ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures*

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

A major challenge in proteomics is to fully identify and characterize the post-translational modification (PTM) patterns present at any given time in cells, tissues, and organisms. Here we present a fast and reliable method (“ModifiComb”) for mapping hundreds types of PTMs at a time, including novel and unexpected PTMs. The high mass accuracy of Fourier transform mass spectrometry provides in many cases unique elemental composition of the PTM through the difference $\Delta M$ between the molecular masses of the modified and unmodified peptides, whereas the retention time difference $\Delta R_T$ between their elution in reversed-phase liquid chromatography provides an additional dimension for PTM identification. Abundant sequence information obtained with complementary fragmentation techniques using ion-neutral collisions and electron capture often locates the modification to a single residue. The ($\Delta M$, $\Delta R_T$) maps are representative of the proteome and its overall modification state and may be used for database-independent organism identification, comparative proteomic studies, and biomarker discovery. Examples of newly found modifications include +12,000 Da (+ C atom) incorporation into proline residues of peptides from proline-rich proteins found in human saliva. This modification is hypothesized to increase the known activity of the peptide.

Molecular & Cellular Proteomics 5:935–948, 2006.

Post-translational modifications (PTMs)$^1$ are key regulators of protein function, localization, and interactions taking place inside the cell (1, 2). PTMs are also required for proper folding of the protein. A major challenge in proteomics is therefore to fully identify and characterize the PTM patterns present at any given time in cells, tissues, and organisms (1, 2). Hundreds of modification sites can be identified in a single MS experiment yielding valuable information in cellular processes (3, 4). The main MS tool in PTM detection is tandem mass spectrometry combined with a database search engine, such as Sequest (5) or Mascot (6). Although modern search engine-based proteomic approaches have been highly successful, they possess a number of significant drawbacks. Before the search, the operator specifies all expected modifications, often marked as “variable,” i.e. not necessarily present. Because the engine considers all peptide sequences with and without variable modifications, including all possible combinations of modifications, database searches with several variable modifications often take a much longer time than it took to collect the experimental data set, creating a bottleneck in high throughput analysis. Allowing for many modifications in the database search increases the rate of false positives (7), eliminating which requires a much higher score threshold for identification of peptides, which leads to an enhanced number of false negative results (misses of present proteins) (7).

To reduce the analysis time and the false positive and negative rates, a typical database search focuses upon a few types of modifications, far fewer compared with the broad variety that potentially can be present in the sample. A database search strategy is limited by nature, and although major improvements have been made over the past couple of years (8), most of the acquired tandem mass spectra remain unidentified through these searches. In a typical LC/MS proteomic-type analysis, the identification success rate usually varies between 5 and 15% (9). Even with FTMS that provides ppm mass accuracy and can use two complementary fragmentation techniques (collisionally activated dissociation (CAD) and electron capture dissociation (ECD) (10)), no more than 30% of MS/MS datasets produce positive identifications (11).

Part of the unidentified mass spectra may be due to unexpected modifications. Fig. 1 shows an example of an endogenous peptide from a human saliva sample sequence suggested by Mascot as peptide WAPGGQQSSQ from an
ModifiComb, a Proteomic Tool for PTM Identification

A) Mass spectrum showing MH$^{2+}$ peaks and CAD in B).

B) MS/MS spectrum of GPQGQGGHQQ with identified y and b ions.

C) Expanded view of B) showing higher resolution of y and b ions.

D) MS/MS spectrum of WAPGGQQ with identified ions.

E) MS/MS spectrum of GPPQQGQGHQQ with identified ions.
unnamed human protein. Although five identified fragments deviated from their theoretical values by less than 11 mDa (Fig. 1, C and D) and the data quality was good (S-score value (11) was four, way above the threshold value of two), the dataset received a Mascot score (M-score) of 18, below the threshold value of 41. A database search using several common variable modifications did not provide a better answer. Subsequently a ModifiComb search (see below) identified the peptide as a modified version of another peptide that eluted from the nano-LC column some 9 min earlier and was 12,000 Da lighter (survey spectrum integrated over the 9-min time interval is depicted in Fig. 1A), identified by Mascot as GP-PQQGGHQQ (Fig. 1B) with M-score of 47. Note that the masses of all 12 identified fragments were internally consistent with experimentally measured masses deviating from the theoretical values by less than 5 mDa and with the deviation changing linearly with the fragment mass (Fig. 1B, inset). The 12,000-Da shift was observed in y8 and y9 fragments as well as in all b fragments. Accurate mass analysis of the mass difference (109.055 Da) between the y8 and y7 fragments revealed the unique elemental composition of the third amino acid (C6H7NO) only 2 mDa away from the theoretical mass (109.053 Da) of the modified proline residue that has the same elemental composition. Thus the identity of the 12,000-Da modified proline was additionally confirmed. Such a proline modification is not reported for humans (12), although analogues of it can be found in the literature (see below). After the insertion of this modification into the Mascot search as a user-defined modification, the modification position was confirmed with M-score of 48 and a nearly perfect fit of 12 fragment masses (Fig. 1E).

This example is typical, and it demonstrates both the problem and the solution. The problem is the presence of modified peptides, sometimes of an unknown type. The solution can be to utilize the fact that peptides within the same LC/MS run may be correlated. Because peptides in the same LC/MS run originate from the same protein mixture (peptide separation prior to LC/MS is not assumed), heterogeneity of PTMs and mutation sites results in the presence within the sample of several closely related variants of the same peptide, a kind of a peptide family. Most PTMs are present in substoichiometric amounts; therefore, each family includes one “base” (unmodified) peptide and one or several “dependent” peptides (modified or/and mutated). Many of the dependent peptides may be expected to elute within a limited time window before or after the base peptide.

