

Extent of Modifications in Human Proteome Samples and Their Effect on Dynamic Range of Analysis in Shotgun Proteomics*[§]

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

The complexity of the human proteome, already enormous at the organism level, increases further in the course of the proteome analysis due to *in vitro* sample evolution. Most of *in vitro* alterations can also occur *in vivo* as post-translational modifications. These two types of modifications can only be distinguished *a posteriori* but not in the process of analysis, thus rendering necessary the analysis of every molecule in the sample. With the new software tool ModifiComb applied to MS/MS data, the extent of modifications was measured in tryptic mixtures representing the full proteome of human cells. The estimated level of 8–12 modified peptides per each unmodified tryptic peptide present at $\geq 1\%$ level is approaching one modification per amino acid on average. This is a higher modification rate than was previously thought, posing an additional challenge to analytical techniques. The solution to the problem is seen in improving sample preparation routines, introducing dynamic range-adjusted thresholds for database searches, using more specific MS/MS analysis using high mass accuracy and complementary fragmentation techniques, and revealing peptide families with identification of additional proteins only by unfamiliar peptides. Extensive protein separation prior to analysis reduces the requirements on speed and dynamic range of a tandem mass spectrometer and can be a viable alternative to the shotgun approach. *Molecular & Cellular Proteomics* 5:2384–2391, 2006.

The human proteome is the most complex system in molecular biology. Today's consensus puts the number of human genes in the range from 20,000 to 25,000 (1). Further complexity is added at several levels mainly in the form of alternative splicing and post-translational modifications (PTMs).¹ It is believed that at least 40–60% of all human genes give alternative splicing isoforms (2, 3). Large scale studies on chromosomes 21 and 22 indicate that over 80% of the genes could undergo alternative splicing (4). Once synthesized on the ribosomes, most proteins undergo a multitude of PTMs, the exact type and position of which often cannot be pre-

dicted based on genetic information alone. These PTMs can consist of protein chains being cleaved, many different chemical groups can be attached to them (e.g. acetyl, methyl, phosphoryl, sugars, and lipids), and finally proteins can be internally or externally cross-linked by e.g. disulfide bonds. More than 200 different types of PTMs are currently known, and many more are yet to be discovered (5). With 20 common amino acids, this figure corresponds to more than 10 different types of PTM per amino acid. When combining the complexity generated by alternative splicing with that produced by PTMs, the current estimate of the number of different protein molecules expressed in a given individual organism is close to a million, which is roughly 50 forms per gene (6). Because an average human protein consists of 500 amino acids, the overall modification rate is approximately one modification per 10 amino acid residues. With all likelihood, this figure is an underestimation, perhaps even a significant one.

The extent of modifications in a proteome sample is of importance not only for biology but also for analytical sciences. For instance, the extent of modifications is very important for shotgun proteomics, which is based on identification of (largely unmodified) peptides derived from enzymatically digested complex protein mixtures (7). The shotgun approach is known to face the so-called dynamic range challenge arising from the fact that concentrations of proteins in whole proteomes or complex mixtures such as blood plasma differ by many orders of magnitude (8). For instance, relative protein concentrations in human plasma range from 1 to 10^{11} (9). The challenge therefore is to detect low abundance peptides in the presence of much more abundant competitors (10). Extensive modifications can exacerbate this problem significantly as traditional database search approaches allow for only a few modifications to be included in the search; these are usually the most frequent modifications, such as deamidation and methionine oxidation. Abundant modifications could reduce the detection probability as modified peptides from abundant proteins might mask peptides from low abundance proteins. This could result in an elevated rate of false positive identifications and/or affect the reproducibility of the proteome analysis, which is rather poor (9).

Sample treatment and preparation for proteomics analysis can also result in *in vitro* modifications. Some of these modifications are relatively easy to recognize, such as carbamidomethylation of cysteine residues (11), but most *in vitro*

From the Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Box 583, Uppsala S-751 23, Sweden

Received, July 5, 2006, and in revised form, September 27, 2006
Published, MCP Papers in Press, October 2, 2006, DOI 10.1074/mcp.M600248-MCP200

¹ The abbreviation used is: PTM, post-translational modification.

modifications can also occur *in vivo*, such as methionine oxidation (12, 13) and asparagine deamidation (14–17). With no method existing for *a priori* distinguishing between *in vitro* and *in vivo* modifications, mass spectrometry has to make a full sample analysis with filtering out *in vitro* modifications after the data processing step. The extent of *in vitro* modifications is difficult to estimate as most of them occur substoichiometrically. In the most extensive to date analysis of PTMs of individual proteins (18), the ratio between *in vivo* and *in vitro* modifications was ~1:3. Assuming this ratio to be typical, a usual shotgun proteome sample would therefore yield 0.4 *in vivo* + *in vitro* modifications per amino acid. This corresponds to four to five modified peptides for each unmodified tryptic peptide with the average length of 10–12 amino acids. Even this figure is likely an underestimation. Reacquiring replicate samples has shown only a 50% overlap in identified peptide sequences (19), and this observed detection probability is different for low abundance modified peptides compared with high abundance unmodified peptides.

