The mass distance fingerprint: a statistical framework for *de novo* detection of predominant modifications using high-accuracy mass spectrometry

Frank Potthast [a,*], Bertran Gerrits [a], Jari Häkkinen [b], Dorothea Rutishauser [c], Christian H. Ahrens [a], Bernd Roschitzki [a], Katja Baerenfaller [d], Richard P. Munton [e], Pascal Walther [f], Peter Gehrig [a], Philipp Seif [g], Peter H. Seeberger [g], and Ralph Schlapbach [a]

[a] *Functional Genomics Center Zürich, Uni/ETH Zürich, Switzerland*

[b] *Computational Biology, Theoretical Physics, Lund University, Sweden*

[c] *Institute of Neuropathology, University Hospital Zürich, Switzerland*

[d] *Institute of Molecular Cancer Research, University of Zürich, Switzerland*

[e] *Institute of Cell Biology, ETH Zürich, Switzerland*

[f] *Institute of Biochemistry, University of Zürich, Switzerland*

[g] *Laboratory for Organic Chemistry, ETH Zürich, Switzerland*

Abstract

We describe a statistical measure, Mass Distance Fingerprint, for automatic *de novo* detection of predominant peptide mass distances, *i.e.,* putative protein modifications. The method's focus is to globally detect mass differences, not to assign peptide sequences or modifications to individual spectra. The Mass Distance Fingerprint is calculated from high accuracy measured peptide masses. For the data sets used in this study, known mass differences are detected at electron mass accuracy or better. The proposed method is novel because it works independently of protein sequence databases and without any prior knowledge about modifications. Both modified and unmodified peptides have to be present in the sample to be detected. The method can be used for automated detection of chemical/post-translational modifications, quality control of experiments and labelling approaches, and to control the modification settings of protein identification tools. The algorithm is implemented as a web application and is distributed as open source software.

Key words: Mass distance fingerprint, Mass distance histogram, Post-translational modification, Protein identification

1 Introduction

In proteomics, high throughput approaches using mass spectrometry have become widely used. These approaches promise to enable researchers to assess, on a large scale, both expression level and functional state of the proteins that carry out most functions in a cell. The success of proteomics experiments, such as studies of protein function and cell signaling pathways, ultimately depends on how well the protein content in samples is identified and annotated. Consequently, a lot of effort is put into identifying the constituent proteins using mass spectrometric methods. The goal is to assign acquired spectra to known peptide sequences and potential co- and post-translational modifications. To this end database search engines were rapidly developed after the introduction of ionization techniques for biological mass spectrometry [1–5]. These approaches depend on sequence databases that are used by the engines to match real spectra to theoretical *in silico* spectra. The matching is complicated by the fact that there are protein modifications and the sequence databases store the unmodified sequences. To resolve this, the researcher typically defines a small set of modifications for inclusion in the matching process. But due to combinatorial explosion, the usage of a large number of variable modifications is inherently difficult, if not impossible, in these approaches. A related alternative approach is error tolerant searching [6–8] that considers a multitude of modifications or mutations.

The need to keep track of protein modifications is readily recognized by the proteomics community, and few repositories of known peptide modifications have been created. The RESID database [9] lists co- and post-translational modifications. Post-translational modifications (PTMs) are also stored in Delta Mass [10] together with information on modifications induced by sample preparation procedures for mass spectrometric analysis, but mass changes are only given as integer values. FindMod [11] also lists some modifications and detects PTMs from this list in conjunction with a protein sequence and a few precursor masses. The most comprehensive collection of chemical and biological modifications being relevant to mass spectrometry can be found in UniMod [12]. The list of known protein modifications is growing; in December 2006, UniMod lists 495 modifications including 144 amino acid substitutions [6].

* Corresponding author. Functional Genomics Center Zurich, Uni-ETH Zurich, Winterthurer Strasse 190, 8057 Zurich, Switzerland. Fax: +41 44 635 39 22.
Email address: peptoscope@fgcz.ethz.ch (Frank Potthast).

The focus of the method presented here, Mass Distance Fingerprint (MDF), is to globally assess predominant precursor mass distances, *i.e.,* finding dominant PTMs in a data set; it is not aimed for assigning peptide sequences or modifications to individual spectra. The MDF is limited to the detection of frequent precursor mass distances and will not detect low abundance modifications. The method to calculate the MDF of a data set has three stages. First, the Mass Distance Histogram (MDH) is calculated. Second, a statistical random background model, also reported in this paper, is subtracted from the experimentally observed MDH. Third, Gaussian distributions are fitted to the remaining signal for accurate determination of mass distances. The resulting list of frequent mass distances and related information is then the Mass Distance Fingerprint.

The novelty of MDF is its independence of both sequence databases and of prior knowledge about modifications, since it uses only precursor mass information. In MDF, MS/MS level data is not used. Approaches that use both MS/MS level and sequence information exist. In the P-mod algorithm [13], MS/MS spectra are compared with *in silico* generated spectra using sequence information provided to the algorithm. In contrast, in MDF the aim is to detect modifications *de novo*. The ICATcher [14] and ModifiComb [15] algorithms, like P-mod, use MS/MS information but work independently of sequence information. In contrast to P-mod, ICATcher relies on both the modified and unmodified peptide being measured. The latter dependency is also valid for the MDF presented here.