Here we report on a software tool (“ModifiComb”) that searches for such peptide families and reveals the PTM and mutation patterns of complex peptide mixtures. The tool “combs out” from large data arrays pairs of peptides with strong sequence similarities, one of which is a base peptide and the other of which is a dependent peptide (Fig. 2). The base peptide is usually identified either via de novo sequencing or database searching, whereas the dependent peptide should not give database identification without variable modifications included. Identification of the base peptide is not critical for the analysis: based on sequence similarity, peptide pairs can be found in a “blind” search without knowing which peptide in the pair is the base one.

Comparing the molecular masses of the peptide inside the pairs, the program builds a ΔM histogram of the differences between them. This ΔM histogram built for one LC/MS run or

Fig. 1. Identification of a novel modification on a peptide belonging to human saliva PRP. A, 9-min integrated survey scan showing two ions separated by 12,000 Da. B, CAD spectrum of the lowest mass ion in the survey scan identified as peptide GP-PQQGHQQ from PRP. The inset shows the mass deviation of the fragment masses for this identification. C, CAD spectrum of the +12,000-Da peptide. Note the similarity between this spectrum and the one depicted in B. Fragment ion assignment corresponds to identification in E. D, fragment mass deviation of peptide sequence suggested by Mascot search engine. All masses are off by ≥11 mDa, and the identification is wrong. E, fragment mass deviation of Mascot identification when novel +12,000-Da modification of proline residue is added to the search. Full sequence cleavage is achieved, and no fragment mass deviates more than 6 mDa.

Fig. 2. Block diagram of the ModifiComb algorithm. See text for additional explanation.
several related runs represents the overall pattern of all mutations and PTMs present in the corresponding sample. For $\Delta M$ values below 100 Da, the mDa mass accuracy of FTMS reveals the corresponding elemental composition of the modification; for $\Delta M > 100$ Da values the high mass accuracy limits the number of possible elemental compositions. Inspection of the MS/MS data can often reveal the position of the modification. Once all the base peptides are identified, the $\Delta M$ histogram takes seconds to build. This identification takes seconds if de novo sequencing is used or minutes if a database search is used (the search is typically used without variable modifications or with a few obvious ones, such as oxidation of methionine). Thus the overall data analysis using the $\Delta M$ histogram is much faster than the data acquisition, removing one of the throughput bottlenecks. The difference in the retention times, $\Delta$RT, between the dependent and base peptides is used as complementary information, although the intrinsic resolution, precision, and accuracy of RT measurements are much below the mass measurements. The ($\Delta M$, $\Delta$RT) pair provides a two-dimensional map of the present PTMs and mutations.

Several earlier analogues of the ModifiComb approach can be found in the literature. Recently approaches have been developed to minimize the computational cost of complete PTM identification by applying database filters (13, 14). The filters are based on peptide sequence tags (15) extracted from the acquired MS/MS data. The tags reduce the database to a much smaller set of sequence candidates that can be searched with multiple variable modifications in a reasonable time. This approach is still largely limited to known protein sequences and modifications and can miss modifications if they occur inside the sequence tag. Additionally this approach is firmly database-oriented, that is rather slow and sensitive to sequence errors that are present in all databases. Finally this approach does not produce sample-specific fingerprint patterns. An ideologically similar strategy has been described by Zhang et al. (16) for a low resolution ion trap. However, that approach only worked on mixtures containing a few proteins.

Recently Tsur et al. (17) described MS-Alignment, a software tool for a blind PTM search in large MS/MS datasets. Although using an impressively sophisticated alignment algorithm, MS-Alignment has a number of limitations. Integer $\Delta M$ values that the algorithm uses mask the underlying complexity of modifications (e.g. modifications with elemental compositions CO (formylation), N$_2$, and C$_2$H$_4$ have the same integer mass of 28). Furthermore MS-Alignment processes CAD-only datasets, and analysis speed requirements limit the $\Delta M$ region ($\pm 100$ to $\pm 160$ Da in Ref. 17). In contrast, ModifiComb uses accurate mass data, has no limit on $\Delta M$ values, and uses combined ECD/CAD datasets. As already mentioned, ModifiComb also makes use of the retention time differences, i.e. uses both dimensions of LC/MS separation, which gives it very high specificity. For instance, a dependent peptide with $\Delta M = +0.977$ Da and a small positive $\Delta$RT is surely a deamidated version of the base peptide, whereas $\Delta M = 1.003$ Da and a large $\Delta$RT is likely due to a monoisotopic mass misassignment in one of the peptides.

There is one more significant different between ModifiComb and other algorithms. The usual approach to reducing the search space for modifications is to identify first the set of proteins present in the sample and then search PTMs in that small database. ModifiComb goes one step further and searches PTMs only for identified peptides, further reducing the search space by an order of magnitude (although ~5 peptides per protein are on average identified in our analysis (18), an average protein produces 50 tryptic peptides). The search space reduction diminishes the probability of false positive PTM identification and obviates the development and validation of a special scoring algorithm (see below). The explicit requirement in the ModifiComb non-blind search for the unmodified peptide to be present is a limitation but not a too narrow one as most PTMs appear in substoichiometric proportions.

In the current work, we tested ModifiComb and built $\Delta M$, $\Delta$RT histograms and ($\Delta M$, $\Delta$RT) maps for several biological samples. Sensitivity, specificity, and repeatability of the approach were evaluated. Because the ability of the program to find new and unexpected modifications by far exceeds our current capacity to characterize them, here our goal was not to report all findings, and we limited the current report to the demonstration and validation of the ModifiComb operation. Several examples of new modifications and sample fingerprinting (M-fingerprinting) are provided as an illustration, and their potential biological importance is discussed.