Even when abundant, unmodified, and well isolated proteins are digested in solution the number of detected peptides exceeds the number of expected tryptic peptides by at least an order of magnitude (20). With low resolution MS/MS instruments using only collision-activated dissociation, the success rate of identification of tryptic peptides with a few trivial modifications allowed is only 5–6% (21), whereas with high resolution instruments successful identifications are obtained for 10–15% of produced MS/MS spectra (22). Success rates of automatic data assignment over 60% have been reported (23), but these results concern analysis of highly purified and rather simple protein mixtures and not shotgun proteomics. The more typical success rate for shotgun proteomics of 5–15% identification has been mostly blamed on the poor quality of many MS/MS spectra. However, recently introduced S-score analysis (24) has revealed that only 30% of data are of insufficient quality, whereas approximately half of the remaining 70% of data were either modified peptides or sequences not present in the database (24). This result has triggered the development of the first proteomics-grade *de novo* sequencing procedure, which increased the amount of assigned MS/MS data to 35%, the record for the shotgun approach (25). Still many good quality MS/MS spectra remained unexplained; this could be due to extensive modifications.

Here we report on a study undertaken to quantify the extent of modifications using the novel ModifiComb approach (26). The ModifiComb program analyzes MS/MS information and assigns to base peptides (unmodified, known peptide sequences) so-called dependent peptides, *i.e.* modified versions of the base peptides. The ΔM histogram that ModifiComb creates contains information on all modifications present in the sample above a certain level, in our case substoichiometric modifications present in a dynamic range

of 1:100. Unexpected and novel modifications can easily be revealed by ModifiComb (23) as well as the number of modified peptides and the types of modifications for any given base peptide.

EXPERIMENTAL PROCEDURES

Sample Preparation—500 μg of A431 human epidermoid carcinoma and SD1 human acute lymphoblastic leukemia cell lysates was loaded onto four independently prepared one-dimensional SDS-PAGE gels (~30–200 kDa). Fixing and staining of the gels were performed using ethanol for the SD1 and one of the A431 cell lysates, whereas methanol was used for the remaining two A431 cell lysates. All protein bands were visualized with colloidal Coomassie Blue. The lanes were excised into 30–35 equally sized fractions, and samples were in-gel reduced, alkylated, and digested with modified sequence-grade trypsin (Promega, Madison, WI) as described previously in the literature (11). Finally the samples were vacuum-centrifuged to remove all organic solvents and reconstituted prior to analysis in 20 μl of HPLC water containing 0.1% TFA (Sigma).

Nanoflow LC/MS/MS—All experiments were performed on a 7-tesla hybrid linear ion trap Fourier transform mass spectrometer (LTQ FT, Thermo Electron, Bremen, Germany) modified with a nano-electrospray ion source (Proxeon Biosystems, Odense, Denmark). The high performance liquid chromatography setup used in conjunction with the mass spectrometer consisted of a solvent degasser, nanoflow pump, and thermostated microautosampler (Agilent 1100 nanoflow system). A 15-cm fused silica emitter (75- μm inner diameter, 375- μm outer diameter; Proxeon Biosystems) was used as analytical column. The emitter was packed in-house with a methanol slurry of reverse-phase, fully end-capped Repronil-Pur C₁₈-AQ 3- μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) using a pressurized “packing bomb” operated at 50–60 bars (Proxeon Biosystems). Mobile phases consisted of 0.5% acetic acid and 99.5% water (v/v) (buffer A) and 0.5% acetic acid and 10% water in 89.5% acetonitrile (v/v) (buffer B). 8 μl of prepared peptide mixture was automatically loaded onto the column and rinsed for 20 min in 4% buffer B at a flow rate of 500 nl/min followed by a 90-min gradient from 4 to 45% buffer B at a constant flow rate of 200 nl/min. MS analysis was performed using unattended data-dependent acquisition mode in which the mass spectrometer automatically switches between a high resolution survey scan (resolution = 100,000; *m/z* range, 200–1600) followed by lower resolution fragmentation spectra (electron capture dissociation (27, 28) followed by collision-activated dissociation; resolution = 25,000) of the two most abundant peptides eluting at a given time.

Peptide and Modification Identification—Acquired RAW files were converted to dta files using Extract_msn through BioWorks Browser (Thermo Electron), and complementary pairs were identified as described previously (22). Base peptides were identified by searching against the International Protein Index (IPI) human database (version 3.14; 57,032 sequences; downloaded January 9, 2006) using the Mascot (29) search engine (version 2.1, Matrix Science, London, UK). Searches were performed with trypsin specificity (30), and mass tolerance for monoisotopic peptide identification was set to 5 ppm and ± 0.02 Da for fragment ions. The instrument setting was “ESI-FTICR,” which only permits *b*, *y*, *b* – NH₃, and *y* – H₂O fragment ion types. Only base peptides having a Mascot score (M-score) above the significant threshold of 26 were used for this study (*p* < 0.05). For identification and validation of the dependent peptide sequences revealed by ModifiComb, all data were researched with the Mascot search engine allowing for the known or user-defined variable modifications. The peptide mass tolerance, mass accuracy window for fragment ions, and enzyme specificity as well as the instrument

settings were kept unchanged. Parsing of data and statistical analysis of the search results reported by Mascot were performed using the open-source software MSQuant (50).

