[27] trypsin does not cleave efficiently at acetylated lysine residues,[29] to name a few examples.

"Missed cleavages" can be defined as partial enzymatic protein cleavages generating peptides with internal missed cleavage sites[14] reflecting the allowed number of sites (targeted amino acids) per peptide that were not cut. This is an error allowance for enzyme inefficiency/partial cleavage (http://phenyx.vital-it.ch/docs/pwi/SubmissionEffects.html). The use of missed cleavage sites in tryptic peptides is a useful tool in peptide identification.[14]

As to the influence of chemical modification of proteins on miscleavages, carboxymethylation, widely used in gel based proteomics, was shown to lead to miscleavages by Sellinger and Wolfson,[30] and so do artifactual protein modifications known to arise from 2-DE and sample processing. The multitude of chemical modifications/artifacts on glial fibrillary acidic protein is shown in Table 2.

2.3. Matrix as Limiting Factor

An ideal matrix for MALDI-TOF/TOF would not generate an interfering chemical background and would provide good sensitivity for peptides. However, there is no single matrix fulfilling these criteria for all kinds of peptides in a setting. Therefore, the scientific community in a first step uses a matrix, α-cyano-4-hydroxycinnamic acid (CHCA), which is suitable for a broad area of different peptides.[31] Subsequently, specifically tailored matrices may be used to optimize a matrix for specific peptides. Different matrices influence ionization behavior, formation of adducts, stability, or fragmentation of analytes. This issue was recently addressed by Tholey and Heinzle.[32] Gonnet and co-workers[33] have been addressing the effect of four different matrices on protein identification and propose to use at least two different matrices in order to increase peptide matches and sequence coverage. More specifically, CHCA is said to be a matrix only good for peptides with mass ions below 2500 Da[34,35] whereas sinapinic acid may be recommended for higher masses.[33,36] Kussmann and co-workers,[37] Land and Kinsel,[38] as well as Yao et al.[39] propose the use of 2,5-dihydroxybenzoic acid (DHB) as the matrix of choice for the identification of hydrophobic peptides or modified peptides. For analysis of glycosylated or phosphorylated peptides, DHB as well as 3-hydroxypicolinic acid (3-HPA) would be suitable.[40] Using the inappropriate matrix, therefore, would represent a serious limitation of unambiguous PI.

2.4. The Target (Sample Support) as a Factor for PI

There are two basic sample supports for MALDI applications, metallic or polymer-based targets. While, for high-throughput analysis, "standard", commercially available anchorchips are frequently in use, some limitations were described, and for optimization, specific targets may be used. McComb and co-workers[41] propose the use of polyurethane targets for the analysis of high molecular weight proteins. Hung et al.[42] described the use of Teflon sample supports claiming to produce homogeneous coverage of the matrix over the sample surface and enhancing sensitivity and salt tolerance. Schuerenberg and co-workers[43] propagate the use

Table 2. Chemical Modifications in Rat GFAP (Glial Fibrillary Acidic Protein) Identified by Q-TOF Analyses Frequently Observed in Our Laboratory

modification	observed modified amino acid	position	mass shift (Δ(*m*/*z*))	source
deamidation	asparagine glutamine	75	0.98	post-translational modification; artifact
oxidation	methionine	19; 40	15.99	post-translational modification; artifact
carboxymethyl	cysteine	292	43.00	artifact
methyl ester	glutamic acid	156	14.01	post-translational modification; artifact
	glutamic acid	162		
	threonine	148		
	threonine	363		
	leucine	161		
	glycine	370		
pyro-Glu	glutamine	176; 286	17.03	post-translational modification; artifact
amidation	arginine	103	−0.98	post-translational modification; artifact
		119		
		134		
		150		
		181		
		171		
		268		
		328		
		365		

of prestructured sample supports based upon a gold/Teflon surface with advantages of increased detection sensitivity by sample concentration.

Redeby et al.[44] propose improved analysis of hydrophobic proteins using a specific target plate using a fluorinated organic solvent and a silicone polymer layer. A major improvement can be seen, as the fluorinated organic solvent enabled even analyte distribution on the target.[45] Kleno and co-workers[46] reported a protocol for on-probe protein digestion suitable for hydrophobic proteins that reduces the number of analytical steps necessary and leads to improved sequence coverage.

It must, however, be mentioned that in our laboratory we never experienced any PI problem for all kinds of proteins due to the use of our sample support (AnchorChip targets, SCOUT MALDI MTP with hydrophilic 600 μm patches in a hydrophobic surrounding; Bruker Daltonics, Bremen, Germany) in the high-throughput application of brain protein extracts.

2.5. Contaminants Hampering the PI Process

During the whole analytical process, a series of contaminants can be introduced and chemical noise is painstaking.[47] A protein sample can be contaminated by other proteins or by chemicals used during one of the steps for PI. Principally, two major forms of contaminations can be differentiated, endogenous and exogenous, and both may seriously interfere with fair PI.

Endogenous contaminations are mainly cross contaminations; that is, two or more proteins derived from the same or different samples are being analyzed.[48] As stated by Ding et al.,[49] contamination accounts for many unmatched masses and it is well-known that most of the interfering masses are derived from keratins and trypsin autolysis products. Subtraction of known contaminants from raw data is of pivotal importance to optimize or even enable reliable PI, but this filtering can never be complete.

Barsnes and co-workers[50] demonstrate Mass Sorter: a tool to filter contaminants including keratins, proteins comigrating with the proteins of interest, and others.

Schmidt and co-workers[51] published the iterative data analysis MS-Screener contaminant searches, calculations of half decimal places, elimination of contaminants, and screening of common masses, and their rankings can be evaluated in one set;[52] the removal of contaminants, e.g., resulted in significant and remarkable improvement of the identification rates of helicobacter pylori proteins.

Samuelsson et al.[53] show how scoring performance varies with contamination levels and protein sequence coverage using the PIUMS (Protein Identification Using Mass Spectrometry) algorithm.

The multitude of known and unknown contaminants makes total elimination of misinterpretations by software programs impossible, and one has to start solving the problem by working at high laboratory (hygienic) standards with well-defined materials (e.g., vials or tubes, etc.) and chemicals and by actively reading in the literature how to avoid contaminations.

For a list of contaminants and chemicals filtered in our mass spectrometry system, see Table 3.

3. Limitations of PI by Instrumentation

3.1. The Role of Calibration for PI

All mass spectrometry techniques rely on calibration, usually performed by the use of external and/or internal calibrants of known molecular masses.[54−58] Miscalibration or poor calibration is one of the main errors leading to misidentification of proteins. There may be, however, pitfalls by the use of calibrants: in some cases, the signal of a calibrant might be suppressed by the analyte peptides. On the other hand, the calibrant signal may partially overlap with the analyte signal, resulting in a false assignment of spectra.

In addition to the methods cited, a method independent of internal and external standards has been reported by Wolski and co-workers.[59] Using the algorithm of their combined MS spectra calibration strategy, the identification rate could be improved by between 5 and 15%. Wu and co-workers[60] developed COFI (Calibration Optimization on Fragment Ions), against being independent of internal and external calibrants. Its use has been achieving an average measured mass accuracy of 2.49 ppm for all identified bovine serum albumin peptides.

Table 3. List of Contaminants and Chemicals Filtered in Our Mass Spectrometry System; Mass Tolerance Is 25 ppm

peak label	*m*/*z*
T (trypsin)_porcine (Promega)	842.5094
	1045.5637
	1713.8084
	1774.8975
	2083.0096
	2211.104
	2283.1802
T (trypsin)_bovine (Roche)	659.38
	2163.057
	2273.16
	2289.155
keratin	1066.44
	1232.62
	1277.7
	1307.68
	1399.53
	1383.69
	1475.78
	1638.86
	1791.73
	2150.08
	2184.1
	2383.93
keratin	2501.25
	2510.13
	2705.16
	2932.52
	3264.52
	2825.4056
keratin 1/II	1179.601
	1300.5302
	1716.8517
	1993.9767
keratin 10	1165.5853
	2825.4056
α-cyano-4-hydroxycinnamic acid	568.13
	855.1
	1060.1
coomassie	804.28
	818.3
angiotensin II	1046.54
angiotensin I	1296.685
substance P	1347.736
bombesin	1619.823
adrenocorticotropic hormone 1-17	2093.0868
adrenocorticotropic hormone 18-39	2465.199

3.2. The Factor "LASER" for PI

A laser appropriate for a certain matrix may not be good for other matrices, and therefore, a laser type fulfilling all requirements in all systems cannot be recommended.

Nd:YAG lasers have been sucessfully employed for MALDI analyses of peptides using α-cyano-4-hydroxycinnamic acid (CHCA) as a matrix, but this laser type is not appropriate when other matrices including sinapinic acid preparations of peptides are being applied.[61] This fact represents a limitation and a pitfall for the many MS users that have acquired standard equipment and are not thinking of varying (or are not able to vary) the laser for individual protein identification experiments. Thus, optimal PI maybe hampered by the use of a single laser bought along with instrumentation.

Important information enabling optimization of laser−matrix combinations is available at http://www.sigmaaldrich.com/Brands/Fluka___Riedel_Home/Bioscience/Peptide_Analysis/MALDI_Mass.html.

3.3. Importance of Maintenance

Although some companies claim that MS instruments are free of maintenance, a typical example of maintenance-induced problems is the cleaning of the ion source.[62] In our MALDI-TOF/TOF instrumentation system running in the high-throughput mode, we have to clean the ion source after ten targets have been used. After analyzing ten targets, peak intensity and peak resolution become gradually worse. This is a limiting factor that has to be taken into account for fair PI, and this limitation has not been addressed in the literature thus far.

Cleaning of the ion source is the most important maintenance step, and it should be carried out regularly, as failure to do so leads to impaired peak stability.

3.4. Selecting the Mass Spectrometry Method

There is no individual method that can identify all proteins, and of course, a major limitation is the factor sensitivity. Even the very best instrumentation and method[63,64] are in everyday life not sensitive enough to reliably identify minor or weak spots by most staining methods. It must be stated that the sensitivity of mass spectrometers has not kept pace with the most sensitive staining of protein spots in the gel. Even if one or another spectrum is generated, weak spots do not contain sufficient material to carry out enough MS/MS spectra for reliable identification.

There is some consensus in the scientifique community that electrospray ionization (ESI) and MALDI are complementary ionization techniques that in combination lead to high protein identification rates.[65−73] In any case, no comparison about superiority can be evaluated. Domon and Aebersold[74] tried to evaluate the characteristics and performances of commonly used MS technologies, and some conclusions can be drawn indirectly from this review and from comparison of methods by Lim and co-workers.[75]

3.4.1. Limitations of MALDI Technology

A series of examples addresses the weaknesses of the MALDI methodology: MALDI does not favor identification of hydrophobic peptides.[76] MALDI may be inferior to ESI methods in terms of quantifications due to an inhomogeneous distribution of peptides in the matrix, and MALDI technologies are more susceptible to interference with chemical noise.[47] The sensitivity of MALDI may be inferior to that of ESI technologies.[77] According to Hansen et al.,[67] peptides of lower molecular mass were generally favored by ESI whereas MALDI tended to identify fewer but larger peptides.

3.4.2. Limitations of ESI Technology

Studies have shown that, in contrast to MALDI methods, identification of basic residues is not favored by ESI.[34,40,78−80] ESI is very sensitive to modest amounts of salt and/or detergents as well as impurities that are more likely to compete successfully for the available charge at the expense of the analyte.[81] ESI, as a flowing technique and unlike the MALDI methodology, consumes the entire amount of a peptide preparation within the component elution time.[82] The complexity of ions generated in ESI modes is enormous and complicates analysis.[83,84] Furthermore, ESI resolution is limited to an effective upper limit of 100000 atomic mass units.[77] The level of expertise needed to assemble the needle structure and pack it with particular supports along with the choice of conductive polymeric materials used for the

Table 4. Comparison of Advantages and Disadvantages of Some Individual Mass Spectrometry Methods[a]

mass spectrometry method	mass acccuracy	resolving power	sensitivity	dynamic range	identification	quantification	throughput	PTMs
IT (ion trap) mass analyzers[b,74]	+	+	+++	+	+++	+	++++	+
QQ (hybrid quadrupole)-TOF (time-of-flight)[c,74]	+++	+++		++	+++	++++	+++	+
TOF—TOF[d,74]	+++	+++	+++	++	+++	+++	++++	+
FT (Fourier transform)-ICR (ion cyclotron resonance)[e,74]	++++	++++	++	++	++++	+++	+++	+
QQQ (triple quadrupole)[f,74]	++	+	+++	+++	+	++++	+++	
QQ-LIT (linear ion trap)[g,74]	++	+	+++	+++	+	++++	+++	++++

[a] +, low or possible; ++, medium; +++, good or high; ++++, excellent or very high. [b] Thevis, M.; Makarov, A. A.; Horning, S.; Schanzer, W. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3369. [c] Williams, J. P.; Nibbering, N. M.; Green, B. N.; Patel, V. J.; Scrivens, J. H. *J. Mass Spectrom.* **2006**, *41*, 1277. [d] Liu, Z.; Schey, K. L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 482. [e] Peng, W. P.; Cai, Y.; Chang, H. C. *Mass Spectrom. Rev.* **2004**, *23*, 443. [f] Hager, J. W.; Yves Le Blanc, J. C. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 1056. [g] Hopfgartner, G.; Varesio, E.; Tschappat, V.; Grivet, C.; Bourgogne, E.; Leuthold, L. A. *J. Mass Spectrom.* **2004**, *39*, 845.

nanospray needle and the use of sanded needles is a serious problem and may lead to unreproducible results.[85−87] In ESI QQ (hybrid quadrupole)-TOF, information-dependent acquisition technology would overlook peptides that cannot be selected for CID (collision-induced dissociation), as they coelute with others, giving stronger signals.[67]

In Table 4 a short comparison of the advantages and disadvantages of different methods is listed.