**EXPERIMENTAL PROCEDURES**

**Sample Collection and Preparation**—Whole human saliva was obtained from a healthy 32-year-old non-smoking Caucasian male taking no medications and with no overt signs of gingivitis or caries. The mouth of the subject was rinsed with water, and the sample was collected 3 h after food intake. To minimize degradation, the sample was collected on ice and during the entire sample preparation procedure kept constantly at 4 °C. A total of 5 ml of saliva was collected of which 2 ml was clarified by centrifugation at 12,000 × g for 10 min, thereby removing debris and cells. The obtained supernatant was loaded onto four 10-kDa mass cutoff filters (Microcon YM-10, 500 µl each) and centrifuged at 14,000 × g for 30 min. In so doing, the endogenous peptides were immediately separated from most of the proteases present in saliva, again minimizing the possibility for further peptide/protein degradation. For the final step of peptide isolation, the flow-through fractions (~10 kDa) were pooled and loaded onto a 3-kDa mass cutoff filter (Microcon YM-3) and again centrifuged at 14,000 × g for 90 min. The final flow-through fractions containing endogenous human saliva peptides were pooled and loaded onto an Oasis HLB Plus cartridge (Waters Corp.) for desalting. The bound peptides were eluted with 90% acetonitrile and 0.5% acetic acid. Finally the peptides were completely dried in a SpeedVac and reconstituted in water containing 0.1% trifluoroacetic acid prior to mass spectrometric analysis. The total time from sample collection to start of the MS analysis was less than 3 h, once again minimizing the chances for sample degradation and therefore removing the need for
protease inhibitors.

K562 human chronic myeloid leukemia, A431 human carcinoma, and Escherichia coli cell lysates were prepared as described previously (11, 18). Briefly 500, 200, and 70 μg of cell lysates, respectively, were loaded onto a one-dimensional SDS gel, and one lane of each lysate was excised into 20–30 gel pieces. The K562 and E. coli samples were prepared a second time following the same procedure, but the E. coli gel was only cut into seven pieces. In-gel reduction, alkylation, and digestion with trypsin (Promega, Madison, WI) were performed as described in the literature (19). Samples were dried to complete dryness using a SpeedVac and reconstituted immediately prior to analysis in 20 μl of water containing 0.1% TFA.

Liquid Chromatography/Mass Spectrometry—Analysis was performed on a 7-tesla hybrid linear ion trap Fourier transform mass spectrometer (LTQ-FT, Thermo, Bremen, Germany) equipped with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark). An HPLC system was used on line with the mass spectrometer. The system (Agilent 1100 nanoflow) consists of a solvent degasser, a nanoflow pump, and a thermostated microautosampler. Solvents used consisted of 99.5% water and 0.5% acetic acid as buffer A and 90% acetonitrile, 9.5% water, and 0.5% acetic acid as buffer B. The peptide sample was automatically loaded at a flow rate of 500 nl/min onto a 15-cm-long Proxeon nano-ESI emitter (75-μm inner diameter, 360-μm outer diameter) packed in house with fully end-capped Reprosil-Pur C18 3-μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). After 20 min of loading time the peptides were eluted from the column with a 90-min gradient (4–45% buffer B) at a 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 300–1500) followed by lower resolution fragmentation spectra (ECD followed by CAD; resolution = 25,000) of the two most abundant peptides eluting at a given time.

Database Search: Peptide Identification—All data were searched using the Mascot search engine (version 2.1, Matrix Science, London, UK) against either the International Protein Index database (human, 3.10; downloaded September 5, 2005) for saliva, K562, and A431 samples or National Center for Biotechnology Information non-redundant database (www.ncbi.nlm.nih.gov; downloaded September 5, 2005) with taxonomy specified to E. coli. For base peptide identification, only oxidized methionine was chosen as a variable modification. Searches were performed with no enzyme specificity for the human saliva sample and with trypsin specificity (20) for K562 human leukemia, A431 human carcinoma, and E. coli cell lysates; 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. 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 the “no enzyme/enzyme specificity” option as well as the instrument settings were kept unchanged. Parsing of data and statistical analysis of the search results reported by Mascot was performed using the open source software MSQUANT (msquant.sourceforge.net).

Algorithm Description: Overview—The block diagram of the ModifiComb algorithm is presented in Fig. 2. The ModifiComb algorithm works under the assumption that both the unmodified as well as the modified version of a peptide are present in the mixture. The ECD and CAD spectra are treated as described in Ref. 11, and a merged fragment list is submitted to the Mascot search engine. A list of Mascot reliably identified (M > 34) base peptide sequences is created for each sample. Another list is created for dependent peptides that were not identified by Mascot (or received a below threshold score). The dependent peptide fragment lists are then compared with those of base peptides, and for each peptide pair, the molecular mass difference ΔM and retention time difference ΔRT are calculated. The ΔRT value is currently approximated by the difference in the scan number of the dependent and base peptides (any given scan duration is a function of the given ion abundance, but the average value of the scan duration fluctuates insignificantly during the peptide elution time). The algorithm determines that the pair “matches” if a certain predefined number (usually four) of fragments of the dependent peptide either coincide within the given mass accuracy with the observed fragments in the base peptide or the corresponding masses are shifted by ΔM. The “matched” peptide pair is reported in the output file. Simultaneously their ΔM and ΔRT data are added to one-dimensional histograms ΔM and ΔRT and a two-dimensional map (ΔM, ΔRT).

Details of the Peptide Matching Algorithm—ModifiComb has two regimes: blind and “open eyed.” In the latter regime, the base peptides are identified either through the Mascot search or de novo sequencing. In the blind regime, the base peptide remains unidentified. Below a detailed description is given for the open eyed regime in the case of base peptide identification by Mascot (the procedure for de novo sequencing is easily extrapolated). First an initial search is performed with no variable modifications except oxidation of methionine for each . The output contains the obtained Mascot score M (M > 0 if Mascot suggested a sequence, and M = 0 if Mascot did not make any suggestion) and the corresponding Mascot-suggested sequence for M > 0. The user defines three parameters.