RESULTS

Tryptic mixtures for whole proteomes of the human cell line A431 were analyzed; 16,130 unique peptide sequences were identified. To determine the influence of the instrument speed on analysis result, the same sample was run again, this time producing 11,301 unique sequences. The overlap between two independent runs was 6,605 sequences, or close to 50%. Thus half of the abundant peptides are not detected in a single run due to the limited speed of the instrument, and each sample needs to be run at least twice with MS/MS data pooled. Fig. 1a shows the ModifiComb ΔM histogram for the pooled A431 data. Two quantities were of particular interest: the average number of modified peptides per base (unmodified) peptide for the whole proteome (N_{av}) and the respective figure (N_{ab}) for an abundant protein. The N_{av} value was found to be 2.4. To verify this result, the data were analyzed for a different cell line (SD1), and a very similar N_{av} value of 2.2 was obtained.

All of the most abundant modifications detected in Fig. 1a may occur both *in vivo* and *in vitro*, including deamidation and oxidation (31), methylation (32), and formylation (33). To highlight the role of sample preparation, methanol was replaced by ethanol during fixing and staining of the one-dimensional gel. This procedure is usually quite time-consuming (overnight procedure) and therefore could be a large contributor to *in vitro* modifications (34). The resultant ModifiComb ΔM histogram is shown in Fig. 1b. Although frequencies for some modifications, first of all methylation ($\Delta M = +14.02$), have dramatically decreased, some other modifications appeared with increased abundances, in particular oxidation of tryptophan into kynurenine ($\Delta M = +3.99$). This modification has been known for over 35 years since it was first identified in hen egg white lysozyme (35). Previously it was assumed that tryptophan oxidation was analogous to methionine oxidation and therefore a common artifact of sample handling (36). But Taylor *et al.* (37) showed that tryptophan oxidation is not correlated to methionine oxidation, and since then it was believed to predominantly be an *in vivo* protein modification. The increased levels of tryptophan oxidation in our experiments indicate that ethanol facilitates *in vitro* formation of this modification, which needs to be further analyzed. But the N_{av} value did not significantly change: it was found to be 2.2 in the case of ethanol *versus* 2.4 for methanol. To be certain that the observed differences between methanol- and ethanol-treated samples was due to the solvent change, a second batch of A431 prepared in ethanol was analyzed. A high correlation was found ($r = 0.93$, data not shown) between ModifiComb histograms of two ethanol samples, whereas between samples treated with different alcohols the correlation was poorer ($r = 0.73$, data not shown). The lowered correlation was

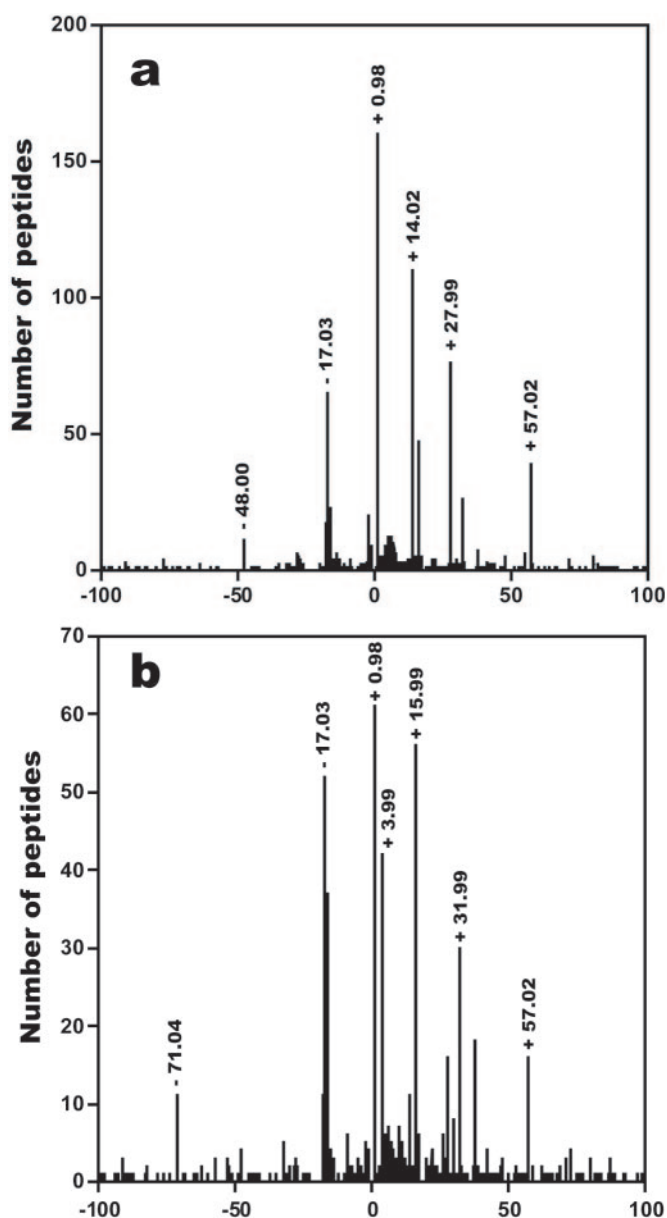


FIG. 1. ModifiComb histograms of ΔM values (ΔM is the molecular mass difference between the modified and unmodified peptides) for A431 cells prepared using methanol (A) and A431 cells prepared using ethanol (B). Highest abundant peaks are labeled with their respective ΔM values.

mainly due to large differences in a few ΔM channels, like methylation and oxidation of tryptophan. This result confirms that the observed difference in modifications states of the samples was indeed due to the solvent change.