The MDF has a different conceptual focus than the methods mentioned above; the MDF provides a measure of modifications on the level of a collection of MS spectra, *i.e.,* MDF is not applicable directly on single level spectra. Extending the MDF to single level spectra is possible with existing technologies comparing MS/MS spectra [14].

Both the MDH and the MDF are implemented as an algorithmic framework called Peptoscope. The Peptoscope source code is distributed under the GPL license version 2 [16].

In essence Peptoscope needs the peptide masses to calculate the MDH. These masses are fed into Peptoscope using is the widely used Mascot Generic File (mgf) [17]. Conveniently, most commercial mass spectrometer software are capable of generating such mgf formatted files and the mgf format is usually used as input to database interrogation software. Of the available data in the mgf, Peptoscope uses the charge and m/z information to determine the peptide masses. Peptoscope output is a list of detected predominant mass distances and includes annotation with known modifications if applicable. However, no prior modification information is used in the calculation of modifications present in the data.

2 Experimental details

Four experimental data sets are used to illustrate the applicability of the MDF for *de novo* detection of chemical and post-translational modifications. The four data sets are published as supplemental material [18].

The tryptic peptide content of the four experiments was separated and analyzed by LC-ESI-MS/MS on a "Finnigan LTQ-FT" (Thermo Electron, Bremen, Germany), a hybrid instrument consisting of a linear ion trap and a Fourier transform ion cyclotron resonance mass spectrometer. As in [19,14], only doubly charged precursor masses were considered for all data sets.

Each data set contains one or more LC runs; each run contains hundreds if not thousands of precursor masses. The data sets used represent the typical variation of proteomic data in terms of both experimental setup as well as the amount of data to be analysed, *i.e.,* the number of tandem mass spectra. The exact number of precursor masses used in one analysis can be found in the Peptoscope output. The experimental details are given for each data set in the results section.

2.1 Data set 1

Data set 1 is derived from an *in vitro* study of the human DNA mismatch repair system. A DNA affinity matrix (DynaBeads derivatised with heteroduplex DNA containing an insertion/deletion mismatch) was incubated with HeLa cell nuclear extract for either 5 (sample 1) or 25 minutes (sample 2) at 25^o C. Proteins that bound to this matrix were subsequently eluted, reduced and labeled with the heavy- and light-cleavable ICAT (isotope-coded affinity tag) reagent [20,21] (Applied Biosystems, Foster City, CA, USA), respectively. The differentially labeled samples were combined and digested with trypsin (Sequencing Grade Modified Trypsin, Promega, Madison, WI, USA) at 37^o C for 24 hours. Peptides were first purified with a cation exchange column (ICAT Cation-Exchange Cartridge, Applied Biosystems) and ICAT-labeled peptides were subsequently extracted with an Avidin affinity column (ICAT Cartridge Avidin, Applied Biosystems). The acid cleavage of the biotin tag and all the remaining steps were performed according to the manufacturer's instructions. Sep-Pak columns (Vac C18 1cc, 50 mg, Waters, Milford, MA, USA) were used for further clean up of the affinity-purified fraction. After mass spectrometric analysis and data processing this data set was analysed by Peptoscope using 4199 precursor masses.

2.2 Data set 2

Data set 2 originated from mouse cortical synaptosomes and was obtained as follows: mouse cortical synaptosomes were prepared by differential centrifugation and sucrose density gradient fractionation as previously described [22]. Synaptic proteins were cleaned by acetone precipitation and solubilised in 7 M urea, 50 mM ammonium carbonate pH 7.8 before cysteine reduction and alkylation. Tryptic digest was performed overnight at 37^oC with a final urea concentration of 1.5 M, and an enzyme to protein ratio of 1:25. The resulting peptide mixture was acidified to pH < 3 with acetic acid containing 25% acetonitrile, centrifuged to remove insoluble matter before fractionation using a polySULFOETHYL A strong cation exchange chromatography HPLC column (PolyLC, USA). After lyophilisation, peptide fractions were desalted using reverse phase trap cartridges. Finally 819 precursor masses were used for Peptoscope analysis.

2.3 Data set 3

Data set 3 is derived from a mouse brain sample with background as follows. Neurotrypsin is a trypsin-like serine protease predominantly expressed in the peripheral and central nervous system (CNS) [23]. A truncating deletion in the human gene results in severe mental retardation [24]. To investigate the role of the proteolytic activity of neurotrypsin in the CNS we generated transgenic mice overexpressing neurotrypsin specifically in neurons starting at birth. In search for neurotrypsindependant changes in the neuronal network, hippocampi of wild-type (wt) and transgenic (tg) mice were prepared and the proteins were analyzed using the ICAT technology [20,21]. Hippocampus homogenate was subjected to two consecutive centrifugation steps each at 3000x g, separating nuclei and cell debris. Then, with a 34000x g centrifugation step S2 (soluble) and P2 (pellet) subcellular fractions were produced. S2 mainly consists of soluble proteins and light membrane particles, such as synaptic vesicles. P2 comprises heavy membrane particles including synaptosomes, Golgi apparatus, endoplasmatic reticulum, mitochondria and plasma membranes. The samples were treated as described in the protocol from Applied Biosystems *Cleavable ICAT Reagent Kit for Protein Labeling*. In brief, the proteins were denatured and reduced, followed by labeling with Cleavable ICAT Reagent by alkylation of free cysteines. The protein mixture was digested with trypsin and the complex sample was fractionated using a cation exchange column. The biotinylated peptides were subsequently purified on an avidin cartridge. Further sample treatment as in data set 1. The input for Peptoscope was an mgf file containing 16177 precursor masses.