4. Limitations of PI Due to Data Processing and Data Mining

4.1. Spectra Quality

A significant number of laboratories are publishing work without primarily respecting spectra quality. Spectra quality is defined by three basic components: (a) charge state differentiation, (b) total signal intensity, and (c) signal-to-noise estimates. Spectra are sent for database searches without prior controlling quality, and this is a major factor for the limitation of PI.

Spectra generated should be refined by spectra quality filters, thus performing MS/MS spectra quality assessment prior to application of PI methods in order to avoid submission of poor quality spectra to databases.[88,89] Tabb and co-workers,[90] for example, introduced preliminary rules for prefiltering, including minimum and maximum thresholds on number of peaks and a minimum threshold on peak intensity; this method may remove about 40% of poor quality spectra. Bern et al.[91] reported an algorithm that is able to remove up to 75% of "bad" spectra while losing only 10% of high quality spectra. One possibility of scoring is proposed by Purvine et al.,[92] showing that differentiation between single and multiple precursor states provides a partial binary score and components b and c provide partial scores which are then subsummarized to form the final score that forms the basis for a SPEQUAL automated quality assessment; this algorithm is based on intensity-based scoring.

In contrast, Bern and co-workers[91] propose the use of a ranking algorithm insofar as rank versus probability fits a negative exponential function: this includes input of ranking and the logarithmic likelihood that the peaks are indeed b- and y-ions.

The basic information on processing and classification of protein mass spectra was recently reviewed in an excellent article that introduces the reader to the general and specific issues.[93]

Although highly sophisticated software exists for the assessment of spectra quality and peak detection,[94−96] the primary goal has to be avoiding factors that generate poor spectra (see above).

4.2. Peptide Identification

4.2.1. PI by Databases

There are two basic principles for matching results from the mass spectrometer with databases.

The first consists of submitting mass peak lists to the databases using, for example, SEQUEST[97] and code developmental programs,[98] MASCOT[99] and STEM,[100] MS-tag,[101] SONAR,[102] TANDEM,[103] ProbID,[104] OMSSA,[105] X!Tandem,[103,106] and Phenyx.[107] The limitation of database-based PI is that unmatched masses cannot be handled: only mass peaks included in the database can be assigned to peptides. This means that this method is very much dependent on database quality and susceptible to database errors and conflicts. This also holds for, for example, mutations, miscleavages, peptide modifications, contaminants, etc.

The second basic principle consists of submitting spectra (system of spectra alignment) and includes the following databases: SpecAlign[108,109] and OMSSA.[105] SpecAlign, for instance, is a graphical computational tool, enabling simultaneous visualization and manipulation of multiple datasets. SpecAlign not only provides all common processing functions but also uniquely implements an algorithm that generates the complete "BLAST-like" alignment of each mass spectrum within a loaded dataset.

There are several algorithms for spectra alignments calculating spectra similarities: (a) cross correlation[110] and (b) spectral contrast angle or dot-product comparison.[111] The second basic principle limitations are, among others, that databases are not comprehensive so far, depend on spectra quality, and in the case of "cross-correlation" are very much dependent on fragmentation patterns. A pitfall of the spectral contrast angle systems is that peptide fragment ion spectra contain more peaks than spectra typically used with dot-product comparison, thus losing discriminatory power.

Databases for searching PTMs by MS data are, for example, UNIMOD,[112] Deltamass,[113] FindMod,[114] and FindPept,[19] and those for MS/MS data are SEQUEST, MOD^{i},[115] and Modificomb.[116] UNIMOD, SEQUEST, Findmod, and Delta mass databases only search for known and a limited number of PTMs. The search results exclusively depend on database quality and errors.

The MOD^{i} database shows a drawback, as the MassPective software has to be used in addition when multiple PTMs are to be analyzed from MS/MS spectra. Another shortcoming of MOD^i is that it is assumed that the number of candidate proteins is limited to 20 or less. MOD^i is furthermore not able to map substoichiometric PTMs, in contrast to ModifiComb, which is able to find novel and unexpected types of modifications.

4.2.2. PI by "Database-Independent" Strategies

De novo sequencing is inferring knowledge about peptide sequences independently of any information from databases. The inferred complete or partial sequences are compared to theoretical sequences using specific similarity search algorithms.

There are several basic principles of "database-independent" de novo sequencing methods:[117,118]

(a) The "pseudo" peptide fragment fingerprinting approach means constructing a "pseudo" sequence database on-the-fly: sequences are generated by determination of all possible amino acid compositions with a total mass matching the experimental precursor mass and subsequently, for each composition, by determining all possible amino acid permutations. Theoretical sequences with the highest scores are the most likely to represent the "original" peptide. The major limitation of this approach is the enormous combinatorial complexity, as the number of possible sequences increases exponentially with the precursor mass although there is some refinement by additional algorithms.[119−121]

(b) The peak succession approach represents an incremental approach: candidate sequences are built in an iterative way, amino acid by amino acid, until complete sequences that account for the precursor mass are obtained. Only partial sequences whose extensions are validated by fragment ions in the spectrum are retained for further extension, thus discarding large subsets of permutations. No sequence gaps are allowed, making the use very much dependent on spectra quality. Further refinements of this principle have been reported and are most useful.[122−130]

(c) The sequence tag approach uses iterative methods as given above; when several consecutive fragmentation positions are missing in the spectrum or when unexpected modifications arise, the path is split into a minimum of two sections and may probably lead to wrong sequences. Therefore, Mann and Wilm[131] proposed to limit de novo sequencing to "islands" of consecutive ions that can generally be observed in the high mass region of spectra, called sequence tags. Guten Tag[132] employs this principle and introduces an enhanced scoring system, extracting the tags by recursively parsing a spectrum graph and assuming all peaks as y-ion types as well as limiting sequence tag lengths. Popitam[133] and MultiTag[134] are also based upon a tag approach.

In their excellent review, Hernandez and co-workers[133] discus the shortcomings and drawbacks of de novo sequencing methodologies, and the reader is referred to their detailed work on this subject.

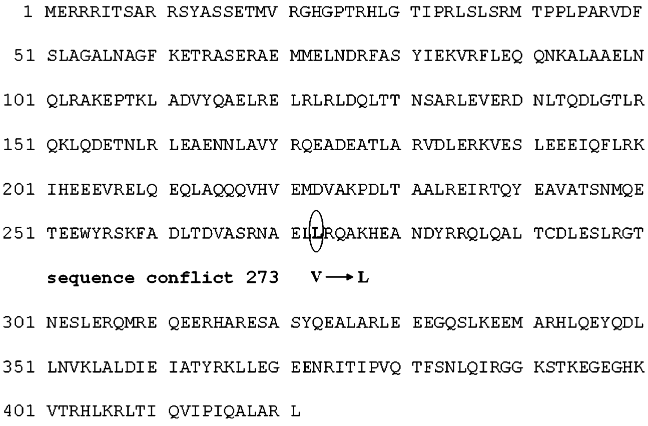
According to Wielsch et al.,[135] de novo sequencing hardly delivers the required accuracy and confidence of produced sequences.[136,137] However, when combined with sequence similarity searching tools, it provided an independent interpretation of MS/MS spectra: The method is considered inherently limited by the inability to produce meaningful sequence candidates from tandem mass spectra either with insufficient fragment presentation or having too complex fragments.

Grossmann and co-workers[138] suggest that the performance of all de novo sequencing software tools inevitably suffers from inherent limitations of MS/MS spectra analysis, making reliable automated de novo sequencing difficult. Mass accuracy, incomplete b- and y-ion series, and chemical noise are major confounding factors for this technique.

4.3. Database Errors—A Major Unsolved Problem for PI and PrI

Technically, peptide and protein analysis is fairly developed. However, only a small percentage of known proteins has been studied experimentally, and work is now hampered by the many errors in databases. Both academia and the biotechnology industry are suffering from this problem,[139] although databases are now actively working on improvement, last but not least, by cooperation with scientists performing protein analyses.[140] It is not only wrong sequences in databases that make scientific life difficult and sometimes unreliable; there are also many other types of entries that are misleading at best.

Like an infection, mistakes can spread through several databases and generate a lot of confusion.[141−144] It is currently not known how extensive database errors are, but a rough estimate would come to a certain percentage[145,146] and systematic comparisons of database annotations have been carried out.[147,148] Although some databases have a high standard owing to their high level of manual curation, a legion of scientists would be required to clean databases and an even stronger interaction is mandatory for a tight cooperation between experimentators and the databases, with responsibility for correctness of data for both. A major problem is the historical data in databases obtained when sequencing technology was not yet as developed as nowadays, when at the nucleic acid level error rates are as low as 1 base in 10000. According to Hadley,[139] the research that will be mainly affected by database errors is probably large-scale/high-throughput studies using large portions of sequence databases (Figure 1).



```
  1 MERRRITSAR RSYASSETMV RGHGPTRHLG TIPRLSLSRM TPPLPARVDF

 51 SLAGALNAGF KETRASERAE MMELNDRFAS YIEKVRFLEQ QNKALAAELN

101 QLRAKEPTKL ADVYQAELRE LRLRLDQLTT NSARLEVERD NLTQDLGTLR

151 QKLQDETNLR LEAENNLAVY RQEADEATLA RVDLERKVES LEEEIQFLRK

201 IHEEEVRELQ EQLAQQQVHV EMDVAKPDLT AALREIRTQY EAVATSNMQE

251 TEEWYRSKFA DLTDVASRNA ELLRQAKHEA NDYRRQLQAL TCDLESLRGT

    sequence conflict 273   V ⟶ L

301 NESLERQMRE QEERHARESA SYQEALARLE EEGQSLKEEM ARHLQEYQDL

351 LNVKLALDIE IATYRKLLEG EENRITIPVQ TFSNLQIRGG KSTKEGEGHK

401 VTRHLKRLTI QVIPIQALAR L
```

Figure 1. Demonstration of a database error in the rat glial fibrillary acidic protein sequence (amino acid position 273) based on conflicting reports from the Swiss-Prot and PIR (Protein Information Resource) databases. Based on Q-TOF analysis, we could validate leucine (273) and reject the report from the PIR database based on mRNA information on GFAP.

The establishment of the quality control task group CODATA[149] may assist in improving systems, but again, only an interaction between sequencing laboratories and databases may solve the problem in the long run. Scientists should be challenged to submit data directly to the databases interactively, and sequence conflicts and data on PTMs ought to be submitted in addition to publication in scientific journals. Ideally, only experimentally verified information, either by direct data submission or the literature, should be included in databases, and UniProt is following this strategy.[150] Still, most information on sequences in databases is inferred by similarity with previously analyzed entities,[151] and this again introduces errors, probably at a large scale.[146,148,152−154] Consensus-based approaches may assist in improving database systems proposing the detection of inconsistencies in the annotation of related proteins forming sequence clusters.[155,156] A series of other solutions to the problem is offered through knowledge discovery techniques based upon rules, anomalies, and common patterns.[157−161] Kretschmann et al.[160] used a novel approach to the problem by the use of automatic learning of rules from a highly curated database and subsequently using them for improving databases.

While knowledge is increasing, databases are working on improvements, and scientists have learned to be cautious when consulting databases, wrong annotations in proteomics still do occur and peptide and protein identifications are not appropriately controlled.

The consequence of erroneous databases for analytical scientists is that even higher sequence coverages must be generated, either by the instrument or even by time and money consuming digestions of proteins using several enzymes, etc.

5. Validation of Peptide and Protein Identification

Validation of peptide and protein identification is an obligatory step, and information on validation is not always provided in the literature or it was not carried out.