The N_{av} value of >2 modified peptides per each unmodified one is an average figure for proteins of vastly different abundances. This value should be much higher for very abundant proteins. As a model for such a protein, actin was chosen,

accounting for 5–10% of the total protein content. Actin is a 43-kDa cytoplasmic protein that serves as a main component of the cytoskeleton system in eukaryotic cells (38, 39). A ModifiComb ΔM histogram for actin (IPI00021438) pooled from three experiments is shown in Fig. 2. Here the N_{ab} value

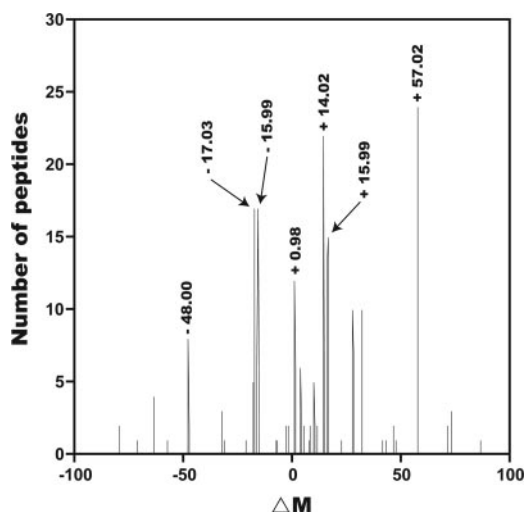


FIG. 2. ModifiComb histogram of ΔM values for actin cytoplasmic 1 protein (combined results from three independently prepared and analyzed A431 cell lysates). Highest abundance peaks are labeled with their respective ΔM values.

(average number of dependent peptides per unmodified one for an abundant protein) is 6.5. The full list of the number of unique dependent peptides for each of the 26 detected base peptides (sequence coverage, 95%) is given in Table I. Note that 11 base peptides have 10 or more dependent peptides.

The peptide with the sequence HQGVMVGMGQK was selected for more detailed investigation. Table II presents data on the dependent peptides found for this peptide, including their relative abundance. Some of the dependent peptides had multiplicity >1 , meaning that several chromatographically separated modified molecules with the same modification mass appeared in the same LC/MS/MS dataset. This could be due to the modification being present at different amino acid residues (e.g. one of the two methionines present). Potentially a structural modification like racemization of amino acids could cause a significant change in the elution time in nano-LC (40). If one takes into account the multiplicity of different isoforms of the dependent peptides, the N_{ab} value increases to 7.8. Again this figure is clearly an underestimation as it embraces only modifications with the mass $-100 \dots +100$ Da, excluding such abundant modifications as glycosylation. Besides, this figure is based on three LC/MS/MS runs, each with a 50% probability of detecting a peptide with average abundance, which gives $\sim 88\%$ probability to detect all peptides. The 12% correction increases N_{ab} to 8.7. Note also that some large and highly modifiable peptides appeared

TABLE I

List of detected base (unmodified) peptides of the abundant protein actin in a human proteome sample (A431 cell line)

Mowse scores (M-score) of unique base peptides and numbers of unique dependent peptides detected for each base peptide by ModifiComb are given. The data are pooled from three independent LC/MS/MS runs.

Base peptide	Position	M-score	Dependent peptides
AGFAGDDAPR	18–27	83	3
AVFPSIVGRPR	28–38	75	4
HQGVMVGMGQK	39–49	111	18
DSYVGDEAQS	50–60	103	13
DSYVGDEAQSKR	50–61	63	8
GILTLK	62–67	35	1
YPIEHGIVTNWDDMEK	68–83	91	10
IWHHTFYNELR	84–94	81	18
VAPEEHPVLLTEAPLNPK	95–112	95	16
TTGIVMDSGDGVTHTVPIYEGYALPHAILR	147–176	135	16
LDLAGR	177–182	35	3
DLTDYLMK	183–190	67	10
GYSFTTTAER	196–205	61	8
DIKEK	210–214	35	1
EKLCYVALDFEQEMATAASSSSLEK	213–237	29	1
LCYVALDFEQEMATAASSSSLEK	215–237	98	2
SYELPDGQVITIGNER	238–253	86	10
CPEALFQPSFLGMESCGIHETTFNSIMK	256–283	39	1
CDVDIRK	284–290	42	5
KDLYANTVLSGGTMYPGIADR	290–311	149	14
DLYANTVLSGGTMYPGIADR	291–311	54	17
EITALAPSTMK	315–326	93	9
IKIIPPER	326–335	35	2
IIAPPER	328–334	31	1
IIAPPERK	328–335	30	1
QEYDESGPSIVHR	359–371	116	10

TABLE II

List of dependent peptides detected for the base peptide HQGVVMVGMGQK from actin in the pool of three datasets from A431 cell lines

Relative (Rel.) abundances of modified peptides are determined as the integral of the chromatographic peak of their most abundant charge state normalized by the corresponding value for the base peptide. ΔRT is the average difference in retention times (in minutes) between the base and the dependent peptide in reversed-phase nano-LC. Multiplicity is the number of distinct chromatographic peaks detected for the same modification mass. For peptides having multiplicity >1 the average ΔRT value is listed.