2.4 Data set 4

Data set 4 is a standard protein mix of proteins as supplied by Applied Biosystems, treated with cleavable ICAT as data set 1. The mix consists of six proteins; bovine serum albumin (Swissprot accession number P02769), β-galactosidase (P00722), α-lactalbumin (P00711), β-lactoglobulin (P02754), lysozyme (P00698), and apotransferrin (P02787). For Peptoscope 912 precursor masses were used.

3 Methods

In this section, the Mass Distance Histogram (MDH) is defined, and a statistical model for the MDH is developed. The statistical model is derived from a simulation of random peptides. The Mass Distance Fingerprint (MDF) is derived from the MDH and the section is concluded with an illustration of the MDF using data set 4.

3.1 The Mass Distance Histogram and the Mass Distance Fingerprint: definition, simulation, model

3.1.1 Mass Distance Histogram: Definition

Given a set of mass spectrometric measurements, the MDH is defined as the distribution of distances between all possible pairs of measured peptide masses. For a number of n masses, there are $n(n-1)/2$ mass distance pairs. For the purpose of this study, Peptoscope analysis was restricted for mass differences up to 100 Da, and a bin size of 0.01 Da was used in the MDH. As an example, data set 1 consists of 4199 precursors yielding 8813701 pairs of which 1791599 have a mass distance in the histogram range.

3.1.2 Simulation and model

To investigate the expected random distribution $R(\Delta m)$ of an MDH, an arbitrary number of 100 million peptide pairs were generated randomly in a computer simulation. The mass distribution for these random peptides follow the natural distribution of peptide masses and for each of these peptide pairs, the mass distance $\Delta m = |m_1 - m_2|$ was calculated. For these mass distances, a histogram $MDH(\Delta m)$ was generated in the range between 0 to 100 Da (see Figure .1). The regular structure of the simulated distribution can be

modeled well by a sum of Gaussian distributions with distance δ and constant width σ_R:

$$MDH(\Delta m) \approx R(\Delta m, \sigma_R) = \frac{1}{\kappa\sigma_R\sqrt{2\pi}} \sum_{i=0}^{\kappa} \exp - \frac{(\Delta m - i \cdot \delta)^2}{2\sigma_R^2}. \quad (1)$$

Similar results are found by performing a random pairing of peptides emanating from an *in silico* tryptic digest of any protein sequence database and using these peptide pair mass differences as the background R. This is a result of the fact that a tryptic digest will yield a peptide mass distribution that is a sum of Gaussian like distributions with centres approximately separated at integer Da values with empty regions between the Gaussians. Therefore the background is best described as the distribution of masses of unmodified random peptide pairs. Peptoscope is relying on a superposition of the measured modified and unmodified peptides, where the modifications will be additional Gaussians on top of the random background. The modification induced Gaussians are much narrower than the background distribution of Gaussians.

The central statement of Equation 1 is that for all nominal mass differences, the fitting Gaussian curves have the same width and the same height with high predictive power. The factor in front of the sum in Equation 1 ensures that $R(\Delta m)$ behaves like a probability measure,

$$\int_0^{\kappa} R(\Delta m, \sigma_R)\, d(\Delta m) = 1. \quad (2)$$

In this study, κ is 100 Da as a consequence of the mass range choice of 0 to 100 Da in the MDH.

The δ in Equation 1 originates from the fact that true peptide masses are distributed in clusters with a mean value of roughly $i \cdot 1.000458$ Da where i is an integer [25,26]. For the mass differences dealt with here, it is found to be approximately 1.00044 Da. The model has a single parameter, σ_R, which is obtained from a measured $MDH(\Delta m)$ by minimizing the square deviation, $E(\sigma_R)$, between the background model and $MDH(\Delta m)$:

$$E(\sigma_R) = \int_0^{\kappa} [MDH(\Delta m) - R(\Delta m, \sigma_R)]^2\, d(\Delta m). \quad (3)$$

Inserting σ_R, obtained by minimizing $E(\sigma_R)$ of Equation 3, into Equation 1 yields the background model $R(\Delta m)$. In Figure .1, both the simulated $MDH(\Delta)$ and the model $R(\Delta m)$ are shown. The overall similarity is good and using Peptoscope, the user can and should visually inspect the quality of the background

model fit compared to the measured MDH.

It should be noted that σ_R must be fitted for each experiment individually, it is not a universal constant. To appreciate this statement we have to delve into how the background masses are built up. All peptide masses are built up by a composition of electron, proton, and neutron masses. This means that for any given mass there are many random ways to compose a molecule that will be close to that mass. The resulting masses typically do not match other masses exactly; a distribution at approximately integer masses (in Da) is built up. This natural width at integer masses will become larger for increasing mass just by the fact that there are so many more combinations contributing to a specific heavy mass (neighbourhood) as compared to a specific light mass. The variation in width will of course be visible also when mass differences are studied.