5.1. Scoring for MS/MS Peptide Identification

Scoring systems of the corresponding tandem MS identification software packages, including MASCOT, SEQUEST, etc., show remarkable limitations of PI.[97,99,135] One problem is the dependence on databases, and indeed, databases may be changing from day to day, thus resulting in biased scoring. Low quality MS/MS results may give high scores by chance, and therefore, even fair identifications should be validated by search-engine independent techniques.[162−166] On the other hand, identifications with low scores are classified as nonidentified and abandoned, with the reason for the low scores being, however, simply poor MS/MS data quality or that the structure is merely not in the database. This is particularly a problem in automated PI, and it is therefore mandatory that manual re-evaluation takes place. The M(ascot)-score, for instance, may thus be complemented by database-independent scoring[86] such as, e.g., the S-score.[162]

Kapp et al.[167] used PeptideProphet, a rescoring algorithm, to increase the performance of the SEQUEST algorithm and to indicate predictable false positive error rates by introduction of “consensus scoring”, the use of multiple (at least two) search algorithms to decrease the rate of false positive results and enable cross-validation of results. Determination of the score threshold for peptide identification based on a reverse sequence database search and calculating scores for each experiment[167] is proposed. Colinge et al.[107] introduced OLAV-scoring based on signal detection theory (imagination of the score as a signal emitted by every match) to better discriminate between true and false positive results as compared to the MASCOT database.[107] Keller et al. introduced different filtering strategies for SEQUEST results[164] combined with the statistical model of the expectation maximization algorithm[168] to distinguish correct from incorrect peptide assignments of MS/MS spectra. Anderson et al.[169] demonstrated that support vector machines (SVM) could improve the outcome from ion trap spectra searches against the SEQUEST algorithm. Ulintz et al. used a machine learning algorithm as a new scoring measure to improve the specificity of peptide identification of MS data.[170] All these approaches were introduced, as scoring may need improvement and, eventually, rescoring may be needed.

5.2. Use of Randomized Databases for the Validation of PI

The use of randomized sequence databases (or reversed or nonsense databases), in particular for coping with false positive results, is an important approach for the validation of PI and PrI, and a series of probability models of protein sequences have been summarized in several publications.[171]

The application of reversed and reshuffled sequences dates back to the 1980s, and they were just incorporated into modern searches a few years ago.[172−176] It became obvious how many protein sequences were matched to these control nonsense databases, thus showing a major mathematical/statistical inherent error of PI and PrI. A concomitant database search between a standard database and a nonsense database is recommended and may improve PI and PrI, but it has not been documented whether the use of nonsense databases does improve the reliability of data,[4] as, in addition to analytical steps, new errors may have been introduced by the use of these methods.[177] Rejtar et al.,[176] however, estimated the increase of false positive identifications by the use of a randomized database to be from 2.7% to 3.9%. Cargile and co-workers[4] report on the significant potential for false positive identifications from large databases using tandem MS: they searched a dataset against an in silico generated random protein database and generated a significant number of positive matches despite the use of filtering criteria. An example for the use of a reversed database along with molecular radius is provided by Park et al.[178] using a simple organism, pseudomonas putida: The proteome was filtered by a reversed sequence database search and correlated by molecular weight obtained at 1-DE. Their “decoy“ approach uses a reverse sequence database, i.e., translated ORFs in reverse orientation,[179] and the authors claimed higher confidence of protein identification. This method is, however, limited by the fact that proteins may show unpredictable and nonanalyzed post-translational or chemical modifications that may well influence the mobility in a 1-DE gel.

Shadforth and co-workers[180] produced an innovative scoring strategy combining the average peptide score method[181] with prefiltering of peptide identifications and the use of a reversed database. The authors claim that the threshold of identification, mainly set at 95%, may be therefore set to zero and is reducing false positive identifications; this is at the expense of losing some correct identifications, but this is a fair price to pay, in our opinion.[167] Qian et al.[175]estimated the rate of false positive results from MS/MS peptide identification from three different sets of human samples both by an independent search against SEQUEST, a reversed human protein sequence database IPI, and by respecting experimental factors. Protein identification was different between the three datasets, indicating that factors such as sample complexity and sample preparation can seriously affect the rate of false positive results per se.

6. Limitations of PI by Protein Properties

6.1. Short and Very Short Proteins

In proteomic practice, short and very short proteins are assigned to a gray zone overlapping with "peptidomics", as if peptides were structurally that different from proteins.[182]

As stated by Frith and co-workers,[183] current catalogues of mammalian proteins exhibit an artifactual discontinuity at a length of about 100 amino acids. The authors identify proteins in the FANTOM collection of mouse cDNAs by analyzing synonymous and nonsynonymous substitutions, confirming that there is no such discontinuity: they propose that about 10% of mouse proteins are shorter than 100 amino acids, although the majority of these are variants or fragments of longer proteins.[184,185] The problem of incorrect ORF annotations can be extrapolated to proteins,[186] and indeed, annotation or even homology searches are not reliable per se.[187]

In principle, identification of short structures by mass spectrometry can be performed in analogy to PI,[188,189] and the major differences are prefractionation (often chromatographical in nature)[190]/preseparation (2-DE is not the method of choice for low molecular weight structures, although approaches to overcome this limitation were published),[191] bioinformatic tools, and databases.

In principle, fragmentation spectra are difficult to interpret and MS/MS searches are only successful when the mass of the peptide fragmented and the sequence are identical.

This is a very limiting factor, and in the study by Clynen et al.,[192] many ion peaks remain unidentified. Svensson and co-workers[193] studied peptides from hypothalamic extracts by MS/MS, bypassing the problem of the use of protein or peptide identification databases. The authors sequenced peptides and in a nonsophisticated way submitted sequences to a basic local alignment search tool (BLAST), thus identifying known peptides and proposing new sequences.

The major problem in peptide identification will be solved when specific and large databases based upon MS/MS data will be available. An approach to solve the existing limitation of current low molecular weight databases was introduced by Falth et al.[194]

This SwePep database is specifically designed for endogenous peptides generated from precursor proteins and is limited to a molecular weight < 10 kDa. No specific databases for identification of very low and low molecular weight proteins per se are available, and mass analysis peptide sequence prediction (MAPSP) is at the prediction level.[195]

6.2. Hydrophobic Proteins

The analysis of hydrophobic proteins is a challenge to proteomics methods in particular, as most pharmaceutical targets are hydrophobic in nature.[196] First of all, we have to clarify that not all hydrophobic proteins are membrane proteins and vice versa, and hydrophobicity, expressed by positive values of the GRAVY index,[197] is not equal to high insolubility and vice versa.[198] Once brought into solution (sample preparation is not discussed herein), hydrophobic proteins are undergoing proteolytic cleavage or cleavage by cyanogen bromide and are identified mainly via their hydrophilic peptides[199] (Figure 2).

1 MLELLPTAVE GVSQAQITGR PEWIWLALGT ALMGLGTLYF LVKGMGVSDP

51 DAKKFYAITT LVPAIAFTMY LSMLLGYGLT MVPFGGEQNP IYWARYADWL

101 FTTPLLLLDL ALLVDADQGT ILALVGADGI MIGTGLVGAL TKVYSYRFVW

151 WAISTAAMLY ILYVLFFGFT SK**AESMRPEV ASTFK**VLRNV TVVLWSAYPV

201 VWLIGSEGAG IVPLNIETLL FMVLDVSAKV **GFGLILLR**SR **AIFGEAEAPE**

251 **PSAGDGAAAT SD**

Figure 2. Identification of the highly hydrophobic and short protein bacteriorhodopsin. Identification is based upon the hydrophilic peptides of the protein by MALDI-TOF/TOF-MS/MS (Ultraflex, Bruker Daltonics, Bremen, Germany) analyses.

The presence of hydrophobic epitopes, however, leads to generation of only a few peptides, and missed cleavages and nonspecific cleavages at different sites are seriously hampering PI. The amount of hydrophobic peptides extracted following tryptic digestion, however, can be increased by the use of a cycloalkyl aliphatic saccharide[200] and in situ liquid−liquid extraction.[201] Also multienzyme digestion using several proteases, such as chymotrypsin, LysC, and AspN, to name a few, leads to higher sequence coverages in our laboratory. The use of nonspecific proteases, such as, e.g., proteinase K, subtilisin, or CNBr,[196] warrants approaches different from those of identifying hydrophilic proteins, and in general, these cannot be used to analyze hydrophobic structures, as, e.g., different matrices have to be used in MS, etc. Using CNBr split products for identification, one has to take into account that methionine is modified to homoserine.[202] Furthermore, unstable intermediates may be formed by the use of CNBr.[203] Limited acid hydrolysis of hydrophobic proteins may represent a nonsophisticated approach to generation of many hydrophobic and hydrophilic peptides with the great advantage that less chemicals are introduced into the system.[204] It is not known, however, how many peptides are lost or are cleaved down to single amino acids. Moreover, ionization of hydrophobic peptides in MALDI-systems can be poor and residual SDS may hamper analysis as well.[205] The use of atmospheric pressure photoionization−mass spectrometry proposes a solution to the ionization problem of hydrophobic, apolar peptides, but experience with this technique is limited yet.[206] Difficulties in eluting hydrophobic proteins from LC columns are commonly encountered problems, and there are no validated standard protocols, although the problem has been addressed by several authors.[207−209]

Of course, almost all hydrophobic peptides can be finally analyzed by MS methodologies, but different analytical approaches have to be used in most cases.

6.3. The Isobaric Amino Acids Problem

Although Kassel and Biemann[210] used tandem mass spectrometrical differentiation between hydroxyproline isobars and isomers as early as in 1990, no high-throughput technique to discriminate isobaric amino acids has been described. A protonated peptide of interest is individually mass selected by using the first mass spectrometer (MS-1) and is introduced into a collision cell region where it undergoes collision-induced decomposition in a neutral gas, such as, e.g., argon or helium. Fragments formed in this atmosphere are then separated and analyzed by a second mass

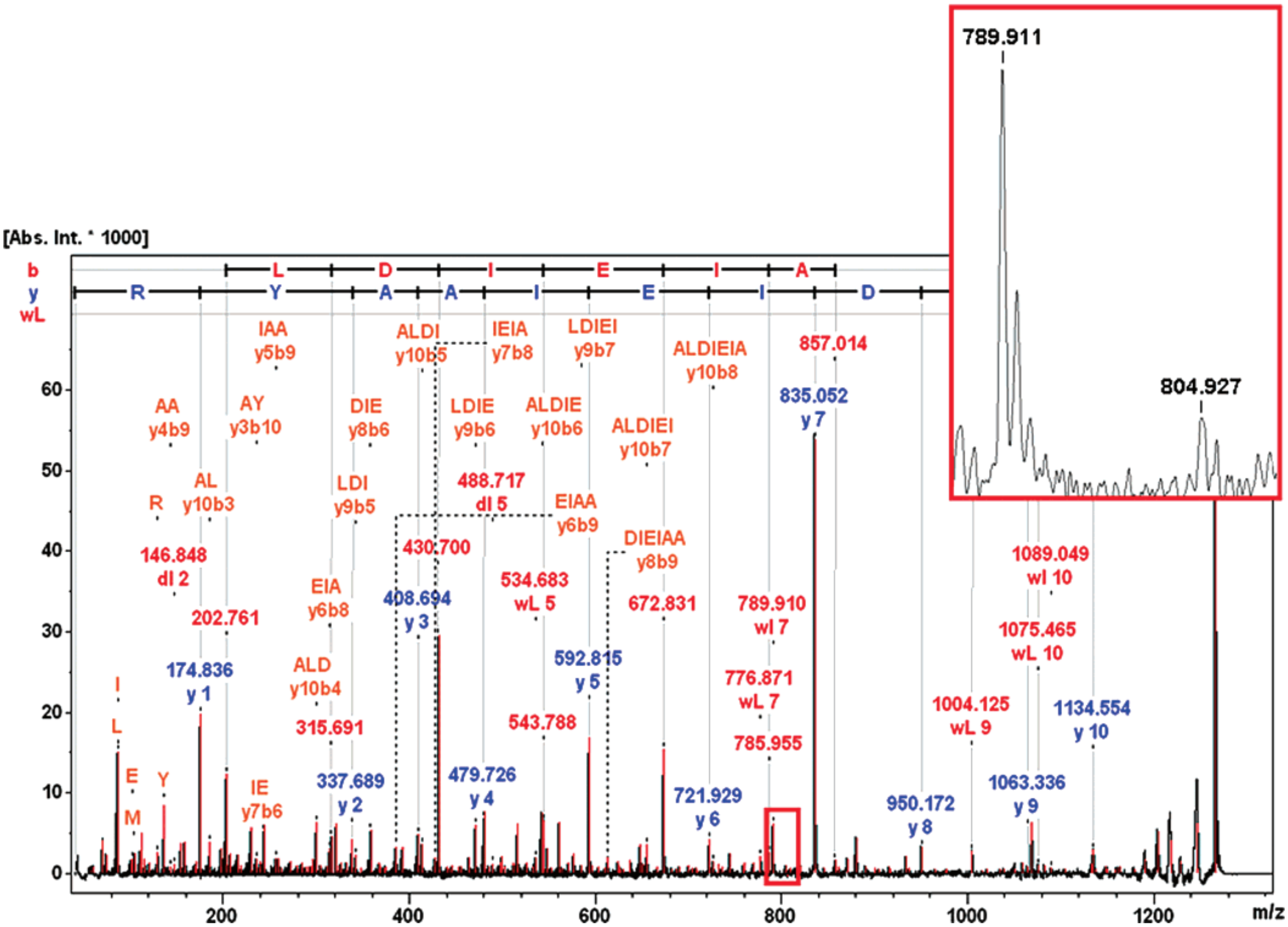


Figure 3. Differentiation between leucine/isoleucine using high-energy collision (CID) mode MS/MS analyses. Argon was used as collision gas for generating tryptic peptides which create additional high-energy w-type ions. The CID spectrum confirmed identification of isoleucine at amino acid number 384 due to diagostic w-ions at *m*/*z* 789.911 and 804.927 in rat neurofilament triplet L protein.

spectrometer (MS-2). Hydroxyproline, e.g., has the same residual mass of 113 as leucine and isoleucine,[211] and the normal sequence ions will therefore not be able to discriminate these three amino acids. Therefore, side-chain sequence ions, such as, e.g., d_n or w_n, have to be generated to discriminate between isoleucine and leucine, particularly by producing the immonium ion[212−214] and the 3- and 4-hydroxyproline isomers (Figure 3).