ΔM	Probable origin	Rel. abundance	ΔRT	Multiplicity
Da		%	min	
-79.994	-CH ₄ O ₂ S	33.2	-3.0	1
-63.998	-CH ₄ OS	17.2	-14.4	1
-48.004	-CH ₄ S	25.8	-19.4	2
-32.011	-H ₂ NO	24.7	-18.4	1
-15.995	-O	4.8	4.3	1
0.983	-NH +O	81.7	0.3	1
15.995	+O	6.5	-3.4	2
27.010	+CHN	2.2	-0.9	1
27.995	+CO	14.7	8.9	3
31.986	+O ₂	1.1	-5.8	1
43.003	+CHNO	1.0	-4.4	1
47.983	+O ₃	1.1	9.2	1
57.019	+C ₂ H ₃ NO	5.0	2.0	2

in Table I with multiplicity of 1 (e.g. peptide 213–237 containing Cys, Gln, and Met and peptide 256–283 containing Cys, Gln, and two Met residues). The dependent peptides for these molecules are probably present in the sample but remained undetected. Thus the true value of N_{ab} should exceed 10 and become comparable with the typical width of a tryptic peptide (10–12 residues).

The likely reason for the low multiplicities of long peptides is their very high acidity (Table I), which leads to high m/z values at which their ions appear in mass spectra (41). The high m/z value is detrimental for detection due to the time-of-flight effect during ion transfer from the linear ion trap to the FT detector. If only peptides not exceeding 20 amino acid residues are considered (21 of 26 peptides in Table I), a decent correlation is obtained between their length and the number of dependent peptides excluding the multiplicity factor (Fig. 3). The correlation factor $r = 0.68$ is far greater than the threshold value of 0.55 that requires for $n = 21$ data points to reject with 99% certainty the null hypothesis that there is no linear correlation (42).

Note that the slope of the linear fit in Fig. 3 is close to one, meaning that one residue length increment results on average in one extra modified peptide form. Thus both the slope in Fig. 3 and the N_{ab} value point toward the same average modification rate: close to one modification per amino acid residue. This new estimate is 2.5 times higher than the conservative estimate based on the 1:3 ratio of *in vivo* and *in vitro* modifications (18).

DISCUSSION

The presence of many substoichiometric modifications does not require the mass distribution of an intact protein to be exceptionally broad if such distribution is observed with an instrument possessing limited dynamic range. Consider an

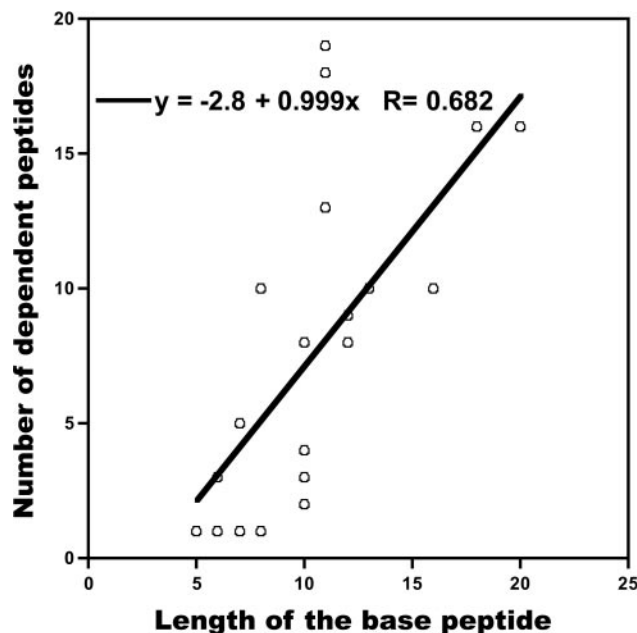


FIG. 3. The number of dependent peptides (excluding multiplicity) per base peptide as a function of the length of the base peptide. The data are taken for 21 of 26 peptides in Table I with the length not exceeding 20 amino acids (AA).

analogy with the isotopic distribution of the small protein glucagon, C₁₅₃H₂₂₄N₄₂O₅₀S. Glucagon molecule contains 470 atoms, close to the number of amino acid residues in an average human protein. Each of the 470 atoms can appear in at least two forms, the light, most abundant isotope and the heavy isotope(s). Neglecting the fine structure of isotopic masses, there are at least 521 main “isoforms” of the same molecule, each with unique molecular mass that differs from any other isoform by at least 1 Da. Yet the “unmodified”

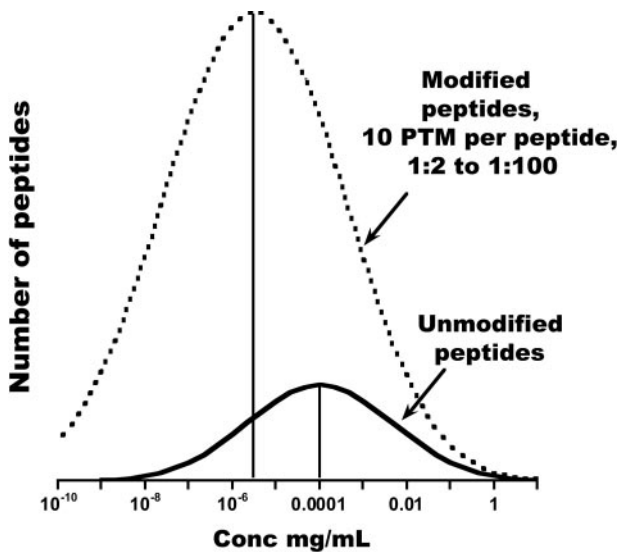


FIG. 4. Theoretical distribution of unmodified tryptic peptide concentrations in a complex biological sample (solid line) and the resulting distribution of modified peptide concentrations (dashed line) assuming 10 modifications per peptide at a substoichiometric range of 1:2 to 1:100.