The M-score threshold M1 above which all suggested sequences are accepted as trustworthy.

The M-score threshold M2 below which all suggested sequences are deemed wrong (because a certain minimal S-score is implicitly required for every tandem MS spectrum ModifiComb considers (see requirement 3), this assumption is correct with a high probability).

The minimum number n of common cleavage sites present when comparing two different peptides. The definition of a common cleavage is given below.

All dta files with M > M1 are considered to belong to base peptides (A), whereas those with M < M2 are viewed as belonging to potential dependent peptides (B). Each possible pair of A and B peptides are considered. For each compared pair, n is calculated in the following way. First ΔM is determined as the difference of the molecular masses between B and A peptides. ΔM is then considered as the mass of the potential modification (this approach intrinsically favors single modifications), which can be assumed as positive as well as negative. The Mascot-suggested peptide sequence for A is used to generate a list of b ion masses {b1, ...,b L} and y ion masses {y1, ...,y L} where L + 1 is the length of the sequence. The masses (m1, ..., m L) in the dta file are already tagged with the likely type of the ion they represent, y, b, or by (either b or y ion) (18). The masses tagged y and b are compared with the theoretical sets {y1, ...,y L} and {b1, ...,b L}, respectively, whereas by-tagged ions are compared with both sets. A “match-1” means that the masses of the theoretical and experimental fragments coincide within 20 mDa, whereas a “match-2” means that they differ by ΔM ± 20 mDa. If some fragment of a defined type (b or y) matches by-match-2, then all subsequent fragments of the same type should also match as match-2. This requirement may seem too stringent because the type of the fragment might be identified incorrectly in which case the above requirement may not be fulfilled for a perfectly legitimate pair A and B.
However, carefully made validation of the fragment type identification procedure has ensured its extremely high reliability, which stems from the high mass accuracy and the consensus-based selection rules (11, 18). After the matching procedure is done, the number \( N \) of matched cleavage sites in the given pair of dta files is counted (matched complementary b and y ions are counted as one cleavage site) and reported. If \( N \geq n \), the peptide pair is considered a hit, and the \( \Delta M \) value, \( N \), the scan number difference between the two dta files, their identities, and Mascot scores are reported.

**Database-independent Comparison**—A blind search is performed when neither the protein database nor reliable de novo sequencing data are available, e.g. it can be used for analysis of a set of unknown proteins. In the database-independent comparison, all dta files are compared with each other. Fragment masses are compared with each other in the same way as described above taking into account their ion type tags, and the number of common cleavage sites is the discriminating parameter. The major difference with the above database-dependent comparison is that here all \( \Delta M \) values are given a positive sign. This is because in the database-independent approach it may be difficult to know which peptide in the pair is unmodified; therefore, the lighter peptide is always selected as the base one.

**RESULTS**

The above ModifiComb procedure was applied to the MS/MS datasets derived from an E. coli cell lysate, a K562 human chronic myeloid leukemia cell lysate, and an A431 human carcinoma cell lysate as well as a sample of <3-kDa endogenous peptides from human saliva.

**A431 Sample**—Fig. 3 shows the \( \Delta M \) histogram acquired with \( n = 4 \) and plotted with a 5-mDa resolution for the region from −100 Da to +100 Da. The two insets show the compressed regions stretching in both directions by 1400 Da. The main histogram contains a number of distinct peaks. The high mass accuracy and resolution of FTMS resolves the peaks and assigns to them unique elemental compositions. Main peaks and their assignments are listed in Table I. The peak abundance is proportional to the abundance of the corresponding modification in the sample; thus the \( \Delta M \) histogram represents the modification spectrum. Some of the peaks are doublets as an inset around +28 Da shows. The lighter peak G is mainly due to +CO contribution (formylation), whereas the heavier peak H is due to +C2H4 (ethylation or dimethylation).

The most abundant modifications listed in Table I contain both expected and unexpected findings. Oxidation (peaks F and I) was probably the most trivial case, although exhaustive analysis of the oxidation peaks to confirm that it was confined to methionine was not performed. Deamidation (peak D) is known to occur in vitro during sample storage and preparation (mostly during digestion) (21, 22) but also sometimes in vivo (23, 24). Carbamidomethylation (peak J) is indistinguishable from addition of glycine; it can occur in vitro during sample preparation.
preparation. Loss of ammonia (peak B) could be attributed to ion decomposition in the ion source if not for the clear shift in the retention time of the dependent ions (see below), indicating that this loss occurs in solution. Such a loss may have a number of origins (see Table I), but the details were not investigated. Methylation (peak E) and dimethylation/ethylation (peak H) are common in vivo (25, 26). In vitro formylation (CO) of serine and threonine residues is associated with CNBr cleavage in combination with formic acid (27), but given that neither of these compounds was used the observed formylation should have taken place in vivo. The methanethiol (CH₄S) loss from methionine (peak A) was in principle expected but only from C-terminal methionine as was reported previously for chemical cleavage with CNBr under acidic conditions (28).

Here we found a number of cases where the loss occurs from an internal methionine. The H₂ loss (peak C) was due to several different modifications. Oxidation of threonine into 2-amino-3-ketobutyric acid is known to occur through metal-catalyzed oxidation (29). A combination of methionine oxidation (±15.994915 Da) in concurrence with conversion of aspartic acid into succinimide (−18.010565 Da) makes up for a joint loss of H₂ in some peptide sequences. Several peptides with loss of H₂ occurring from the methionine residue were observed. The methionine in these sequences is located at the N terminus of the peptide and most probably due to an in vitro phenomenon of the digestion process (ring forming between the side chain and N terminus). One should notice that a PTM difference of −2 Da could be mistaken for a mutation of the methionine residue to glutamic acid. High mass accuracy is therefore essential to distinguish between the ΔM due to mutation (ΔM = 1.9979 Da) and due to loss of H₂ (ΔM = 2.0156 Da).