However, these sequence ions, in analogy to d-series ions, are not always generated.[215] Low-energy ESI-Trap MS(n) is another nonsophisticated technique to cope with the leucine−isoleucine difficulties.[216] Another way out may be consecutive reaction mass spectrometry[217] or hot electron capture dissociation in Fourier transform ion cyclotron resonance mass spectrometry, methods that are, however, highly time consuming and not readily available.[218]

6.4. The Problem of Low Complexity Regions of Proteins

Single amino acid repeats comprise a single homopolymeric tract of a particular amino acid. Uncontrolled genetic expansions of such stretches are linked to a series of human diseases, disorders such as, e.g., poly-Q and poly-A tracts,[219] although long repeats, such as serine or proline, are known to occur physiologically in certain specific brain proteins, including synapsins or poly-A tracts in a fibroin from the araneoid egg case silk[220] and polyproline in proline-rich histone H1 in the skin mucus of the atlantic salmon[221] (Figure 4).

```
  1 MMNFLRRRLS DSSFIANLPN GYMTDLQRPE PQQPPPAPGP GTATASAATS

 51 AASPGPERRP PPAQAPAPQP APQPAPTPSV GSSFFSSLSQ AVKQTAASAG

101 LVDAPAPSAA SRKAKVLLVV DEPHTDWAKC FRGKKILGDY DIKVEQAEFS

151 ELNLVAHADG TYAVDMQVLR NGTKVVRSFR PDFVLIRQHA FGMAENEDFR

201 HLVIGMQYAG LPSINSLESI YNFCDKPWVF AQMVAIFKTL GGEKFPLIEQ

251 TYYPNHREML TLPTFPVVVK IGHAHSGMGK VKVENHYDFQ DIASVVALTQ

301 TYATAEPFID AKYDIRVQKI GNNYKAYMRT SISGNWKTNT GSAMLEQIAM

351 SDRYKLWVDA CSEMFGGLDI CAVKAVHGKD GKDYIFEVMD CSMPLIGEHQ

401 VEDRQLITDL VISKMNQLLS RTPALSPQRP LTTQQPQSGT LKEPDSSKTP

451 PQRPAPQGGP GQPQGMQPPG KVLPPRRLPS GPSLPPSSSS SSSSSSSSSA

501 PQRPGGPTST QVNASSSSNS LAEPQAPQAA PPQKPQPHPQ LNKSQSLTNA

551 FSFSESSFFR SSANEDEAKA ETIRSLRKSF ASLFSD
```

Figure 4. Demonstration of the rat synapsin II protein sequence with a polyserine stretch at 486−499. By MALDI-TOF/TOF[71] the polyserine-stretch remained unidentified.

Although Edman degradation of the low complexity region containing proteins would lead to satisfying results, no high-throughput performance is possible and the amounts of proteins required would not be available. Therefore, it is of importance to work on the development of mass spectrometry methods for the identification of homopolymeric tracts. In the experience of our laboratory, the main problem is the appropriate cleavage. Often, even multienzymatic cleavage does not lead to satisfactory cleavage; therefore, mass spectrometrical analysis of a single amino acid repeat fails relatively often. The presence of (poly)prolines in the sequence causes the so-called "proline effect", characterized by a labile amide bond on the N-terminal side of P and a stable amide bond on its C-terminal side, with the consequence that tryptic cleavage is modified[222] and that the presence of P significantly affects peptide fragmentation with only a minor cleavage at the C-terminal side of the residue. Cleavage N-terminally to P regularly leads to formation of a Y′′ and Y doublet with Y being deficient in two H atoms.[214] Tryptic digestion of proline-rich proteins theoretically produces either very small peptides or very large hydrophobic peptides, and unusual cleavage sites have been described for these proteins before.[223] This, in turn, requires nonenzymatic SEQUEST searches where all possible cleavage sites have to be respected. These problems exist for other single amino acid repeats containing proteins and represent a shortcoming of PI.

6.5. Hypothetical Proteins, Proteins with Unknown Function, and Unknown Proteins

Again, the glossary has to be recalled, and to make it more complicated, there are also predicted/hypothetical ORFs.[224] Hypothetical proteins (HPs) are proteins predicted from nucleic acid sequences based upon low identity to characterized structures and that have not been shown to exist at the protein level.[197,225] Proteins with unknown function may be fully sequenced proteins without a known functional domain or domain(s) of unknown function (DUFs). Unknown proteins are those where no corresponding nucleic acid sequence (no corresponding ORF) is available.

A major part of unassigned proteins belongs to HPs and proteins with unknown function. Giometti et al.[226] reported that around 32% of identified *Methanococcus jannaschii* belongs to HPs. Li et al.,[227] studying the methanosarcina acetivorans proteome, identified 412 proteins, representing nearly 10% of the ORFs and containing approximately 30% of HPs. Vanden Wymelenberg et al.[228] have shown that 43% of identified structures showed no similarity to known proteins. The problem is aggravated by the estimate that 10−30% of ORFs do not actually encode proteins.[229]

A recent survey of 120 genomes showed that one out of three proteins in the NCBI database is annotated as hypothetical,[230] highlighting the challenge for proteomic analysis of HPs.

One important problem in the analysis of HPs is that ideally the full sequence has to be determined. If the sequence of the HP is still not showing high identity to a known protein, this structure can be considered a hypothetical or unknown or novel protein.

This means that MS/MS techniques and de novo sequencing are forming the basis for studies on HPs[231] and have to be extended if no full sequences are analyzed.

Edman degradation may be necessary to complement MS data.[232]

Extension of techniques may as well include the use of the accurate mass tags method respecting accurate masses and times of elution,[233−235] and these are reported to lead to an enormous increase of annotations of HPs in proteomes. The use of genomic technology, such as, e.g., the generation of recombinant (hypothetical) proteins, is mandatory in a certain percentage to reliably identify an HP.[236] The combined use of genomic and proteomic information is reflected by work from Fermin and co-workers:[237] they described novel gene and gene model detection using whole genome ORF analysis based upon Poisson statistics. Confidence of identification is assessed by estimating the significance of multipeptide identifications incorporating the length of the matching sequence, the number of spectra searched, and the size of the target sequence database.

If a sequence of interest is not present in any database, peptides can be deduced by de novo interpretation of MS/MS and used for designing degenerate oligonucleotide probes.[238] Peptide sequences from MS/MS analysis may also be used for protein identification by sequence similarity searches,[231,239] although MS results and sequence similarity searches are not easy to combined. Moreover, BLAST and FASTA are designed for alignments of sequences longer than peptide sequences obtained from MS/MS, raising the problem that hits may be statistically invalid. MS BLAST may solve the problem, as the scoring matrix was optimized for MS/MS-derived peptides. In addition, peptide sequences obtained from different instruments can be imported. Habermann and co-workers[240] described limitations of cross-species protein identification by MS-driven sequence similarity searches, and this publication is strongly recommended to scientists in the proteomic area with an interest in novel HPs.

The deluge of sequences generated by MS is challenging databases and development of programs.[241] Predictome, for example, a database of putative functional links between proteins, is a good representative of such a tool to cope with the bulk of data from a manifold of 44 genomes and forms a basis for protein identification of HPs and unknown proteins. All these systems represent, however, only machines generating hypotheses again, however valuable they are.

6.6. Protein Modifications: Artifacts and Post-translational Modifications

Modifications of proteins account for a large series of nonidentified proteins. While the known modifications can be readily determined, unsuspected modifications can be analyzed by a significant extra workload. Changes of molecular masses may reflect protein modifications. Chemical modifications of amino acids (artifacts) occur during all steps of sample preparation and sample processing until spotting onto the target, and they represent a confounding factor and pitfall for PI—very often, these artifacts cannot even be discriminated from PTMs without an enormous workload. The same modification may be generated by chemicals or procedures of the analytical method/sample processing, or they are genetically determined PTMs including pyro-glutamate, methylations, and deamidations. A comprehensive list of frequent modifications, both natural and artificial, is provided in the UNIMOD database (http://www.unimod.org/).[112] This contribution lists accurate and verifiable values for mass differences derived from elemental compositions. Other sources for modifications (in addition to the Mascot modification list) are Delta-Mass (http://www.abrf.org/index.cfm/dm.home), PIR-RESID,[242] FindMod (http://www.expasy.ch/tools/findmod/findmod_masses.html), ProSight PTM,[243] ModifiComb,[116] and MOD^i,[115] to name a few.

An inherent problem in the determination of PTMs is that many are escaping analysis due to the fact that they are lost during sample preparation (such as, e.g., dephosphorylation by the presence of phosphatases in the samples) or in the mass spectrometer itself. The existence of labile modifications such as, for example, sulfation and some forms of glycosylation, is hampering detection of PTMs in addition.[244] A series of modifications are highly heterogeneous, and further enzymatic studies are required to cope with this problem.

6.6.1. Chemical Modifications

Chemical modifications are considered a main confounding factor limiting PI, and from our own experience, they account for a significant part of nonidentified structures (Table 2). This problem can be only partially overcome by filtering the expected modifications that would be produced, such as, for example, those resulting from derivatization, from acrylamide—propionamide adducts of cysteine, or from polyacrylamide gel electrophoresis,[245] oxidations, or methyl esterification of glutamic or aspartic acid.[246] Oxidation of thiol-containing amino acids may be generated by the presence of residual persulfate in the gel as well.[247] Klarskov et al.[248] reported a mass increase by addition of a single β-mercaptoethanol moiety to a free cysteine residue, etc. The number and nature of so far unknown chemical modifications remains so far elusive. It has not been studied, however, how many of these artifactual modifications do occur and how many may be searched for at the same time: for filtering PTMs, however, best results are provided when PTMs are searched individually in the same search.

6.6.2. Post-translational Modifications

Mann and Jensen[249] and Jensen[250] have written excellent outlines of the proteomic analysis of post-translational modifications. PTMs account for the vast diversity of proteins,[251] and no conclusions can be drawn from nucleic acid sequences or predictions. PTMs are responsible for protein functions ranging from activation (e.g., phosphorylation[252]) to inhibition of a protein, localization, metabolism/turnover (e.g., ubiquitination), interactions, cross-linking, and homing properties (sialidation), to name a few.[253,254] For the sake of correctness, there are also pre- and cotranslational modifications of a protein, but the limitations of the analysis remain identical.

There is no universal concept or strategy to identify PTMs, and all approaches have their limitations. Analysis of PTMs from two-dimensional gel electrophoresis has the advantage that some modifications can be prescreened by immunological techniques, such as, e.g., by the use of phosphotyrosine, phosphothreonine, or phosphoserine antibodies or samples that can be run in the absence and presence of specific enzymes known to remove a specific modification. The protein amount of a single spot is, however, very often low and does not allow generation of many MS/MS spectra; the use of the identical spot from several gels run in parallel is therefore needed. Enrichment of the protein to be studied by affinity based methods may be a way out[255] but is often time-consuming. Another problem is comigration of proteins in gels forming an apparently single spot, and this protein mixture is prone to errors.