(monoisotopic) form composes 13% of the total isotopic abundance, and the most abundant form contains just two “modifications” (its mass is 2 Da above the monoisotopic mass). With the dynamic range of analysis 1:10, only six isoforms (isotopic peaks) are observed in the mass spectrum, with 1:100 nine isoforms are observed, with 1:1000 12 isoforms are observed, and so on. Although the most common “modification” increases the atomic mass by 8.3% (^{13}C compared with ^{12}C), the overall mass increase due to modifications is only 0.06%.

Bearing this analogy in mind, one should not be surprised that extensive substoichiometric modifications of human proteins do not result in extraordinary broad isoform distribution or large increase of the molecular mass. Only with a very large dynamic range of detection does a multitude of isoforms become apparent. In the case of glucagon, >30 isotopic isoforms become detectable with the dynamic range of measurements 1:10¹¹. Thus the high modification rate of tryptic peptides will have the most impact on measurements performed with high dynamic range.

Fig. 4 depicts a theoretical distribution of concentrations of unmodified tryptic peptides in a complex biological sample, such as a whole proteome or blood plasma. The distribution is assumed to be normal on a logarithmic scale (43) with concentrations starting at 10 mg/ml and peaking at 10⁻⁴ mg/ml, some 5 orders of magnitude below that of the unmodified peptides (solid line). Assuming that 10 modified forms of every tryptic peptide have concentrations 2–100 times below that of the respective unmodified molecule (the distribution of concentrations of modified molecules is homogeneous), a distribution of modified peptides is obtained (dashed line). This

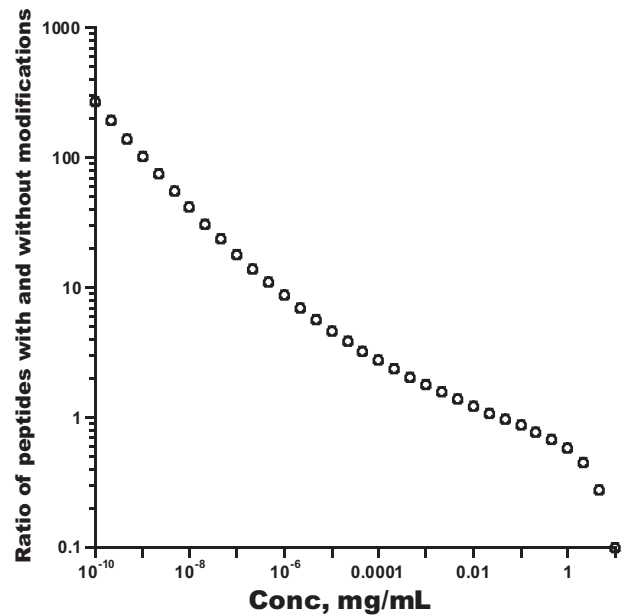


FIG. 5. The average number of modified peptides per single unmodified peptide at a given concentration. The distributions of modified and unmodified peptides are shown in Fig. 4.

distribution is broader than the former, and it peaks near 10⁻⁶ mg/ml, almost two orders of magnitude below that of the unmodified peptides.

The ratio between the number of modified and unmodified peptides (not necessarily of the same sequence) as a function of concentration derived from Fig. 4 is shown in Fig. 5. This ratio, initially small, changes dramatically with concentration. When measurements are performed with the dynamic range 1:1000 (above concentrations of 0.01 mg/ml), the number of unmodified peptides detected exceeds that of modified peptides. Thus modified peptides can be ignored without severe effect on the database search efficiency and false positive rate of protein identification. However, in the interval of the dynamic ranges between 1:1000 and 1:10⁷ (concentrations down to 10⁻⁶ mg/ml), several modified peptides will be detected per each unmodified one. In the region below 1:10⁷, the ratio rapidly grows in favor of modified peptides, exceeding 100 at 1:10⁹. In these regions, the presence of modified peptides can no longer be ignored and must be addressed by choosing an appropriate strategy.

The strategy for eliminating the difficulty posed by the multitude of modified peptides cannot be based on more extensive peptide separation or on increasing the dynamic range and speed of the mass spectrometer used. One obvious effort is to determine the extent of *in vitro* modifications and to optimize the sample preparation routine to minimize the side reactions. Another feasible way is to apply *much more stringent criteria* for identification of low abundance peptides compared with high abundance molecules. Note that modified peptides have in general a high rate of matching with unmodified sequence libraries, including reversed sequence data-

bases (44, 45). The impact of an elevated rate of false positives on reliability of protein identification can be significant for low abundance proteins. In a recent study, transition from 1 to 5% false positive rate for peptides resulted in erroneous assignment of eight of nine of the additionally (mis)identified proteins (46).