To further test the MDH background model, we artificially digested a number of protein databases using the Perl script fasta2MDH.pl (obtainable from the authors by email). The script fasta2MDH.pl calculates an MDH using a protein FastA file as input; it uses the range 0-100 Da with 10000 bins, exactly as used in this paper. We found that the statistical description outlined in this section is valid also for this type of artificial data that describes the underlying experimental background well (data not shown).

Obviously we do not have a completely random distribution of peptides since we are measuring a specific composition of proteins. This is actually the important idea of Peptoscope, the non-random component of the measured mass distributions will be on top of the random background noise. However, in any given experiment we do not know whether the background is build up by light or heavy masses but we know that the background will vary between experiments (and be a mix of heavy and light masses). This is the reason to introduce the fitting parameter σ_R – to compensate for the unknown background. From simulations of the background we learn that σ_R varies between 0.1 and 0.4 Da up to 6000 Da (data not shown).

As mentioned above, a real measurement is of course expected to contain more information than a simple random background signal. Depending on the experimental details, a number of modifications are likely to be present, and sometimes both the modified and unmodified form of a peptide will be measured. Thus, a better approximation of a real measurement would be an extension of Equation 1, where the effect of the background $R(\Delta m)$ and Gaussian signals induced by modification mass shifts Δm_j are summed:

$$MDH(\Delta m) \approx R(\Delta m) + \sum_{j=1}^{\#mod} \frac{s_j}{\sigma_j\sqrt{2\pi}} \exp -\frac{(\Delta m - \Delta m_j)^2}{2\sigma_j^2}. \quad (4)$$

For each modification term j in Equation 4, three Gaussian parameters Δm_j, s_j, and σ_j are obtained by minimizing the deviation between the model and the experimental $MDH(\Delta m)$ in the vicinity of Δm_j. Here the fitting parameters corresponds to the mass distance, Δm_j, the intensity of the signal, s_j, and the width of the fitted peptide peak, σ_j.

3.1.3 The Mass Distance Fingerprint

The MDF for an experiment contains several triplets, Δm_j, s_j, and σ_j, as obtained from Equation 4; triplets with $s_j > 1/3 * R(\Delta_m)$ are reported. Two numbers extend each triplet of the MDF; The first number extending the MDF is the *estimated number of true pairs*: the core of the approach is to think of the Gaussians described by m_j, s_j, and σ_j as originating from modification induced effects. The corresponding area of the Gaussian is calculated from σ_j and s_j. Using the total number of pairs under the curve, this can be expressed in terms of the *estimated number of true pairs*. The second number extending the MDF is the estimated $\pm 2\sigma$ true positive classification rate. The latter two numbers are illustrated in Figure .2.

Within the Peptoscope framework, the MDF is further annotated with mass differences induced by known peptide modifications (retrieved from UniMod [12]). When the measured Δm_j is close to a known PTM induced mass change, this is reported; a list of known isobaric modifications at that Δm_j is given together with the mass deviations from the listed mass difference.

MDF can not distinguish isobaric modifications, nor does it state if a modification adds or subtracts mass from the peptide, and both the unmodified and modified form of a peptide have to be measured for the corresponding modification to be included in the MDF.

3.2 Illustration: Experimental Results

This section illustrates above concepts with results from data set 4 where 912 precursor peaks were examined. 912 precursors correspond to 415416 possible pairs, and of these pairs, 52478 precursor mass distances (Δm) are in the range between 0 and 100 Da. Minimizing Equation 3, the optimal σ_R in Equation 1 for this data set was found to be 0.055 Da. The measured $MDH(\Delta m)$, and the background model $R(\Delta m)$ are displayed in Figure .3. There are two sources of deviations between the model and the measurement. The first deviation is that the $MDH(\Delta m)$ is not close to zero in the regions between the nominal masses differences where $R(\Delta m)$ is close to zero. The effect of this minor discrepancy is negligible for the MDF, and the source of the difference may be an effect of erroneous precursor charge determinations or noise from the

instrument. The second deviation is sharp Gaussian peaks of which some are close to known modification mass differences. This second effect is addressed by the modification terms in Equation 4.

To check whether the sharp peaks on top of $R(\Delta m)$ are likely to correspond to PTMs, Gaussian distributions were fitted to the peaks as described in Equation 4. While the binning of the $MDH(\Delta m)$ is performed with a bin width of 0.01 Da resulting in 10000 bins totally, the Gaussian fitting for the signal peaks is performed with 10 bins (each with an width of 0.0015 Da) for every signal peak. The fitting is achieved by minimizing the least mean square error between the Gaussian model and the signal peak; resulting in the three MDF parameters Δm_j, s_j, and σ_j for each signal peak. The background, $R(\Delta m)$, is subtracted from the measured $MDH(\Delta m)$ before the fitting procedure.