A specific example representing a common pitfall in the determination of PTMs is lysine acetylation in proteins. Acetylation/deacetylation is recognized as a regulatory signal in many cellular processes and is thought to be fairly well analyzed by mass spectrometrical techniques, i.e., mainly by MS/MS. In order to determine which lysine residue of a peptide was acetylated by an acetyltransferase, fragmentation analysis was carried out. However, neither MALDI MS/MS nor ESI MS/MS was able to produce a consecutive ion series because of its sequence and amino acid composition leading to strong internal fragmentation and subsequent complete loss of the b- and y-ion series.[256] This means that the key element for site-specific analysis of acetylation, MS/MS, may not provide fair results. Although peptide cleavage analysis or peptide sequence analysis (Edman degradation) represents a way out of the problem, this example shows a significant limitation of the method.

On the other hand, in peptide cleavage analysis, nonspecific remodeling and fragmentation events may result in generation of spectra that are too complex to be interpreted and what is thought to be an easy determination of the PTM, protein acetylation by mass spectrometry, turns out to be solved only by peptide mutation analysis.

A pitfall in the interpretation of, for example, in vitro acetylation experiments is nonenzymatic cysteine acetylation, and often acetylation sites are mapped on the assumption that only lysines show acceptor function.[257] Another pitfall is given by the fact that proteolytic cleavage of a peptide may lead to generation of new acetylation sites.[258]

Another shortcoming of mass spectrometry screens for O—N-acetylglucosamine modification of proteins is reported by Chalkley and Burlingame,[259] indicating that only MS/MS focused studies can demonstrate the presence of this PTM. LC-MS alone would not identify the PTM, and it has to be taken into account that this modification as well as others may prevent proteolytic cleavage by Pro-C and, most probably, other proteases.

Protein nitration is a frequent modification and is caused by nitric oxide attack. Tyrosine nitration is a very well-known and documented PTM, but care has to be taken if no tyrosine nitration is detected. In our laboratory, we recently observed amino-tyrosine but no nitro-tyrosine in spinal cords from rats with spinal cord injury (manuscript in preparation). As all steps were carried out under reducing conditions, we interpret the result as chemical modification due to reduction of nitrotyrosine (unpublished results). On the other hand, protein S-nitrosylation may not be detectable because this modification is not stable enough under the conditions of SDS-polyacrylamide gel electrophoresis.[260]

Glycopeptides are often suppressed in the presence of their non-glycosylated counterparts, and the presence of sialic acid causes a metastable state and fragmentation.[261]

Multiple PTMs may generate very complicated MS/MS datasets that are difficult to interpret.[262]

Even modification of chirality has been described as a form of PTM.[263] Buczek and co-workers[264] have shown that D- instead of L-phenylalanine at position 46 in a toxic peptide can be observed and that this isomerization moderated biological activity. Moreover, the many so far unknown PTMs[265] may represent confounding factors and analytical shortcomings. It remains to be shown whether new methods such as electron capture dissociation, proposed to preserving PTMs[266] and electron-transfer dissociation,[267] can cope with the challenge of PTM analysis.

6.7. Splice Variants/Isoforms

74% of all human genes are alternatively spliced,[268] accounting for the molecular diversity of proteins. A drawback of proteomics is the fact that mass spectrometrical techniques mostly identify only a part of the protein sequence, and these parts may not be assigned to a specific splice variant (SV). Therefore, MS analysis of the complete sequence to identify a SV is time-consuming and costly, as only MS/MS can be used reliably. Analysis of a SV is of utmost importance[269] because individual SVs from a single protein may show different functions, as illustrated by expression of different SVs during different physiological states. Additional work at the nucleic acid level is often required when truncation is present. Many proteins are truncated as a PTM or during processing, and it is not always readily possible to decide whether a fragment represents a SV or a PTM. This is even becoming more complicated when under certain conditions, such as, for example, under hypoxia, proteins are cleaved by caspases or other proteases.[270] Some authors consult the M_r of the corresponding protein from 2-DE to evaluate the molecular weight differences, but this is not appropriate, as mobility in gels is not determined by the M_r only. Of course, a significant part of a SV is not even known and therefore cannot be readily identified or assigned.

A first screening method for the probable presence of isoforms was published recently by Alm and co-workers.[271] Mass spectra are matched against each other by the use of extracted mass peaks and hierarchical clustering. The outcome is presented in dendrograms in which isoforms (not only SVs) cluster together. An important step forward may be high-throughput alternative splicing evaluation by primer extension and MALDI technology.[272] The analysis of known or suspected SVs using PCR, primer extension, and MALDI-TOF uses reverse-transcribed mRNA amplification with primers surrounding the site of alternative splicing, followed by a primer extension reaction and MS of the primer extension products. This method also corrects potential pitfalls from proteins run on 2-DE (two-dimensional gel electrophoresis), where heteroduplexes formed from different SVs can produce false results. The authors used MALDI-TOF, and therefore, the method may be significantly upgraded by the use of MS/MS; in addition, the method is not capable of identifying unknown SVs.

Although all databases providing information on SVs are naturally incomplete, we refer to Stamm et al.,[273] who summarize corresponding database resources. Pevzner and co-workers[274] as well as Roth et al.[275] address the efficiency of database searches for identification of mutated and modified proteins based upon MS/MS analyses.

To address the problem of glossary misuse: isoforms and paralogs have to be—and can be—discriminated at the protein level.[276] Polymorphisms are a related problem and may account for misidentifications as well.[277] A lot of effort and resources are mandatory to cope with the sheer endless work to be done in the SV area.

7. Protein Identification

Mass spectrometry became the method of choice for protein identification (PrI) and characterization, although linking genomic and proteomic data is inevitable.[278,279] Data from MS analysis are used to identify peptides, and peptides are matched to proteins in databases.[99] The limitations of PI have been discussed above, and the shortcomings of PrI are due to related underlying reasons, as PrI mainly relies on bioinformatic tools.[280]

PrI, based upon peptide mass fingerprinting (PMF),[97] is a main concept but does not identify proteins unambiguously unless very high sequence coverage is obtained.[281,282] Many proteins are very much related, are members of protein families, and show large numbers of isoforms/splice variants or are present with truncated structures; finally, peptides can be assigned to several or to no proteins. PMF is not providing reliable results for low molecular weight proteins (peptides) and very high molecular weight proteins.

The major limitations for PMF are false positives, and more limitations are listed by Perkins and co-workers,[99] such as, for example, that the probability-based scoring algorithm provides a quantitative measure of the significance of a match but is based on certain assumptions. Moreover, duplicate mass values are due to the large mass error window. Finally, atypical sequence entries are hampering PrI.

Stead et al.[283] tried to cope with the limitations of PMF by introducing universal metrics for quality assessment of PrI by mass spectrometry. Three simple and universal metrics to describe different aspects of the PrI by mass spectrometry were developed: Hit ratio (HR) gives an indication of the signal/noise ratio in a mass spectrum, mass coverage (MC) measures the amount of matched protein sequence, and excess of limit-digested peptides (ELDP) reflects the completeness of the digestion preceding PMF.

Another approach consists of using MS/MS data from one or more peptides[284] or, alternatively, using mixed datasets from MS analysis along with physicochemical data, amino acid analysis, or de novo sequencing programs.[131]

As to MS/MS-based protein identification in analogy to PMF, a predicted fragment ion(s) from each peptide of a database sequence is calculated, and the calculated and the observed ion masses are compared and a score is assigned. The individual peptide scores are combined to calculate a score for protein identification. Therefore, the main problem of this type of PrI consists of production of the types of daughter ions that are fully dependent on instrumentation and the analytical procedure used.

From the mid-1990s, MS/MS spectra were matched against sequence tags predicted for all proteins of a database, i.e., short series of fragment ions that could be attributed to coherent sequences of amino acids corresponding to subsets of the predicted peptide as implemented in programs such as, e.g., Protein Prospector's MS tag. Recently, searches are based upon comparisons between the experimentally oberved fragment ions and all predicted fragments for all hypothetical peptides of the appropriate molecular mass as based upon fragmentation rules. Only MS/MS data from more than a single peptide would reliably and correctly identify a protein. In combination with high sequence coverage analyzed by MS/PMF, fair identification of a protein may be obtained also by a single MS/MS peptide. Several databases should be searched, and PrI becomes safer when a protein is identified in more databases; however, thus far, there is no perfect computational tool for quality control of published data.[3] Current protein identification is mainly based upon tandem mass spectrometry (MS/MS) combined with database searching, and the term "high confidence PrI" is abandoned. Stringent validation of data is therefore mandatory for fair PrI.

8. Conclusion

Although proteomics technologies are still holding center stage and are the most valuable tools, there are shortcomings, drawbacks, and pitfalls at all levels of analysis. A concise review on problems would fill books, and therefore, a selection had to be mentioned and many first class publications from these areas were not cited. Moreover, in some subdisciplines that were not addressed, useful reviews are available. It is shown herein that pitfalls in PI can be expected in sample preparation for MS, including spot picking, protein in-gel digestion, target-matrix selection, and contamination. Problems of PI by instrumentation range from calibration errors, use of inappropriate lasers, and even insufficient maintenance of the instruments. A basic problem is the selection of instrumentation, and in most cases two MS principles are required for unambiguous PI. A main confounding factor is, however, data processing and data mining: assessment of spectra quality is an issue to be considered, and PI by databases is a serious limitation due to errors and incompleteness of databases. We have addressed validation of PI, and the use of appropriate scoring systems and random database searches are proposed. Protein composition and properties per se are an inherent limitation of PI, and several strategies have to be applied to overcome the multitude of specific pitfalls, such as, for example, in the analysis of short, hydrophobic structures or those containing isobaric amino acids or low complexity regions. The analytical problem of hypothetical and/or unknown proteins is enormous, and the presence of chemical artifacts, PTMs, and splice variants increases protein diversity and variety almost logarithmically.

Peptide and protein identification is highly professional, and the scientific community has to work on fair existing technology and proteomic know-how and practice and has to provide new strategies coping with limitations in protein chemistry. In particular, strong and effective interactions with databases and bioinformatics[285] are mandatory. As indicated above, enormous efforts are necessary to cope with protein identification, and proteomics techniques, however valuable they are, are not fully developed.

9. Acknowledgments

We are highly indebted to Jae-Kyung Myung, Ph.D., Julius John Pradeep, M.Sc., and Weiqiang Chen, Ph.D., all from our laboratory, for analytical and technical assistance.