Applying more stringent criteria for low abundance peptides may require taking into account all possible modifications while doing a database search, although this is a formidable task. Inclusion of variable modifications greatly extends the database in which the search is performed. Because the reversed (“decoy”) database will expand respectively, the false positive rate determined as the frequency of random matches to the reversed database may also increase significantly. For instance, addition of just one modification, phosphorylation affecting Ser, Thr, or Tyr, can double the false positive rate at the same threshold score (46). To avoid explosion of false positives, increasing the threshold acceptance score will be needed; this can be detrimental for the sensitivity of the analysis and for the rate of false negative identifications.

Another aspect is the speed of data processing. With conventional search engines, inclusion of more than 8–10 variable modifications can slow the search by several orders of magnitude. “Blind” searches that do not assume *a priori* modification types are more time-efficient than searches with explicitly listed variable modifications (26, 47), but they are also more demanding in terms of data quality. Thus the benefits of high mass accuracy, including that in MS/MS, and complementary fragmentation techniques (22) are becoming even more apparent (48).

An alternative to the inclusion of a myriad of PTMs in the database search and an increase in the acceptance threshold can be extensive presorting of MS/MS data before database search. There are number of strategies available for determining similarity between MS/MS data, e.g. ModifiComb (26, 47). All mutually similar MS/MS data can be classified by these approaches as belonging to one peptide “family.” Thus the whole dataset can be separated into a large number of peptide families, each family being significantly different from any other family according to the criteria used. To each family a unique peptide sequence can be assigned through the database search or *de novo* sequencing. Each protein can only be identified by at least one unique peptide family. Such analysis should severely reduce the risk for a modified version of an abundant peptide to be misidentified as an unmodified peptide with a different sequence. However, the same prefiltering may put in the same family non-modified sequences that are homologues but different, thus increasing the rate of false negatives.

Extensive separation of proteins or protein isoforms (with depletion of abundant proteins as an optional first step) prior to digestion ensuring that concentrations of molecules in every fraction are more or less similar may be a viable alternative

to the shotgun strategy. Such analysis will be fully compatible with the dynamic ranges 1:100 to 1:1000 that are typical for most modern mass spectrometers. The medium dynamic range analysis will detect only major modifications for each protein and ignore rare isoforms; this is suitable and even desirable in some applications as many *in vitro* modifications will be kept below the detection threshold and thus remain undetected. Complexity of the protein fraction should be such that the number of protein isoforms per fraction does not exceed 10–100. Modern two-dimensional techniques can routinely separate complex protein mixtures into a few hundred non-overlapping fractions (7, 49), which may be enough to reduce the average fraction complexity to the desired level.

Realization that the needle of unmodified, low abundance peptides can be easily lost in the haystack of abundant modified peptides should not be viewed as a fatal blow to the shotgun strategy that has proved its utility in many studies. To the contrary, the challenge should be perceived as a strong driving force for the development of better sample preparation and protein separation techniques and novel, more informative MS/MS approaches. In our view, such approaches should be based on high mass accuracy in both MS and MS/MS and complementary fragmentation techniques (48).

Acknowledgments—Thomas Köcher, Frank Kjeldsen, and Christopher Adams are acknowledged for insightful discussion.

* This work was supported by the Knut and Alice Wallenberg Foundation and Wallenberg Consortium North (Grant WCN2003-UU/SLU-009 and instrumental grant for FTMS) as well as the Swedish research council (Grants 621-2004-4897, 621-2002-5025, and 621-2003-4877). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ The on-line version of this article (available at <http://www.mcponline.org>) contains supplemental material.

‡ To whom correspondence should be addressed. Tel.: 46-18-471-5729; Fax: 46-18-471-7209; E-mail: Roman.Zubarev@bmms.uu.se.

REFERENCES

1. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945
2. Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002) Alternative splicing and genome complexity. *Nat. Genet.* **30**, 29–30
3. Modrek, B., and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19
4. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H., and Gingeras, T. R. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342
5. Creasy, D. M., and Cottrell, J. S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536
6. Meri, S., and Baumann, M. (2001) Proteomics: posttranslational modifications, immune responses and current analytical tools. *Biomol. Eng.* **18**, 213–220
7. Wolters, D. A., Washburn, M. P., and Yates, J. R. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690
8. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**,