For the method to work it is necessary that σ_j is much smaller than σ_R, *i.e.,* the modification induced peaks must be much sharper than the underlying background distribution. The sharpness of the modification induced peaks, σ_j, is an effect of the intrinsic accuracy of the instrument in MS mode. If the mass accuracy is improved by some factor, σ_j would decrease by the same factor. As described in Section 3.1.2 the width of the background distribution, σ_R, is mainly determined by the length distribution of the measured peptides and is typically above 0.05 Da. Thus, the accuracy of the mass spectrometer should be somewhat below 0.01 Da for the method to be reliable. For peptides with masses between 1000 and 3000 Da, this translates to ppm measurement accuracy or better.

The above method yields the MDF for data set 4 presented in Table .1. Of the 16 signals, nine can be annotated with known modifications from UniMod. Eight of the nine annotated mass differences are closer than 0.0006 Da to the mass differences that would be the effect of the corresponding modifications acting on peptides. To set above number into numerical context, an electron has a mass of 0.000549 Da. The modifications used for annotating the MDF are taken from the UniMod web site http://www.unimod.org [12] (December 2006). In the mass range from 0–100 Da, there were 269 modifications corresponding to 152 different mass distances. The degeneracy is due to isobaric modifications. In the MDF report, the *PSI-MS Names* are preferred over the *Interim Name* from UniMod (*cf.* http://www.unimod.org). The complete list of UniMod modifications used in this study is contained in the supplemental material [18].

4 Results and discussion

In this section we present and discuss the MDFs obtained for the four data sets used in this study. A detailed validation of the results using MS/MS information is reported, and the importance of mass accuracy for Peptoscopes success is illustrated and quantified.

4.1 MDFs for the four data sets

Results for data set 1

The list of Peptoscope detected signals is shown in Table .2 for data set 1. This data set consists of two runs, a SIM scan and an MS survey scan producing 4199 precursors yield 8813701 pairs of which 1791500 fall in the range between 0 and 100 Da. σ_R of Equation 1 is found to be 0.105 Da. The two strongest signals belong to the repetitive monomeric unit of polyethyleneglycol (C_2H_4O) with a mono-isotopic weight of 44.0262 Da (*i.e.,* 88.0524 Da for a $C_2H_4O - C_2H_4O$ unit). The third signal is at 17.0265 corresponding to an elemental difference of (H_3N_1), most probably pyroglutamic acid formed from glutamine. This modification is frequently seen in protein samples measured in our laboratory. Three MDF signals, ++C2H7ON, ++71.02619, and 2*Ethanolyl, are modifications that have been detected repeatedly in our laboratory, however these three modification are not listed in UniMod. Each of the top five signals in Table .2 are estimated to correspond to more than 10000 pairs, with the $\pm 2\sigma$ true positive rate as illustrated in Figure .2 being above 90 percent. The next two signals in the MDF, at $\Delta m = 4.9554$ Da and $\Delta m = 26.9988$ Da, are not close to known modifications. The last signal in the list, $\Delta m = 21.9815$ Da is supposedly a sodium adduct of chemical composition $H(-1)Na$. All, except Asp↔His, MDF signals that could be annotated in this data set match the exact mass value with smaller deviation than the weight of an electron, 0.000549 Da.

In order to test the significance of the results, a Peptoscope analysis was performed for this data set, but with the mass accuracy being artificially decreased; Random mass shifts uniformly distributed in the range [-0.1,0.1] Da were added to all precursor masses, and nominal masses were shifted uniformly between 0 and 4 Da. Peptoscope was run on the resulting synthetic data set. No signal corresponding to known modifications, having a signal stronger than the weakest listed in Table .2, was found. Also, in the low mass accuracy LCQ (3D ion trap) data set mentioned below, no signals were found.

Results for data set 2

The MDF results for data set 2, originating from a single run, are the weakest of this study, as can be seen from Table .3. 819 precursors give 334971 pairs of which 101945 are in the MDH mass range. The optimal is $\sigma_R = 0.055$ Da. Three out of four predominant annotated mass shifts are found within sub-electron mass accuracy. The Gamma-carboxylation annotated mass distance is just 0.0001 Da off the true value.

Results for data set 3

Originating from 33 runs, data set 3 had by far the most doubly charged precursors, 16177. This results in more than 130 million pairs of which more than 19 million are within the MDH range. σ_R is 0.105 Da and ten out of eleven signals as listed in Table .4 were annotated. All signals except that in the guanidination are detected with sub-electron mass accuracy. Among the top signals, oxidation, ICAT, and pyroglutamic acid signals were found.

Computing time for this data set is roughly 30 minutes CPU time. Computationally, it is on the edge of what the current implementation of Peptoscope can handle on usual desktop computer equipment. However, the current implementation is not optimized for speed and can be modified to run significantly faster. Dealing with more than 100 runs for a single MDF is possible.

Results for data set 4

Data set 4 originates from a single run with ICAT-light and ICAT-heavy peptides being mixed at a ratio of 1:1. The MDF obtained with Peptoscope was derived from 912 doubly charged precursors *i.e.,* 415416 pairs of which 52478 were in the MDH mass range. The optimal σ_R was found to be 0.105 Da for this data set. As can be seen from Table .1, Peptoscope annotated nine of the sixteen MDF signals with known modification information. Mass accuracy is about electron mass or better. The mass shift at 25.0252 Da is not annotated, but has been repeatedly seen in Peptoscope analyses in our laboratory.