10. References

(1) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269.
(2) Kislinger, T.; Emili, A. *Expert Rev. Proteomics* **2005**, *2*, 27.
(3) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell Proteomics* **2004**, *3*, 531.
(4) Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3*, 1082.
(5) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850.
(6) Ossipova, E.; Fenyo, D.; Eriksson, J. *Proteomics* **2006**, *6*, 2079.
(7) Baldwin, M. A. *Mol. Cell. Proteomics* **2004**, *3*, 1.
(8) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. *Mol. Cell. Proteomics* **2006**, *5*, 787.
(9) Jahn, O.; Hesse, D.; Reinelt, M.; Kratzin, H. D. *Anal. Bioanal. Chem.* **2006**, *386*, 92.
(10) Nordhoff, E.; Egelhofer, V.; Giavalisco, P.; Eickhoff, H.; Horn, M.; Przewieslik, T.; Theiss, D.; Schneider, U.; Lehrach, H.; Gobom, J. *Electrophoresis* **2001**, *22*, 2844.
(11) Choudhary, G.; Wu, S. L.; Shieh, P.; Hancock, W. S. *J. Proteome Res.* **2003**, *2*, 59.
(12) Olsen, J. V.; Ong, S. E.; Mann, M. *Mol. Cell. Proteomics* **2004**, *3*, 608.
(13) Cagney, G.; Amiri, S.; Premawaradena, T.; Lindo, M.; Emili, A. *Proteome Sci.* **2003**, *1*, 5.
(14) Siepen, J. A.; Keevil, E. J.; Knight, D.; Hubbard, S. J. *J. Proteome Res.* **2007**, *6*, 399.
(15) Fischer, F.; Poetsch, A. *Proteome Sci.* **2006**, *2*, 2.
(16) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., III. *Nat. Biotechnol.* **2003**, *21*, 532.
(17) van Montfort, B. A.; Doeven, M. K.; Canas, B.; Veenhoff, L. M.; Poolman, B.; Robillard, G. T. *Biochim. Biophys. Acta* **2002**, *1555*, 111.
(18) Li, A.; Sowder, R. C.; Henderson, L. E.; Moore, S. P.; Garfinkel, D. J.; Fisher, R. *J. Anal. Chem.* **2001**, *73*, 5395.
(19) Gattiker, A.; Bienvenut, W. V.; Bairoch, A.; Gasteiger, E. *Proteomics* **2002**, *2*, 1435.
(20) Hara, S.; Rosenfeld, R.; Lu, H. S. *Anal. Biochem.* **1996**, *243*, 74.
(21) Keil, B. *Protein Seq. Data Anal.* **1987**, *1*, 13.
(22) Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 496.
(23) Konig, S.; Zeller, M.; Peter-Katalinic, J.; Roth, J.; Sorg, C.; Vogl, T. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 1180.
(24) Beck, H. C.; Nielsen, E. C.; Matthiesen, R.; Jensen, L. H.; Sehested, M.; Finn, P.; Grauslund, M.; Hansen, A. M.; Jensen, O. N. *Mol. Cell. Proteomics* **2006**, *5*, 1314.
(25) Matthiesen, R.; Trelle, M. B.; Hojrup, P.; Bunkenborg, J.; Jensen, O. N. *J. Proteome Res.* **2005**, *4*, 2338.
(26) Chung, T. L.; Hsiao, H. H.; Yeh, Y. Y.; Shia, H. L.; Chen, Y. L.; Liang, P. H.; Wang, A. H.; Khoo, K. H.; Shoei-Lung, L. S. *J. Biol. Chem.* **2004**, *279*, 39653.
(27) Ervin, L. A.; Ball, L. E.; Crouch, R. K.; Schey, K. L. *Invest. Ophthalmol. Vis. Sci.* **2005**, *46*, 627.
(28) Hogan, J. M.; Pitteri, S. J.; McLuckey, S. A. *Anal. Chem.* **2003**, *75*, 6509.
(29) Warren, E. N.; Jiang, J.; Parker, C. E.; Borchers, C. H. *Biotechniques* **2005**, Suppl, 7−11.
(30) Sellinger, O. Z.; Wolfson, M. F. *Biochim. Biophys. Acta.* **1991**, *1080*, 110.
(31) Vestal, M. L.; Campbell, J. M. *Methods Enzymol.* **2005**, *402*, 79.
(32) Tholey, A.; Heinzle, E. *Anal. Bioanal. Chem.* **2006**, *386*, 24.
(33) Gonnet, F.; Lemaitre, G.; Waksman, G.; Tortajada, J. *Proteome Sci.* **2003**, *1*, 2.
(34) Cohen, S. L.; Chait, B. T. *Anal. Chem.* **1996**, *68*, 31.
(35) Jensen, C.; Haebel, S.; Andersen, S. O.; Roepstorff, P. *Int. J. Mass Spectrom. Ion Processes* **1997**, *160*, 339.
(36) Lewis, J. K.; Wei, J.; Siuzdak, G. *Encyclopedia of Analytical Chemistry*; John Wiley & Sons Ltd: Hoboken, NJ, 2000; p 5880.
(37) Kussmann, M.; Lassing, U.; Sturmer, C. A.; Przybylski, M.; Roepstorff, P. *J. Mass Spectrom.* **1997**, *32*, 483.
(38) Land, C. M.; Kinsel, G. R. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 726.
(39) Yao, J.; Scott, J. R.; Young, M. K.; Wilkins, C. L. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 805.
(40) Zhu, Y. F.; Lee, K. L.; Tang, K.; Allman, S. L.; Taranenko, N. I.; Chen, C. H. *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1315.
(41) McComb, M. E.; Oleschuk, R. D.; Manley, D. M.; Donald, L.; Chow, A.; O'Neil, J. D.; Ens, W.; Standing, K. G.; Perreault, H. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1716.
(42) Hung, K. C.; Ding, H.; Guo, B. *Anal. Chem.* **1999**, *71*, 518.
(43) Schuerenberg, M.; Luebbert, C.; Eickhoff, H.; Kalkum, M.; Lehrach, H.; Nordhoff, E. *Anal. Chem.* **2000**, *72*, 3436.
(44) Redeby, T.; Roeraade, J.; Emmer, A. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 1161.
(45) Redeby, T.; Emmer, A. *Anal. Bioanal. Chem.* **2005**, *381*, 225.
(46) Kleno, T. G.; Andreasen, C. M.; Kjeldal, H. O.; Leonardsen, L. R.; Krogh, T. N.; Nielsen, P. F.; Sorensen, M. V.; Jensen, O. N. *Anal. Chem.* **2004**, *76*, 3576.
(47) Krutchinsky, A. N.; Chait, B. T. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 129.
(48) Parker, K. C.; Garrels, J. I.; Hines, W.; Butler, E. M.; McKee, A. H.; Patterson, D.; Martin, S. *Electrophoresis* **1998**, *19*, 1920.
(49) Ding, Q.; Xiao, L.; Xiong, S.; Jia, Y.; Que, H.; Guo, Y.; Liu, S. *Proteomics* **2003**, *3*, 1313.
(50) Barsnes, H.; Mikalsen, S. O.; Eidhammer, I. *BMC Bioinformatics* **2006**, *7*, 42.
(51) Schmidt, F.; Schmid, M.; Jungblut, P. R.; Mattow, J.; Facius, A.; Pleissner, K. P. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 943.
(52) Krah, A.; Schmidt, F.; Becher, D.; Schmid, M.; Albrecht, D.; Rack, A.; Buttner, K.; Jungblut, P. R. *Mol. Cell Proteomics* **2003**, *2*, 1271.
(53) Samuelsson, J.; Dalevi, D.; Levander, F.; Rognvaldsson, T. *Bioinformatics* **2004**, *20*, 3628.
(54) Gobom, J.; Mueller, M.; Egelhofer, V.; Theiss, D.; Lehrach, H.; Nordhoff, E. *Anal. Chem.* **2002**, *74*, 3915.
(55) Bantscheff, M.; Duempelfeld, B.; Kuster, B. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1892.
(56) Moskovets, E.; Chen, H. S.; Pashkova, A.; Rejtar, T.; Andreev, V.; Karger, B. L. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2177.
(57) Chamrad, D. C.; Koerting, G.; Gobom, J.; Thiele, H.; Klose, J.; Meyer, H. E.; Blueggel, M. *Anal. Bioanal. Chem.* **2003**, *376*, 1014.
(58) Levander, F.; Rognvaldsson, T.; Samuelsson, J.; James, P. *Proteomics* **2004**, *4*, 2594.

(59) Wolski, W. E.; Lalowski, M.; Jungblut, P.; Reinert, K. *BMC Bioinformatics* **2005**, *6*, 203.

(60) Wu, S.; Kaiser, N. K.; Meng, D.; Anderson, G. A.; Zhang, K.; Bruce, J. E. *J. Proteome Res.* **2005**, *4*, 1434.

(61) Holle, A.; Haase, A.; Kayser, M.; Hohndorf, J. *J. Mass Spectrom.* **2006**, *41*, 705.

(62) Peele, G. L.; Brent, D. A. *Anal. Chem.* **1977**, *49*, 674.

(63) Du, Y.; Meng, F.; Patrie, S. M.; Miller, L. M.; Kelleher, N. L. *J. Proteome Res.* **2004**, *3*, 801.

(64) 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*, 2010.

(65) Stapels, M. D.; Barofsky, D. F. *Anal. Chem.* **2004**, *76*, 5423.

(66) Heller, M.; Mattou, H.; Menzel, C.; Yao, X. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 704.

(67) Hansen, K. C.; Schmitt-Ulms, G.; Chalkley, R. J.; Hirsch, J.; Baldwin, M. A.; Burlingame, A. L. *Mol. Cell Proteomics* **2003**, *2*, 299.

(68) Baldwin, M. A.; Medzihradszky, K. F.; Lock, C. M.; Fisher, B.; Settineri, T. A.; Burlingame, A. L. *Anal. Chem.* **2001**, *73*, 1707.

(69) Medzihradszky, K. F.; Leffler, H.; Baldwin, M. A.; Burlingame, A. L. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 215.

(70) Jonscher, K. *Genomics Proteomics* **2003**, *3*, 31.

(71) Bodnar, W. M.; Blackburn, R. K.; Krise, J. M.; Moseley, M. A. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 971.

(72) Krutchinsky, A. N.; Zhang, W.; Chait, B. T. *J. Am. Soc. Mass. Spectrom.* **2000**, *11*, 493.

(73) Chen, W. Q.; Kang, S. U.; Lubec, G. *Nat. Protoc.* **2006**, *1*, 1446.

(74) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212.

(75) Lim, H.; Eng, J.; Yates, J. R., III; Tollaksen, S. L.; Giometti, C. S.; Holden, J. F.; Adams, M. W.; Reich, C. I.; Olsen, G. J.; Hays, L. G. J. *Am. Soc. Mass Spectrom.* **2003**, *14*, 957.

(76) Cech, N. B.; Enke, C. G. *Anal. Chem.* **2000**, *72*, 2717.

(77) Hop, C. E.; Bakhtiar, R. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1049.

(78) Sadeghi, M.; Olumee, Z.; Tang, X.; Vertes, A.; Jiang, Z. X.; Henderson, A. J.; Lee, H. S.; Prasad, C. R. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 393.

(79) Krause, E.; Wenschuh, H.; Jungblut, P. R. *Anal. Chem.* **1999**, *71*, 4160.

(80) Valero, M.; Giralt, E.; Andreu, D. *Peptides*; Mayflower Scientific: Kingswinford, 1996; p 855.

(81) Baldwin, M. A. *Methods Enzymol.* **2005**, *402*, 348.

(82) Glish, G. L.; Vachet, R. W. *Nat. Rev. Drug Discov.* **2003**, *2*, 140.

(83) Levin, D. S.; Vouros, P.; Miller, R. A.; Nazarov, E. G.; Morris, J. C. *Anal. Chem.* **2006**, *78*, 96.

(84) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36*, 849.

(85) Moore, R. E.; Licklider, L.; Schumann, D.; Lee, T. D. *Anal. Chem.* **1998**, *70*, 4879.

(86) Guzzetta, A. W.; Thakur, R. A.; Mylchreest, I. C. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2067.

(87) Wetterhall, M.; Nilsson, S.; Markides, K. E.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 239.

(88) Venable, J. D.; Yates, J. R., III. *Anal. Chem.* **2004**, *76*, 2928.

(89) Li, F.; Sun, W.; Gao, Y.; Wang, J. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 1655.

(90) Tabb, D. L.; Eng, J. K.; Yates, J. R., III. *Proteome Research: Mass Spectrometry*; Springer: Berlin, 2000; p 125.

(91) Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., III. *Bioinformatics* **2004**, *20* (Suppl 1), I49.

(92) Purvine, S.; Kolker, N.; Kolker, E. *OMICS* **2004**, *8*, 255.

(93) Hilario, M.; Kalousis, A.; Pellegrini, C.; Muller, M. *Mass Spectrom. Rev.* **2006**, *25*, 409.

(94) Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2006**, *22*, 2059.

(95) Wu, F. X.; Gagne, P.; Droit, A.; Poirier, G. G. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1199.

(96) Baczek, T.; Bucinski, A.; Ivanov, A. R.; Kaliszan, R. *Anal. Chem.* **2004**, *76*, 1726.

(97) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.

(98) Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1*, 211.

(99) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551.

(100) Shinkawa, T.; Taoka, M.; Yamauchi, Y.; Ichimura, T.; Kaji, H.; Takahashi, N.; Isobe, T. *J. Proteome Res.* **2005**, *4*, 1826.

(101) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871.

(102) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36.

(103) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466.

(104) Zhang, N.; Aebersold, R.; Schwikowski, B. *Proteomics* **2002**, *2*, 1406.

(105) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3*, 958.

(106) Duncan, D. T.; Craig, R.; Link, A. J. *J. Proteome Res.* **2005**, *4*, 1842.

(107) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. *Proteomics* **2003**, *3*, 1454.

(108) Wong, J. W.; Durante, C.; Cartwright, H. M. *Anal. Chem.* **2005**, *77*, 5655.

(109) Wong, J. W.; Cagney, G.; Cartwright, H. M. *Bioinformatics* **2005**, *21*, 2088.

(110) Yates, J. R., III; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557.

(111) Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85.

(112) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2004**, *4*, 1534.

(113) Lehmann, W. D.; Bohne, A.; von Der Lieth, C. W. *J. Mass Spectrom.* **2000**, *35*, 1335.

(114) Wilkins, M. R.; Gasteiger, E.; Gooley, A. A.; Herbert, B. R.; Molloy, M. P.; Binz, P. A.; Ou, K.; Sanchez, J. C.; Bairoch, A.; Williams, K. L.; Hochstrasser, D. F. *J. Mol. Biol.* **1999**, *289*, 645.