- 845–867
9. Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R., and Lohley, A. (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* **3**, 311–326
 10. De Godoy, L. M., Olsen, J. V., De Souza, G. A., Li, G., Mortensen, P., and Mann, M. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **7**, R50
 11. Wilm, M., Shevchenko, A., Houthaev, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466–469
 12. Chao, C. C., Ma, Y. S., and Stadtman, E. R. (1997) Modification of protein surface hydrophobicity and methionine oxidation by oxidative systems. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 2969–2974
 13. Levine, R. L., Mosoni, L., Berlett, B. S., and Stadtman, E. R. (1996) Methionine residues as endogenous antioxidants in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 15036–15040
 14. Chelius, D., Rehder, D. S., and Bondarenko, P. V. (2005) Identification and characterization of deamidation sites in the conserved regions of human Immunoglobulin Gamma antibodies. *Anal. Chem.* **77**, 6004–6011
 15. Huang, L. H., Li, J. R., Wroblewski, V. J., Beals, J. M., and Riggan, R. M. (2005) In vivo deamidation characterization of monoclonal antibody by LC/MS/MS. *Anal. Chem.* **77**, 1432–1439
 16. Karty, J. A., and Reilly, J. P. (2005) Deamidation as a consequence of beta-elimination of phosphopeptides. *Anal. Chem.* **77**, 4673–4676
 17. Robinson, A., and Robinson, L. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 8880–8884
 18. Wilmarth, P. A., Tanner, S., Dasari, S., Nagalla, S. R., Riviere, M. A., Bafna, V., Pevzner, P. A., and David, L. L. (2006) Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility. *J. Proteome Res.* **5**, 2554–2566
 19. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., and Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **4**, 1487–1502
 20. Jones, R. C., Thypambal, S., Taylor, J. T., and Edmondson, R. D. (2006) False discovery rates in protein identification, in *54th American Society for Mass Spectrometry Conference on Mass Spectrometry, Seattle, May 27–June 1, 2006*, American Society for Mass Spectrometry, Santa Fe, NM
 21. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
 22. Nielsen, M. L., Savitski, M. M., and Zubarev, R. A. (2005) Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **4**, 835–845
 23. Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzhradszky, K. F., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: I. How much of the data is theoretically interpretable by search engines. *Mol. Cell. Proteomics* **4**, 1189–1193
 24. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol. Cell. Proteomics* **4**, 1180–1188
 25. Savitski, M. M., Nielsen, M. L., Kjeldsen, F., and Zubarev, R. A. (2005) Proteomics-grade de novo sequencing approach. *J. Proteome Res.* **4**, 2348–2354
 26. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006) ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteomics* **5**, 935–948
 27. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **120**, 3265–3266
 28. Savitski, M. M., Kjeldsen, F., Nielsen, M. L., and Zubarev, R. (2006) Complementary sequence preferences of electron-capture dissociation and vibrational excitation in fragmentation of polypeptide polycations. *Angew. Chem. Int. Ed. Engl.* **45**, 5301–5303
 29. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567
 30. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**, 608–614
 31. Stadtman, E. R. (1992) Protein oxidation and aging. *Science* **257**, 1220–1224
 32. Ambler, R. P., and Rees, M. W. (1959) Epsilon-N-Methyl-lysine in bacterial flagellar protein. *Nature* **184**, 56–57
 33. Goodlett, D. R., Armstrong, F. B., Creech, R. J., and van Breemen, R. B. (1990) Formylated peptides from cyanogen bromide digests identified by fast atom bombardment mass spectrometry. *Anal. Biochem.* **186**, 116–120
 34. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685
 35. Previero, A., Coletti-Previero, M. A., and Jolles, P. (1967) Localization of non-essential tryptophan residues for the biological activity of lysozyme. *J. Mol. Biol.* **24**, 261–268
 36. Potgieter, H. C., Ubbink, J. B., Bissbort, S., Bester, M. J., Spies, J. H., and Vermaak, W. J. (1997) Spontaneous oxidation of methionine: effect on the quantification of plasma methionine levels. *Anal. Biochem.* **248**, 86–93
 37. Taylor, S. W., Fahy, E., Murray, J., Capaldi, R. A., and Ghosh, S. S. (2003) Oxidative post-translational modification of tryptophan residues in cardiac mitochondrial proteins. *J. Biol. Chem.* **278**, 19587–19590
 38. Janmey, P. A., Hvidt, S., Oster, G. F., Lamb, J., Stossel, T. P., and Hartwig, J. H. (1990) Effect of ATP on actin filament stiffness. *Nature* **347**, 95–99
 39. Pollard, T. D., and Cooper, J. A. (1986) Actin and actin-binding proteins. A critical evaluation of mechanisms and functions. *Annu. Rev. Biochem.* **55**, 987–1035
 40. Adams, C. M., and Zubarev, R. A. (2005) Distinguishing and quantifying peptides and proteins containing D-amino acids by tandem mass spectrometry. *Anal. Chem.* **77**, 4571–4580
 41. Nielsen, M. L., Savitski, M. M., Kjeldsen, F., and Zubarev, R. A. (2004) Physicochemical properties determining the detection probability of tryptic peptides in Fourier transform mass spectrometry. A correlation study. *Anal. Chem.* **76**, 5872–5877
 42. Bewick, V., Cheek, L., and Ball, J. (2003) Statistics review 7: correlation and regression. *Crit. Care* **7**, 451–459
 43. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003) Global analysis of protein expression in yeast. *Nature* **425**, 737–741
 44. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50
 45. Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P., Ens, W., and Standing, K. G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73**, 1917–1926
 46. Lee, H. K., Picotti, P., Domon, B., and Aebersold, R. (2006) Novel approach to identify low abundance biomarkers in serum, in *54th American Society for Mass Spectrometry Conference on Mass Spectrometry, Seattle, May 27–June 1, 2006*, American Society for Mass Spectrometry, Santa Fe, NM
 47. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567
 48. Zubarev, R. A. (2006) Protein primary structure using orthogonal fragmentation techniques in Fourier transform mass spectrometry. *Expert Rev. Proteomics* **3**, 251–261
 49. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021
 50. Schulze, W. X., and Mann, M. (2004) A novel proteomic screen for peptide-protein interactions. *J. Biol. Chem.* **279**, 10756–10764