The total number of different types of modifications detected is hard to estimate because besides the peaks marked in Fig. 3 even peaks with a single count may mean a modification. Given that there are a total of 651 counts in the histogram excluding the 10 peaks marked A through J and assuming on average three counts per peak, we conclude that the histogram contains at least 217 unique modifications. Such a number of types of modifications is much larger than previously reported for a single sample, but it is not unexpected given the huge complexity of the human proteome. As reported above, some of the observed mass differences (modifications) are due to several modifications simultaneously present in the same peptide sequence. This complicates the analysis, but only slightly, because a second ModifiComb pass can be performed with all dependent sequences found in the first pass regarded now as base peptides.

False Positive Rate—The estimate of the number of modification types is only valid if the histogram does not contain many false (random) counts. However, random counts should be distributed on the ΔM scale much more evenly than the insets in Fig. 3 show, indicating that the majority of counts in Fig. 3 are real. To evaluate the rate of false positives more precisely, all data on the base peptides of the A431 sample were replaced by the same number of different base peptides from the E. coli sample (see below), and the base-dependent peptide matching routine was repeated with the new dataset. Thus obtained peptide pairs may include “true” coincidence cases due to sequence homology; therefore, this method overestimates the actual false positive rate. In the region

**Table I**

| Peak | ΔM     | Elemental composition | Modification types | Amino acid |
|------|--------|-----------------------|--------------------|-----------|
| A    | −48.003371 | −CH₄S | Loss from internal methionine | Met |
| B    | −17.026549 | −NH₃ | Oxobutanoic acid from N-terminal Thr Pyro-Glu from Gln | Thr, Gin |
|      |         |       | Pyruvic acid from N-terminal Ser S-Carboxymethylcysteine cyclization (N-terminus) | Ser, N terminus |
| C    | −2.015650 | −H₂   | Didehydro Threonine oxidation to 2-amino-3-oxo butanoic acid | Ser, Tyr |
| D    | 0.984016  | +O    | Deamidation | Asn, Gin |
| E    | 14.015650 | +CH₂  | Methylation Methyly ester | Cys, His, Lys, Asn, Gln, Arg, Ile, Leu, N terminus |
| F    | 15.994915 | +O    | Oxidation | Met, His, Trp |
| G    | 27.994915 | +CO   | Formylation | Lys, Ser, Thr, N terminus |
| H    | 28.031300 | +CH₄   | Ethylation | Lys, Glu, N terminus |
| I    | 31.989829 | +O₂   | Dimethylation | Lys, Arg, Asn, N terminus |
| J    | 57.021464 | +CH₂NO | Carbamidomethyl | Cys, Lys, His, Asp, Glu, N terminus |

List of the most abundant modifications found in human A431 cell lysate and denoted in Fig. 2

Modification types are derived from UniMod (12).

| Elemental composition | Modification types | Amino acid |
|-----------------------|--------------------|-----------|
| CH₄S                  | Loss from internal methionine | Met |
| NH₃                   | Oxobutanoic acid from N-terminal Thr Pyro-Glu from Gln | Thr, Gin |
|                       | Pyruvic acid from N-terminal Ser S-Carboxymethylcysteine cyclization (N-terminus) | Ser, N terminus |
| −H₂                   | Didehydro Threonine oxidation to 2-amino-3-oxo butanoic acid | Ser, Tyr |
| +O                    | Deamidation | Asn, Gin |
| +CH₂                  | Methylation Methyly ester | Cys, His, Lys, Asn, Gln, Arg, Ile, Leu, N terminus |
| +O                    | Oxidation | Met, His, Trp |
| +CO                   | Formylation | Lys, Ser, Thr, N terminus |
| +CH₄                  | Ethylation | Lys, Glu, N terminus |
| +O₂                   | Dimethylation | Lys, Arg, Asn, N terminus |
| +CH₂NO                | Carbamidomethyl | Cys, Lys, His, Asp, Glu, N terminus |

**TABLE I**

List of the most abundant modifications found in human A431 cell lysate and denoted in Fig. 2

Modification types are derived from UniMod (12).
The detected modifications could be confined to a relatively few base peptides that are for some reasons prone to modifications or could be more or less homogeneously spread over many base peptides. To find out the actual situation, the modification multiplicity values (number of different dependent peptides per unique base peptide) were calculated for every base peptide. There were 824 cases with multiplicity 1, 36 with multiplicity 2, and four cases with multiplicity 3. As expected, the "one peptide-one modification" model dominated. Analysis showed that the base peptides with the highest multiplicity values came from the most abundant proteins.

**Efficiency**—Identification of each base/dependent pair of peptides means that another MS/MS dataset that has previously been ignored is now explained. Thus the total efficiency of the proteomic analysis increases by 3.5%, an increase comparable with the efficiency of full-length de novo sequencing. Because of the differences in the extent of modification, this increment depended upon the analyzed organism (see below) and was higher for complex organisms (humans) than for primitive ones (*E. coli*).

The efficiency of ModifiComb in terms of false negatives (misses of true, present modifications) was tested for methylations. A Mascot search with methylation as variable modifications that took much longer than the ModifiComb run produced the same number of hits and did not reveal any new methylated peptide.

**The Role of ΔRT**—The retention time histogram resembles that of ΔM (data not shown), but the RT resolving power of nanoflow HPLC is much lower than the mass resolution of the FTMS instrument. Nevertheless this resolution is often suffi-
cient to provide fine details on the position of modification. For example, Fig. 4a shows at least two peaks separated on the ΔRT scale, both corresponding to the same ΔM value of 27.995 Da (formylation). The first peak, at approximately +180 scans, corresponds to formylation of a side chain of serine and threonine (70% of peptides contributing to the peak carry this modification), whereas the peak at ≈+600 scans corresponds to modification of the N terminus (85% of the contributing peptides). The methylation peak (Fig. 4b) also features two components, a sharp peak at ≈50 scans and a smaller, broader peak at ≈300 scans. The heterogeneity of methylation is considered below in more detail.