4.2 Validation of MDF results using MS/MS

The novel aspect of the MDF is that it uses peptide information only to detect predominant mass distances. Here, MS/MS information is used to validate and confirm MDF results. Detailed data of this validation can be found in the supplemental material [18].

MDF results are validated with the deltaMasses [27] program originating from the ICATcher algorithm [14]. deltaMasses compares MS/MS spectra against each other, looking for pairs of unmodified/modified peptides using a true statistical scoring scheme. For the validation of MDF reports, we used a deltaMasses probability cutoff $p < 0.00001$ *i.e.,* one would expect one false positive pair in 100000 detected pairs. deltaMasses does not detect pairs of MS/MS spectra below a mass difference of 4.5 Da. Therefore, low Δm MDF entries do not have a corresponding MS/MS value.

The question raised and answered by this validation is which part of the true pairs detected by the MDF are confirmed by a MS/MS validation test. Table .5 lists the number of detected pairs for both the MDF and the technology using MS/MS results. It was not possible to analyse data set 3 with the MS/MS approach because the software currently has an input limit of 10000 MS/MS spectra. The numbers in the table are of similar order of magnitude which underlines the validity of MDF results.

4.3 Importance of mass accuracy

To illustrate the importance of high mass accuracy, the Peptoscope algorithm was applied to an ICAT data set generated with a Thermo Electron LCQ (3D ion trap) mass spectrometer [14]. The Peptoscope distribution for this LCQ data set, having a precursor mass accuracy below 1 Da [14], is shown in Figure .4. Due to the low mass accuracy, Peptoscope is unable to detect any signal; the observed LCQ $MDH(\Delta m)$ is statistically seen flat while the high mass accuracy LTQ-FT MDH clearly follows the statistical model described by $R(\Delta m)$ and the model of Equation 4 (see Figure .3).

The importance of high mass accuracy is also illustrated in Figure .2. Imagine that the mass accuracy would be increased by a factor of two. The green curve would have half σ and double height while the light grey area would remain constant. However, the dark grey area (false positives) would decrease by a factor of 2. Thus, high mass accuracy is pivotal for the discriminative power of the method.

Both the unmodified and modified form of a peptide have to be measured for the corresponding modification to be detected. A possible strategy to detect complete modifications in a low complexity protein mix, with known protein identities, would be to synthesize the corresponding peptides. Measuring the synthesized peptides and performing a combined MDF would make it possible to also detect constant modifications with the MDF. In this context, the question arises what proportion of biologically meaningful PTMs will feature both modified and unmodified forms. Many modifications are known to be

reversible, for example phosphorylation of serine, threonine, and tyrosine, and acetylation of lysine. Since the reverse reactions are catalyzed by enzymes, the unmodified form of the peptides is usually present. In particular phosphorylations have low stoichiometries, *i.e.,* phosphorylated amino acids are generally less abundant than the corresponding nonphosphorylated residues [28]. Stable, irreversible protein modifications such as arginine methylation tend to be more complete, however, hypomethylated proteins are usually still detectable [29]. The different stoichiometry is one of the reasons why phosphorylation is not detected but methylation is.

5 Concluding remarks

A statistical model for the distribution of peptide mass distances has been presented; the corresponding histogram is called the Mass Distance Histogram (MDH). A model for the expected random background of the MDH is given in a form of Gaussian distributions. From this model for the MDH, the list of mass deviations is calculated; this is the mass distance fingerprint (MDF). Both the MDH and the MDF are calculated from precursor mass data only. The method depends on the use of mono-isotopic masses, *i.e.,* the knowledge of precursor charges. m/z values alone would be insufficient because wrong mass determination would blur signals of the MDH. No MS/MS or sequence information is used, and knowledge about known chemical or post-translational modifications is not required. Thus, the MDF is a true *de novo* PTM detection approach. It is shown that the entries of the MDF are frequently corresponding to weight shifts induced by known chemical or post-translational modifications of the peptides.

The detection of a modification on a single peptide measured only once seems difficult using the MDF approach. However, imagine a case where a peptide is measured 10 times, 5 times unmodified and 5 times modified. For this case, there are 25 pairs. If this is shifted to the situation of 9 unmodified and 1 modified, there are still 9 pairs. This way, the MDF compensates for the rare modification problem to some extent.

The nature of the MDF allows it to be used as a fast quality control of labelling approaches where both light and heavy form of a peptide should be predominantly present; Not detecting the expected mass shift would indicate a failure in the labelling procedure.

The MDH bin width used in this script was 0.01 Da. We found this to work well throughout our LTQ-FT experiments. For instruments like an Orbitrap, which can achieve better accuracy with about 1/2 ppm accuracy using lock spray calibration, it might be good to decrease the bin width accordingly.

The current version of Peptoscope works with simple histograms; if a true signal falls exactly in between two bins, the corresponding mass distance signal would be split up in two adjacent bins. More advanced binning approaches are possible but are not the scope of this work.