(115) Kim, S.; Na, S.; Sim, J. W.; Park, H.; Jeong, J.; Kim, H.; Seo, Y.; Seo, J.; Lee, K. J.; Paek, E. *Nucleic Acids Res.* **2006**, *34*, *W258*.

(116) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *Mol. Cell. Proteomics* **2006**, *5*, 935.

(117) Hernandez, P.; Muller, M.; Appel, R. D. *Mass Spectrom. Rev.* **2006**, *25*, 235.

(118) Shui, W.; Liu, Y.; Fan, H.; Bao, H.; Liang, S.; Yang, P.; Chen, X. *J. Proteome Res.* **2005**, *4*, 83.

(119) Heredia-Langner, A.; Cannon, W. R.; Jarman, K. D.; Jarman, K. H. *Bioinformatics* **2004**, *20*, 2296.

(120) Spengler, B. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 703.

(121) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337.

(122) Zidarov, D.; Thibault, P.; Evans, M. J.; Bertrand, M. J. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 13.

(123) Ishikawa, K.; Niwa, Y. *Biol. Mass Spectrom.* **1986**, *13*, 373.

(124) Yates, J. R., III; Griffin, P. R.; Hood, L. E.; Zhou, J. X. *Techniques in protein chemistry II*; Academic Press: San Diego, CA, 1991; p 477.

(125) Johnson, R. S.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945.

(126) Scarberry, R. E.; Zhang, Z.; Knapp, D. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 947.

(127) Bartels, C. *Biomed. Environ. Mass. Spectrom.* **1990**, *19*, 363.

(128) Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1991**, *3*, 326.

(129) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327.

(130) Fernandez-de-Cossio, J.; Gonzalez, J.; Besada, V. *Comput. Appl. Biosci.* **1995**, *11*, 427.

(131) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390.

(132) Tabb, D. L.; Saraf, A.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 6415.

(133) Hernandez, P.; Gras, R.; Frey, J.; Appel, R. D. *Proteomics* **2003**, *3*, 870.

(134) Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A.; Shevchenko, A. *Anal. Chem.* **2003**, *75*, 1307.

(135) Wielsch, N.; Thomas, H.; Surendranath, V.; Waridel, P.; Frank, A.; Pevzner, P.; Shevchenko, A. *J. Proteome Res.* **2006**, *5*, 2448.

(136) Standing, K. G. *Curr. Opin. Struct. Biol.* **2003**, *13*, 595.

(137) Shevchenko, A.; Loboda, A.; Shevchenko, A.; Ens, W.; Standing, K. G. *Anal. Chem.* **2000**, *72*, 2132.

(138) Grossmann, J.; Roos, F. F.; Cieliebak, M.; Liptak, Z.; Mathis, L. K.; Muller, M.; Gruissem, W.; Baginsky, S. *J. Proteome Res.* **2005**, *4*, 1768.

(139) Hadley, C. *EMBO Rep.* **2003**, *4*, 829.

(140) Orchard, S.; Hermjakob, H.; Apweiler, R. *Mol. Cell. Proteomics* **2005**, *4*, 435.

(141) Kyrpides, N. C.; Ouzounis, C. A. *Mol. Microbiol.* **1999**, *32*, 886.

(142) Gilks, W. R.; Audit, B.; De Angelis, D.; Tsoka, S.; Ouzounis, C. A. *Bioinformatics* **2002**, *18*, 1641.

(143) Learn, G. H., Jr.; Korber, B. T.; Foley, B.; Hahn, B. H.; Wolinsky, S. M.; Mullins, J. I. *J. Virol.* **1996**, *70*, 5720.

(144) Lesk, A. M.; Boswell, D. R.; Lesk, V. I.; Lesk, V. E.; Bairoch, A. *Protein Seq. Data Anal.* **1989**, *2*, 295.

(145) Brenner, S. E. *Trends Genet.* **1999**, *15*, 132.

(146) Galperin, M. Y.; Koonin, E. V. *In Silico Biol.* **1998**, *1*, 55.

(147) Devos, D.; Valencia, A. *Proteins* **2000**, *41*, 98.

(148) Wilson, C. A.; Kreychman, J.; Gerstein, M. *J. Mol. Biol.* **2000**, *297*, 233.

(149) CODATA Task Group on biological macromolecules and colleagues. Committee on Data for Science and Technology of the International Council of Scientific Unions. *Bioessays* **2000**, *22*, 1024.

(150) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. *Nucleic Acids Res.* **2005**, *33*, D154.
(151) Artamonova, I. I.; Frishman, G.; Gelfand, M. S.; Frishman, D. *Bioinformatics* **2005**, *21* (Suppl 3), iii49.
(152) Smith, R. F. *Genome Res.* **1996**, *6*, 653.
(153) Bork, P.; Bairoch, A. *Trends Genet.* **1996**, *12*, 425.
(154) Hegyi, H.; Gerstein, M. *Genome Res.* **2001**, *11*, 1632.
(155) Xie, H.; Wasserman, A.; Levine, Z.; Novik, A.; Grebinskiy, V.; Shoshan, A.; Mintz, L. *Genome Res.* **2002**, *12*, 785.
(156) Kaplan, N.; Vaaknin, A.; Linial, M. *Nucleic Acids Res.* **2003**, *31*, 5617.
(157) Hu, Z. Z.; Narayanaswamy, M.; Ravikumar, K. E.; Vijay-Shanker, K.; Wu, C. H. *Bioinformatics* **2005**, *21*, 2759.
(158) Michailidis, G.; Shedden, K. *J. Comput. Biol.* **2003**, *10*, 689.
(159) Eisenhaber, F.; Bork, P. *Bioinformatics* **1999**, *15*, 528.
(160) Kretschmann, E.; Fleischmann, W.; Apweiler, R. *Bioinformatics* **2001**, *17*, 920.
(161) Yu, G. X. *Bioinf. Comput. Biol.* **2004**, *2*, 615.
(162) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *Mol. Cell. Proteomics* **2005**, *4*, 1180.
(163) Olsen, J. V.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13417.
(164) Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *OMICS* **2002**, *6*, 207.
(165) Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768.
(166) Chen, Y.; Kwon, S. W.; Kim, S. C.; Zhao, Y. *J. Proteome Res.* **2005**, *4*, 998.
(167) Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. *Proteomics* **2005**, *5*, 3475.
(168) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383.
(169) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2*, 137.
(170) Ulintz, P. J.; Zhu, J.; Qin, Z. S.; Andrews, P. C. *Mol Cell Proteomics* **2006**, *5*, 497.
(171) Higdon, R.; Hogan, J. M.; Van Belle, G.; Kolker, E. *OMICS* **2005**, *9*, 364.
(172) Allet, N.; Barrillat, N.; Baussant, T.; Boiteau, C.; Botti, P.; Bougueleret, L.; Budin, N.; Canet, D.; Carraud, S.; Chiappe, D.; Christmann, N.; Colinge, J.; Cusin, I.; Dafflon, N.; Depresle, B.; Fasso, I.; Frauchiger, P.; Gaertner, H.; Gleizes, A.; Gonzalez-Couto, E.; Jeandenans, C.; Karmime, A.; Kowall, T.; Lagache, S.; Mahe, E.; Masselot, A.; Mattou, H.; Moniatte, M.; Niknejad, A.; Paolini, M.; Perret, F.; Pinaud, N.; Ranno, F.; Raimondi, S.; Reffas, S.; Regamey, P. O.; Rey, P. A.; Rodriguez-Tome, P.; Rose, K.; Rossellat, G.; Saudrais, C.; Schmidt, C.; Villain, M.; Zwahlen, C. *Proteomics* **2004**, *4*, 2333.
(173) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378.
(174) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43.
(175) Qian, W. J.; Liu, T.; Monroe, M. E.; Strittmatter, E. F.; Jacobs, J. M.; Kangas, L. J.; Petritis, K.; Camp, D. G., II; Smith, R. D. *J. Proteome Res.* **2005**, *4*, 53.
(176) Rejtar, T.; Chen, H. S.; Andreev, V.; Moskovets, E.; Karger, B. L. *Anal. Chem.* **2004**, *76*, 6017.
(177) Rudnick, P. A.; Wang, Y.; Evans, E.; Lee, C. S.; Balgley, B. M. *J. Proteome Res.* **2005**, *4*, 1353.
(178) Park, G. W.; Kwon, K. H.; Kim, J. Y.; Lee, J. H.; Yun, S. H.; Kim, S. I.; Park, Y. M.; Cho, S. Y.; Paik, Y. K.; Yoo, J. S. *Proteomics* **2006**, *6*, 1121.
(179) Karplus, K.; Barrett, C.; Hughey, R. *Bioinformatics* **1998**, *14*, 846.
(180) Shadforth, I.; Dunkley, T.; Lilley, K.; Crowther, D.; Bessant, C. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3363.
(181) Chepanoske, C. L.; Richardson, B. E.; von Rechenberg, M.; Peltier, J. M. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 9.
(182) Fricker, L. D.; Lim, J.; Pan, H. E.; Che, F. Y. *Mass Spectrom. Rev.* **2006**, *25*, 327.
(183) Frith, M. C.; Forrest, A. R.; Nourbakhsh, E.; Pang, K. C.; Kai, C.; Kawai, J.; Carninci, P.; Hayashizaki, Y.; Bailey, T. L.; Grimmond, S. M. *PLoS Genet.* **2006**, *2*, e52.
(184) Chou, K. C.; Shen, H. B. *J. Proteome Res.* **2006**, *5*, 1888.
(185) Hortin, G. L.; Jortani, S. A.; Ritchie, J. C., Jr.; Valdes, R., Jr.; Chan, D. W. *Clin. Chem.* **2006**, *52*, 1218.
(186) Linial, M. *Trends Biotechnol.* **2003**, *21*, 298.
(187) Bienkowska, J. R.; Hartman, H.; Smith, T. F. *Protein Eng.* **2003**, *16*, 897.
(188) Hummon, A. B.; Amare, A.; Sweedler, J. V. *Mass Spectrom. Rev.* **2006**, *25*, 77.
(189) Schoofs, L.; Baggerman, G. *Briefings Funct. Genomics Proteomics* **2003**, *2*, 114.
(190) Ramstrom, M.; Bergquist, J. *FEBS Lett.* **2004**, *567*, 92.
(191) Kitta, K.; Ohnishi-Kameyama, M.; Moriyama, T.; Ogawa, T.; Kawamoto, S. *Anal. Biochem.* **2006**, *351*, 290.
(192) Clynen, E.; Baggerman, G.; Veelaert, D.; Cerstiaens, A.; Van der Horst, D.; Harthoorn, L.; Derua, R.; Waelkens, E.; De Loof, A.; Schoofs, L. *Eur. J. Biochem.* **2001**, *268*, 1929.
(193) Svensson, M.; Skold, K.; Svenningsson, P.; Andren, P. E. *J. Proteome Res.* **2003**, *2*, 213.
(194) Falth, M.; Skold, K.; Norrman, M.; Svensson, M.; Fenyo, D.; Andren, P. E. *Mol. Cell. Proteomics* **2006**, *5*, 998.
(195) Eisenacher, M.; de Braaf, J.; Konig, S. *Bioinformatics* **2006**, *22*, 1002.
(196) Wu, C. C.; Yates, J. R., III. *Nat. Biotechnol.* **2003**, *21*, 262.
(197) Lubec, G.; Afjehi-Sadat, L.; Yang, J. W.; John, J. P. *Prog. Neurobiol.* **2005**, *77*, 90.
(198) Zhang, L.; Xie, J.; Wang, X.; Liu, X.; Tang, X.; Cao, R.; Hu, W.; Nie, S.; Fan, C.; Liang, S. *Proteomics* **2005**, *5*, 4510.
(199) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242.
(200) Katayama, H.; Tabata, T.; Ishihama, Y.; Sato, T.; Oda, Y.; Nagasu, T. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2388.
(201) Kjellstrom, S.; Jensen, O. N. *Anal. Chem.* **2003**, *75*, 2362.
(202) Chao, C. C.; Ma, Y. S.; Stadtman, E. R. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2969.
(203) Zhang, X.; Dillen, L.; Vanhoutte, K.; Van Dongen, W.; Esmans, E.; Claeys, M. *Anal. Chem.* **1996**, *68*, 3422.
(204) Zhong, H.; Zhang, Y.; Wen, Z.; Li, L. *Nat. Biotechnol.* **2004**, *22*, 1291.
(205) Zischka, H.; Gloeckner, C. J.; Klein, C.; Willmann, S.; Swiatek-de Lange, M.; Ueffing, M. *Proteomics* **2004**, *4*, 3776.
(206) Delobel, A.; Halgand, F.; Laffranchise-Gosse, B.; Snijders, H.; Laprevote, O. *Anal. Chem.* **2003**, *75*, 5961.
(207) Craft, D.; Li, L. *Anal. Chem.* **2005**, *77*, 2649.
(208) Hixson, K. K.; Rodriguez, N.; Camp, D. G., II; Strittmatter, E. F.; Lipton, M. S.; Smith, R. D. *Electrophoresis* **2002**, *23*, 3224.
(209) Speers, A. E.; Blackler, A. R.; Wu, C. C. *Anal. Chem.*, in press.
(210) Kassel, D. B.; Biemann, K. *Anal. Chem.* **1990**, *62*, 1691.
(211) Nachman, R. J.; Russell, W. K.; Coast, G. M.; Russell, D. H.; Predel, R. *Peptides* **2005**, *26*, 2151.
(212) Aubagnac, I. L.; El Amarani, B.; Devienne, F. M.; Conbarieu, R. *Org. Mass Spectrom.* **1985**, *20*, 428.
(213) Heerma, W.; Bathelt, E. R. *Biol. Mass Spectrom.* **1986**, *13*, 205.
(214) Papayannopoulos, I. A. *Mass Spectrom. Rev.* **1995**, *14*, 49.
(215) Johnson, R. S.; Martin, S. A.; Biemann, K.; Stults, J. T.; Watson, J. T. *Anal. Chem.* **1987**, *59*, 2621.
(216) Armirotti, A.; Millo, E.; Damonte, G. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 57.
(217) Tomer, K. B.; Guenat, C. R.; Deterding, L. J. *Anal. Chem.* **1988**, *60*, 2232.
(218) Kjeldsen, F.; Haselmann, K. F.; Sorensen, E. S.; Zubarev, R. A. *Anal. Chem.* **2003**, *75*, 1267.
(219) Faux, N. G.; Bottomley, S. P.; Lesk, A. M.; Irving, J. A.; Morrison, J. R.; de la Banda, M. G.; Whisstock, J. C. *Genome Res.* **2005**, *15*, 537.
(220) Hu, X.; Lawrence, B.; Kohler, K.; Falick, A. M.; Moore, A. M.; McMullen, E.; Jones, P. R.; Vierra, C. *Biochemistry* **2005**, *44*, 10020.
(221) Luders, T.; Birkemo, G. A.; Nissen-Meyer, J.; Andersen, O.; Nes, I. F. *Antimicrob. Agents Chemother.* **2005**, *49*, 2399.
(222) Wang, Y.; Johansson, J.; Griffiths, W. J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2182.
(223) Wilmarth, P. A.; Riviere, M. A.; Rustvold, D. L.; Lauten, J. D.; Madden, T. E.; David, L. L. *J. Proteome Res.* **2004**, *3*, 1017.
(224) Shaw, A. C.; Gevaert, K.; Demol, H.; Hoorelbeke, B.; Vandekerckhove, J.; Larsen, M. R.; Roepstorff, P.; Holm, A.; Christiansen, G.; Birkelund, S. *Proteomics* **2002**, *2*, 164.
(225) Lapidus, A.; Galleron, N.; Sorokin, A.; Ehrlich, S. D. *Microbiology* **1997**, *143*, 3431.
(226) Giometti, C. S.; Reich, C.; Tollaksen, S.; Babnigg, G.; Lim, H.; Zhu, W.; Yates, J.; Olsen, G. *J. Chromatogr., B* **2002**, *782*, 227.
(227) Li, Q.; Li, L.; Rejtar, T.; Karger, B. L.; Ferry, J. G. *J. Proteome Res.* **2005**, *4*, 112.
(228) Vanden Wymelenberg, A.; Minges, P.; Sabat, G.; Martinez, D.; Aerts, A.; Salamov, A.; Grigoriev, I.; Shapiro, H.; Putnam, N.; Belinky, P.; Dosoretz, C.; Gaskell, J.; Kersten, P.; Cullen, D. *Fungal Genet. Biol.* **2006**, *43*, 343.
(229) Ochman, H. *Trends Genet.* **2002**, *18*, 335.
(230) Kolker, E.; Makarova, K. S.; Shabalina, S.; Picone, A. F.; Purvine, S.; Holzman, T.; Cherny, T.; Armbruster, D.; Munson, R. S., Jr.; Kolesov, G.; Frishman, D.; Galperin, M. Y. *Nucleic Acids Res.* **2004**, *32*, 2353.
(231) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917.