**Heterogeneous Modifications: the Case of Methylation—**

Proteins undergo post-translational methylation at nucleophilic side chains, most often on nitrogen and oxygen atoms. O-Methylation reduces the negative charge of glutamic and aspartic acids by creating a methyl ester on their carboxylate side chain and thereby increases the hydrophobicity. The lifetime of this modification varies with hydrolysis back to Glu and Asp as the outcome. N-Methylations do not alter the charge but increase the hydrophobicity. The modification can take place at the ε-amine on lysine, the imidazole group of histidines, the guanidine moiety of arginine, and the side-chain nitrogen of glutamine and asparagines. It is a permanent modification not readily reversible under physiological conditions.

The fact that different methylation sites form clearly separated peaks in the ΔRT spectrum (Fig. 4b) can be explored to study the heterogeneity of methylation sites. As an example of such a study, Fig. 5 shows the case of a base peptide, TATPQQAEVHEK, identified from an A431 human carcinoma cell lysate through Mascot search with a very high M-score of 76. Manual validation of both CAD and ECD spectra (Fig. 5A and inset therein) confirmed the identification, revealing complete cleavage coverage of the sequence. The peptide was detected at RT = 36.95 min, and the ModifiComb found two modified forms of this peptide eluting ≈3.5 and 5.0 min later.
(Fig. 5B) with the same $\Delta M = +14.015$ Da corresponding to methylation. From the extracted ion chromatogram in Fig. 5B, the abundance ratio between these two species is 1:0.8. The $\Delta RT$ values (+240 and +345 scans, respectively) correspond to different components of the right-hand tail of the $\Delta RT$ distribution in Fig. 4b. Manual inspection of the CAD and ECD spectra of these two dependent peptides (Fig. 5, C, D, and insets therein) located the positions of methylation at Glu$^{186}$ and His$^{185}$, respectively. Note that ECD was instrumental in locating these modifications due to abundant c$\text{11}$ and c$\text{12}$ ions (b$\text{11}$ and especially b$\text{12}$ ions in CAD were unreliable because of the poor signal/noise and the presence of losses from them that made interpretation equivocal). It is hardly a coincidence that adjacent amino acids so different in their chemical properties are methylated, whereas no sign of modification is found on the proximate Glu$^{183}$ residue.

**Two-dimensional Map**—The ($\Delta M$, $\Delta RT$) dots can be put on a two-dimensional map to provide a total overview of the modification state of the sample. Such a ($\Delta M$, $\Delta RT$) map is shown in Fig. 6a. The $\Delta M$ scale consists of 5-mDa wide channels; to simplify the map and highlight the most abundant modifications, only channels containing at least two counts are displayed. The map contains a total of 908 dots corresponding to 45 $\Delta M$ channels. The $\Delta RT$ trends for each modification are clearly discernable.

**Human Versus E. coli**—To appreciate the modification complexity (M-complexity) of the human sample, compare Fig. 6a with a similar map created for E. coli samples and displayed in Fig. 6b. There are 53 dots here in 12 $\Delta M$ channels, a clear testimony to the much lower M-complexity of the bacterium.

On the map, the E. coli modification pattern may resemble a reduced human pattern, but this resemblance is superficial. Fig. 7a shows the $\Delta M$ plots of both samples in comparison from which their differences are apparent. The product/mom-ent correlation (Pearson correlation (30)) analysis confirms the dissimilarity of these two patterns ($r = 0.45$).

**Repeatability of $\Delta M$ Analysis**—The obtained low value of $r$ can be compared with the correlation obtained between two independently prepared and analyzed E. coli samples (Fig. 7b) as well as K562 human leukemia cell line samples (Fig. 7c), which were much better ($r = 0.87$ in both cases). This testifies to the high repeatability of $\Delta M$ analysis. Note that for the high repeatability of $\Delta M$ analysis it is not essential that the instrument picks up exactly the same pairs of peptides in each analysis nor is the difference in absolute retention times of peptides important. Of course, at different sample loads the signal to noise ratio in $\Delta M$ plots will change, and some peaks may change their abundance or even disappear if they are small. This, however, is of lesser importance for the correlation analysis as the product-moment correlation factor $r$ picks up the overall similarity between the patterns.

**Human Versus Human**—Fig. 7d presents a comparison between different human cell lines, A431 and K562. The obtained correlation $r = 0.49$ is just a little higher than that between human and E. coli samples, indicating that ModifiComb $\Delta M$ analysis can be used for assessing modification states of the same organism. Detailed analysis of the differences in the modification states between these and other human cell lines will be reported separately. Here we just note that the two largest differences observed between the A431 and K562 cell lines are in the extent of methylation and loss of NH$_3$.

**Database-independent Search**—As already mentioned, identification of the sequence of the base peptide is not essential for the ModifiComb algorithm. To test the performance in the blind mode, the human A431 cell line $\Delta M$ histogram with base peptides obtained through Mascot search was compared with the histogram where peptide pairs were found through blind search in which the sequence of the base peptide was not known. As already mentioned, in such a search there is no difference between the base and dependent peptide, and because of that, absolute $\Delta M$ values were plotted. The obtained $\Delta M$ histogram was compared with the respectively modified $\Delta M$ histogram from Fig. 7a. To make the comparison fair, the oxygen peak F was removed from the blind search spectrum (this peak is mainly due to oxidized methionine and is not prominent in the database-dependent spectrum because the Mascot search did not include methio-
tion analysis confirms the high degree of similarity between the two patterns \((r = 0.97)\). Note that the similarity between the database search and blind search \(\Delta M\) spectra is much larger than that between two analyses of independently prepared samples of the same protein mixture. Thus both blind and open eyed methods can be used for fingerprinting the sample through \(\Delta M\) histograms or \((\Delta M, \Delta R T)\) maps.