For the high accuracy precursor masses used in this study which were generated by an LTQ-FT, modifications are identified at electron mass accuracy or better. The MDH and the MDF are implemented in a framework called Peptoscope. Peptoscope provides an annotation of the *de novo* detected mass differences, this annotation is taken from the list of known modifications documented in UniMod. Peptoscope is implemented as a web application and is distributed as open source software under the GNU public license. In this study, only precursor masses were used. Obviously, the MDF is not limited to this special case, it can be applied to the collection of all MS signals irrespective of MS/MS measurements being performed or not. Obtaining an MDF with Peptoscope does not require any user interaction or parameters. Therefore, it is possible to automate MDF generation directly on the mass spectrometer instrument's computer. We believe that the MDF has the potential to become a standard technology on high accuracy mass spectrometers.

6 Acknowledgements

Thanks to John Cottrell, David Creasy, Hubert Rehrauer, and Mike Scott for helpful discussions. This work is in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

[1] J. K. Eng, A. L. McCormack, J. R. r. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J Am Soc Mass Spectrom 5 (11) (1994) 976–989.

[2] K. R. Clauser, P. Baker, A. L. Burlingame, Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching, Anal Chem 71 (14) (1999) 2871–2882.

[3] D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (18) (1999) 3551–3567.

[4] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (6928) (2003) 198–207.

[5] J. R. r. Yates, Mass spectrometry and the age of the proteome, J Mass Spectrom 33 (1) (1998) 1–19.

[6] D. M. Creasy, J. S. Cottrell, Error tolerant searching of uninterpreted tandem mass spectrometry data, Proteomics 2 (10) (2002) 1426–1434.

[7] C. L. Gatlin, J. K. Eng, S. T. Cross, J. C. Detter, J. R. r. Yates, Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry, Anal Chem 72 (4) (2000) 757–763.

[8] P. A. Pevzner, Z. Mulyukov, V. Dancik, C. L. Tang, Efficiency of database search for identification of mutated and modified proteins via mass spectrometry, Genome Res 11 (2) (2001) 290–299.

[9] J. S. Garavelli, The RESID Database of Protein Modifications: 2003 developments, Nucleic Acids Res 31 (1) (2003) 499–501.

[10] DeltaMass – Association of Biomolecular Resource Facilities, http://www.abrf.org/index.cfm/dm.home.

[11] M. R. Wilkins, E. Gasteiger, A. A. Gooley, B. R. Herbert, M. P. Molloy, P. A. Binz, K. Ou, J. C. Sanchez, A. Bairoch, K. L. Williams, D. F. Hochstrasser, High-throughput mass spectrometric discovery of protein post-translational modifications, J Mol Biol 289 (3) (1999) 645–657.

[12] D. M. Creasy, J. S. Cottrell, UniMod: Protein modifications for mass spectrometry, Proteomics 4 (6) (2004) 1534–1536.

[13] B. T. Hansen, S. W. Davey, A.-J. L. Ham, D. C. Liebler, P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data, J Proteome Res 4 (2) (2005) 358–368.

[14] F. Potthast, J. Ocenasek, D. Rutishauser, M. Pelikan, R. Schlapbach, Database independent detection of isotopically labeled MS/MS spectrum peptide pairs, J Chromatogr B Analyt Technol Biomed Life Sci 817 (2) (2005) 225–230.

[15] M. M. Savitski, M. L. Nielsen, R. A. Zubarev, ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures, Molecular and Cellular Proteomics 5 (5) (2006) 935–948.

[16] GNU General Public License, http://www.gnu.org/copyleft/gpl.html.

[17] mgf format reference, http://www.matrixscience.com/help/data_file_help.html.

[18] Available from http://www.peptoscope.ms.

[19] D. L. Tabb, A. Saraf, J. R. r. Yates, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model, Anal Chem 75 (23) (2003) 6415–6421.

[20] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, R. Aebersold, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, Nat Biotechnol 17 (10) (1999) 994–999.

[21] J. Li, H. Steen, S. P. Gygi, Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response, Mol Cell Proteomics 2 (11) (2003) 1198–1204.

[22] R. K. Carlin, D. J. Grab, R. S. Cohen, P. Siekevitz, Isolation and characterization of postsynaptic densities from various brain regions: enrichment of different types of postsynaptic densities, J Cell Biol 86 (3) (1980) 831–845.

[23] T. P. Gschwend, S. R. Krueger, S. V. Kozlov, D. P. Wolfer, P. Sonderegger, Neurotrypsin, a novel multidomain serine protease expressed in the nervous system, Mol Cell Neurosci 9 (3) (1997) 207–219.

[24] F. Molinari, M. Rio, V. Meskenaite, F. Encha-Razavi, J. Auge, D. Bacq, S. Briault, M. Vekemans, A. Munnich, T. Attie-Bitach, P. Sonderegger, L. Colleaux, Truncating neurotrypsin mutation in autosomal recessive nonsyndromic mental retardation, Science 298 (5599) (2002) 1779–1781.

[25] M. Mann, in: Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, GA (1995) 693.

[26] M. Wehofsky, R. Hoffmann, M. Hubert, B. Spengler, Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples, Eur J Mass Spectrom 7 (2001) 39–46.

[27] deltaMasses is available from http://www.detectorvision.com/deltaMasses.

[28] D. Arnott, M. A. Gawinowicz, R. A. Grant, T. A. Neubert, L. C. Packman, K. D. Speicher, K. Stone, C. W. Turck, ABRF-PRG03: phosphorylation site determination, J Biomol Tech 14 (3) (2003) 205–215.