(232) Johnson, R. S.; Walsh, K. A. *Protein Sci.* **1992**, *1*, 1083.
(233) Lipton, M. S.; Pasa-Tolic', L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic', N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049.
(234) Ding, Y. H.; Hixson, K. K.; Giometti, C. S.; Stanley, A.; Esteve-Nunez, A.; Khare, T.; Tollaksen, S. L.; Zhu, W.; Adkins, J. N.; Lipton, M. S.; Smith, R. D.; Mester, T.; Lovley, D. R. *Biochim. Biophys. Acta* **2006**, *1764*, 1198.
(235) Elias, D. A.; Monroe, M. E.; Marshall, M. J.; Romine, M. F.; Belieav, A. S.; Fredrickson, J. K.; Anderson, G. A.; Smith, R. D.; Lipton, M. S. *Proteomics* **2005**, *5*, 3120.
(236) Wang, X. R.; Zhou, Y. B.; Liu, F.; Wang, K. S.; Shen, Y.; Liu, J. H.; Han, Z. G. *Cell. Mol. Biol. Lett.* **2006**, *11*, 161.
(237) Fermin, D.; Allen, B. B.; Blackwell, T. W.; Menon, R.; Adamski, M.; Xu, Y.; Ulintz, P.; Omenn, G. S.; States, D. J. *Genome Biol.* **2006**, *7*, R35.
(238) Shevchenko, A.; Chernushevic, I.; Shevchenko, A.; Wilm, M.; Mann, M. *Mol. Biotechnol.* **2002**, *20*, 107.
(239) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594.
(240) Habermann, B.; Oegema, J.; Sunyaev, S.; Shevchenko, A. *Mol. Cell. Proteomics* **2004**, *3*, 238.
(241) McCarthy, F. M.; Cooksey, A. M.; Wang, N.; Bridges, S. M.; Pharr, G. T.; Burgess, S. C. *Proteomics* **2006**, *6*, 2759.
(242) Garavelli, J. S. *Nucleic Acids Res.* **2003**, *31*, 499.
(243) LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. *Nucleic Acids Res.* **2004**, *32*, W340.
(244) Wolfender, J. L.; Chu, F.; Ball, H.; Wolfender, F.; Fainzilber, M.; Baldwin, M. A.; Burlingame, A. L. *J. Mass Spectrom.* **1999**, *34*, 447.
(245) Bordini, E.; Hamdan, M.; Righetti, P. G. *Electrophoresis* **2000**, *21*, 2911.
(246) Haebel, S.; Albrecht, T.; Sparbier, K.; Walden, P.; Korner, R.; Steup, M. *Electrophoresis* **1998**, *19*, 679.
(247) Fantes, K. H.; Furminger, I. G. *Nature* **1967**, *216*, 71.
(248) Klarskov, K.; Roecklin, D.; Bouchon, B.; Sabatie, J.; Van Dorsselaer, A.; Bischoff, R. *Anal. Biochem.* **1994**, *216*, 127.
(249) Mann, M.; Jensen, O. N. *Nat. Biotechnol.* **2003**, *21*, 255.
(250) Jensen, O. N. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7*, 391.
(251) Baumann, M.; Meri, S. *Expert Rev. Proteomics* **2004**, *1*, 207.
(252) D'Ambrosio, C.; Salzano, A. M.; Arena, S.; Renzone, G.; Scaloni, A. *J. Chromatogr., B*, in press.
(253) Seet, B. T.; Dikic, I.; Zhou, M. M.; Pawson, T. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7*, 473.
(254) Farriol-Mathis, N.; Garavelli, J. S.; Boeckmann, B.; Duvaud, S.; Gasteiger, E.; Gateau, A.; Veuthey, A. L.; Bairoch, A. *Proteomics* **2004**, *4*, 1537.
(255) Stensballe, A.; Andersen, S.; Jensen, O. N. *Proteomics* **2001**, *1*, 207.
(256) Dormeyer, W.; Ott, M.; Schnolzer, M. *Mol. Cell. Proteomics* **2005**, *4*, 1226.
(257) Dormeyer, W.; Dorr, A.; Ott, M.; Schnolzer, M. *Anal. Bioanal. Chem.* **2003**, *376*, 994.
(258) Pasheva, E.; Sarov, M.; Bidjekov, K.; Ugrinova, I.; Sarg, B.; Lindner, H.; Pashev, I. G. *Biochemistry* **2004**, *43*, 2935.
(259) Chalkley, R. J.; Burlingame, A. L. *Mol. Cell. Proteomics* **2003**, *2*, 182.
(260) Patton, W. F. *J. Chromatogr., B* **2002**, *771*, 3.
(261) Seo, J.; Lee, K. J. *J. Biochem. Mol. Biol.* **2004**, *37*, 35.
(262) Larsen, M. R.; Trelle, M. B.; Thingholm, T. E.; Jensen, O. N. *Biotechniques* **2006**, *40*, 790.
(263) Kreil, G. *Annu. Rev. Biochem.* **1997**, *66*, 337.
(264) Buczek, O.; Yoshikami, D.; Bulaj, G.; Jimenez, E. C.; Olivera, B. M. *J. Biol. Chem.* **2005**, *280*, 4247.
(265) Hansen, B. T.; Davey, S. W.; Ham, A. J.; Liebler, D. C. *J. Proteome Res.* **2005**, *4*, 358.
(266) Zubarev, R. A. *Curr. Opin. Biotechnol.* **2004**, *15*, 12.
(267) Coon, J. J.; Ueberheide, B.; Syka, J. E.; Dryhurst, D. D.; Ausio, J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 9463.
(268) Johnson, J. M.; Castle, J.; Garrett-Engele, P.; Kan, Z.; Loerch, P. M.; Armour, C. D.; Santos, R.; Schadt, E. E.; Stoughton, R.; Shoemaker, D. D. *Science* **2003**, *302*, 2141.
(269) Godovac-Zimmermann, J.; Kleiner, O.; Brown, L. R.; Drukier, A. K. *Proteomics* **2005**, *5*, 699.
(270) Chen, G.; Gharib, T. G.; Huang, C. C.; Taylor, J. M.; Misek, D. E.; Kardia, S. L.; Giordano, T. J.; Iannettoni, M. D.; Orringer, M. B.; Hanash, S. M.; Beer, D. G. *Mol. Cell. Proteomics* **2002**, *1*, 304.
(271) Alm, R.; Johansson, P.; Hjerno, K.; Emanuelsson, C.; Ringner, M.; Hakkinen, J. *J. Proteome Res.* **2006**, *5*, 785.
(272) McCullough, R. M.; Cantor, C. R.; Ding, C. *Nucleic Acids Res.* **2005**, *33*, e99.
(273) Stamm, S.; Riethoven, J. J.; Le Texier, V.; Gopalakrishnan, C.; Kumanduri, V.; Tang, Y.; Barbosa-Morais, N. L.; Thanaraj, T. A. *Nucleic Acids Res.* **2006**, *34*, D46.
(274) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11*, 290.
(275) Roth, M. J.; Forbes, A. J.; Boyne, M. T., II; Kim, Y. B.; Robinson, D. E.; Kelleher, N. L. *Mol. Cell. Proteomics* **2005**, *4*, 1002.
(276) Spitzer, M.; Lorkowski, S.; Cullen, P.; Sczyrba, A.; Fuellen, G. *BMC Bioinf.* **2006**, *7*, 110.
(277) Wroblewski, M. S.; Wilson-Grady, J. T.; Martinez, M. B.; Kasthuri, R. S.; McMillan, K. R.; Flood-Urdangarin, C.; Nelsestuen, G. L. *FEBS J.* **2006**, *273*, 4707.
(278) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440.
(279) Yates, J. R., III. *Trends Genet.* **2000**, *16*, 5.
(280) Parker, K. C.; Patterson, D.; Williamson, B.; Marchese, J.; Graber, A.; He, F.; Jacobson, A.; Juhasz, P.; Martin, S. *Mol. Cell. Proteomics* **2004**, *3*, 625.
(281) Baldwin, M. A. *Mol. Cell. Proteomics* **2004**, *3*, 1.
(282) Eriksson, J.; Fenyo, D. *J. Proteome Res.* **2004**, *3*, 979.
(283) Stead, D. A.; Preece, A.; Brown, A. J. *Mol. Cell. Proteomics* **2006**, *5*, 1205.
(284) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426.
(285) Russell, S. A.; Old, W.; Resing, K. A.; Hunter, L. *Int. Rev. Neurobiol.* **2004**, *61*, 127.

CR068213F