**New E. coli Modification**—The most abundant bacterium modifications include those found in the human sample, loss of NH₃ (peak A), deamidation (peak B), formylation (peak C), and dihydroxy (peak D) but also a new modification listed in Table II. This modification (peak E) is located on histidine residues and gives a mass shift of +98.073 Da corresponding to the chemical formula \(\text{C}_6\text{H}_{10}\text{O}\). This modification was not found in the literature; its actual structure is not known. Because this modification was found only in \(E. coli\) samples, it is more likely to occur in vivo than in vitro.

**Novel Modification in Human Saliva**—Human saliva is a very accessible sample; yet it is also a very interesting one for M-fingerprinting due to the presence there of many antimicrobial peptides that can undergo modifications during their contacts with microorganisms (31). A partial \(\Delta M\) histogram of the saliva peptide sample is shown in Fig. 9. The most abundant detected modification corresponds to \(\Delta M = +12,000\) Da, an unexpected value not found among common modifications (12). The mass shift can only correspond to a carbon atom. One of the contributing peptide pairs to that modification has already been shown in Fig. 1; the modification is located on a proline residue. Here we discuss the possible biological implications of this finding stemming from the fact that at least six base peptides related to this modification were identified as belonging to proline-rich proteins (PRPs). One of the primary functions of saliva PRPs in mammals is to precipitate tannins, polyphenolic compounds commonly found in certain beverages, fruits, and berries. Tannins exhibit a variety of harmful effects ranging from reduction of the nutritional value of food to causing esophageal cancer (32). Strong binding of tannins to PRPs is believed to be the first line of defense against these harmful compounds (33, 34). Transformation of the proline pyrrole five-member ring into a pyrrolidine six-member ring (\(+12,000\) Da) makes the modified endogenous saliva peptides slightly more hydrophobic than the unmodi-
fied counterparts. Although the base and +12.000-Da peptides are shown in the same integrated spectrum in Fig. 1A, the chromatographic peak of the dependent peptide was delayed by 9 min compared with the base peptide. Because the main interaction between the tannins and PRPs is the hydrophobic stacking of the phenolic ring of a polyphenol over the pyrrole rings of the Pro residues (33, 35), an increase in the hydrophobicity of the latter makes the stacking interaction stronger, which may improve the defense against tannins. This modification may be similar in structure to the rare amino acid baikiain (36), which has not been reported previously in humans.

To additionally validate this modification, correlation analysis was performed between the abundances of 12 identified fragments in mass spectra presented in Fig. 1, B and C. Because the assumed proline modification is hydrophobic, its influence on the fragment abundances was expected to be small. If, on the other hand, the spectra belonged to alto-

**TABLE II**

| Peak | ΔM     | Elemental composition | Modification type                                      | Amino acid          |
|------|--------|-----------------------|------------------------------------------------------|---------------------|
| A    | -17.026549 | -NH₃                  | Oxobutanoic acid from N-terminal Thr                  | Thr                 |
|      |        |                       | Pyro-Glu from Gln                                     | Gln                 |
|      |        |                       | Pyruvic acid from N-terminal Ser                      | Ser                 |
|      |        |                       | S-Carboxamidomethylcysteine cyclization (N terminus)  | N terminus          |
| B    | 0.984016 | +O                    | Deamidation                                           | Asn, Gin            |
|      |        | −NH                   |                                                      |                     |
| C    | 27.994915 | +CO                   | Formylation                                           | Lys, Ser, Thr, N terminus |
| D    | 31.986829 | +O₂                   | Dihydroxy                                             | Pro, Arg, Lys, Phe  |
| E    | 43.005814 | +CHNO                 | Carbamylation                                         | Lys, Arg, Cys, N terminus |
| F    | 98.073165 | +C₆H₁₀O              | Unknown                                               | His                 |

**Fig. 9.** The distribution of ΔM values in the region from −100 to 100 Da (resolution, 0.005 Da) for the human saliva sample. The inset shows magnification of the 11.900–12.100-Da region.
gether different peptides, little correlation was expected between the abundances. The obtained correlation factor \( r = 0.88 \) left no doubt that the peptides in Fig. 1, B and C, are closely related.

**CONCLUSION AND DISCUSSION**

A new software tool has been created that utilizes correlations in peptide fragmentation to reveal PTM patterns of complex biological samples. Using database-dependent identifications with common proteomic search engines or performing blind searches yields numerous known and novel peptide modifications. A high resolution \( \Delta M \) histogram reveals the overall modification pattern. Additionally \( \Delta RT \) information confirms the nature of the modifications. A two-dimensional (\( \Delta M, \Delta RT \)) map demonstrates repeatable features for the same biological sample independent from the method of map generation (database or blind search). On the other hand, the maps are different for different organisms and samples, which may be used for comparative proteomic work and searching for biomarkers.

The (\( \Delta M, \Delta RT \)) map immediately reveals the PTMs present in the sample without making a priori assumptions of their chemical composition or the site of attachment. All abundant present modifications are detected, including ones that are unexpected and novel. The sensitivity of ModifiComb should be higher than that of the database search with variable modifications as ModifiComb separates the function of the peptide identification from the function of the PTM assignment and is usually satisfied with lower information content from MS/MS data of modified peptides (at least four fragments) than is required for reliable database identification of unmodified peptides (at least six to seven fragments). The speed of ModifiComb is much faster than that of any database search engine. The information in the PTM spectrum derived from the sample can be used to minimize the number of modifications allowed as variable parameters in the subsequent database search thereby speeding it up as well as reducing the rate of false positive and false negative identifications.

**Acknowledgments**—Magnus Molin from the Department of Genetics and Pathology at Uppsala University is acknowledged for sample preparation of the K562 human chronic leukemia cell line. Additionally we thank Frank Kjeldsen, Thomas Köcher, Oleg Silivra, Chris Adams, and Alexander Misharin for fruitful discussions.

* This work was supported by the Knut and Alice Wallenberg Foundation and Wallenberg Consortium North Grant WCN2003-UU/SLLU-009 (to R. Z.) as well as Swedish Research Council Grants 621-2004-4897 and 621-2003-4877 (to R. Z.). 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.

‡ Both authors contributed equally to this work.

§ To whom correspondence should be addressed: Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Box 583, S-75123 Uppsala, Sweden. Tel.: 46-18-471-5729; Fax: 46-18-471-5729; E-mail: Mikhail.Savitski@bmms.uu.se.

**REFERENCES**

1. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. Nat. Biotechnol. 21, 255–261.
2. Jensen, O. N. (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. 8, 33–41.
3. Gruler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol. Cell. Proteomics 4, 310–327.
4. Bearssole, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J. X., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc. Natl. Acad. Sci. U. S. A. 101, 12130–12135.
5. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequence in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989.
6. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence data-bases using mass spectrometry data. Electrophoresis 20, 3551–3567.
7. Ong, S. E., Mittler, G., and Mann, M. (2004) Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. Nat. Methods 1, 119–126.
8. Shevchenko, A., Loboda, A., Eks, W., and Stading, K. G. (2000) MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteome research. Anal. Chem. 72, 2132–2141.
9. Keller, A., Nesvizhskii, A. l., 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.
10. 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.
11. 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 Mouse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. Mol. Cell. Proteomics 4, 1180–1188.
12. Creasy, D. M., and Cottrell, J. S. (2004) Unimod: protein modifications for mass spectrometry. Proteomics 4, 1534–1536.
13. Tannen, S., Shu, H. J., Frank, A., Wang, L. C., Zandi, E., Mumbry, M., Pevzner, P. A., and Bafna, V. (2005) InsPect: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. 77, 4626–4639.
14. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun. Mass Spectrom. 17, 2310–2316.
15. Mann, M., and Wilm, M. (1994) Error tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. 66, 4390–4399.
16. Zhang, N., Li, X.-j., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. (2005) ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics 5, 4096–4106.
17. Tsur, D., Tannen, 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.
18. 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.
19. Wilm, M., Shevchenko, A., Houthaev, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1999) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature 379, 466–469.
20. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Tryptic cleaves exclusively C-terminal to arginine and lysine residues. Mol. Cell. Proteomics 3, 608–614.
21. Chelius, D., Rehder, D. S., and Bondarenko, P. V. (2005) Identification and characterization of deamination sites in the conserved regions of human immunoglobulin \( \gamma \) antibodies. Anal. Chem. 77, 6004–6011.
22. Kary, J. A., and Reilly, J. P. (2005) Deamidation as a consequence of \( \beta \)-elimination of phosphopeptides. Anal. Chem. 77, 4673–4676.

**Molecular & Cellular Proteomics 5.5** 947
23. 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
24. Huang, L. H., Li, J. R., Wroblewski, V. J., Beals, J. M., and Riggin, R. M. (2005) In vivo deamidation characterization of monoclonal antibody by LC/MS/MS. Anal. Chem. 77, 1432–1439
25. Ambler, R. P., and Rees, M. W. (1959) ϵ-N-Methyl-lysine in bacterial flagellar protein. Nature 184, 56–57
26. Bedford, M. T., Frankel, A., Yaffe, M. B., Clarke, S., Leder, P., and Richard, S. (2000) Arginine methylation inhibits the binding of proline-rich ligands to Src homology 3, but not WW, domains. J. Biol. Chem. 275, 16030–16036
27. 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
28. Gross, E., and Witkop, B. (1961) Selective cleavage of the methionyl peptide bond in ribonuclease with cyanogen bromide. J. Am. Chem. Soc. 83, 1510–1511
29. Stadtman, E. R. (1993) Oxidation of free amino acids and amino acid residues in proteins by radiolysis and by metal-catalyzed reactions. Annu. Rev. Biochem. 62, 797–821
30. Pearson, K. (1896) Mathematical contributions to the theory of evolution: III. Regression, heredity, and panmixia. Philos. Trans. R. Soc. Lond. 187, 253–318
31. Zasloff, M. (2002) Antimicrobial peptides of multicellular organisms. Nature 415, 389–395
32. Warner, T. F., and Azen, E. A. (1988) Tannins, salivary proline-rich proteins and esophageal cancer. Med. Hypotheses 26, 99–102
33. Baxter, N. J., Lilley, T. H., Haslam, E., and Williamson, M. P. (1997) Multiple interactions between polyphenols and a salivary proline-rich protein repeat result in complexation and precipitation. Biochemistry 36, 5566–5577
34. Mehansho, H., Butler, L. G., and Carlson, D. M. (1987) Dietary tannins and salivary proline-rich proteins—interactions, induction, and defense mechanisms. Annu. Rev. Nutr. 7, 423–440
35. Charlton, A. J., Baxter, N. J., Khan, M. L., Moir, A. J., Haslam, E., Davies, A. P., and Williamson, M. P. (2002) Polyphenol/peptide binding and precipitation. J. Agric. Food Chem. 50, 1593–1601
36. King, F. E., King, T. J., and Warwick, A. J. (1950) The chemistry of extractives from hardwoods. Part III. Baikiala, an amino acid present in Baikiaea plurijuga. J. Chem. Soc. 3590–3597