[29] J. D. Gary, S. Clarke, RNA and protein interactions modulated by protein arginine methylation, Prog Nucleic Acid Res Mol Biol 61 (1998) 65–131.

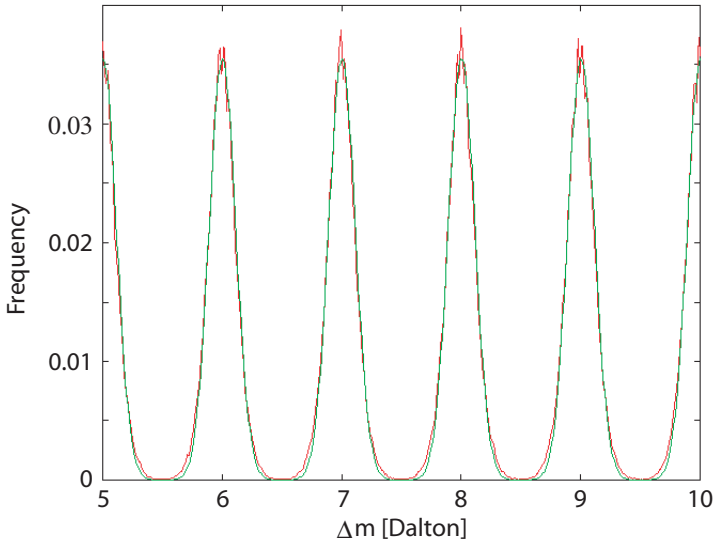

Fig. .1. Comparison of a simulated $MDH(\Delta m)$ (red curve) and the model $R(\Delta m)$ of Equation 1 (green curve) in mass range 5 to 10 Da. The MDH was obtained from 100 million randomly generated peptide pairs. The quantitative agreement is similar in the whole range between 0 to 100 Da.

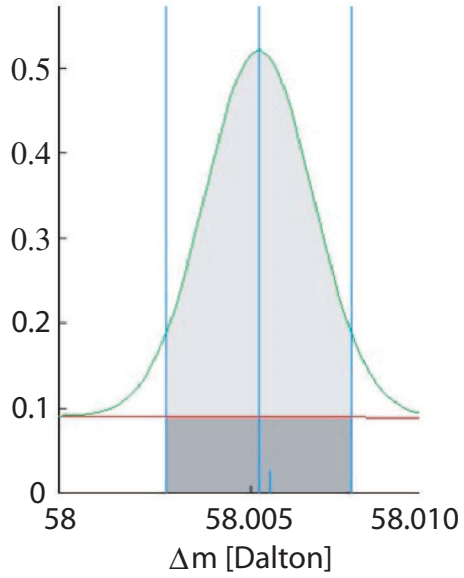

Fig. .2. Illustration of the method. Red curve: $R(\Delta m)$ having width σ_R. Green curve: the modification induced part having width σ_j. The red curve seems flat because $\sigma_R >> \sigma_j$ This peak close to $\delta_m = 58.005$ Da is induced by a modification. The area between the green and the red curve corresponds to the estimated number of true pairs. 95 percent of these true pairs are within $\pm 2\sigma_j$ distance from 58.005 Da. The $\pm 2\sigma_j$ boundaries are shown with blue lines.

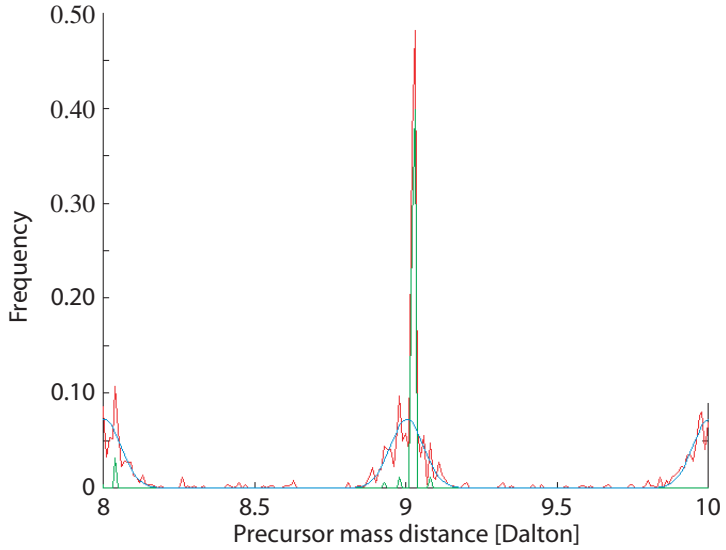

Fig. .3. Illustration of the MDF concept with data set 4, mass distances ranging from 8 to 10 Da. Red curve: the measured $MDH(\Delta m)$. Blue curve: $R(\Delta m)$ with $\sigma_R = 0.055$ Da. Green curve: signal used for deriving the Mass Distance Fingerprint. In this case, Peptoscope finds a modification at $\Delta m = 9.02967$ Da which is 0.00052 Da off from the theoretical value of the cleavable ICAT modification Δm of 9.03019 Da. (See Table .1 for the complete MDF results of data set 4.)

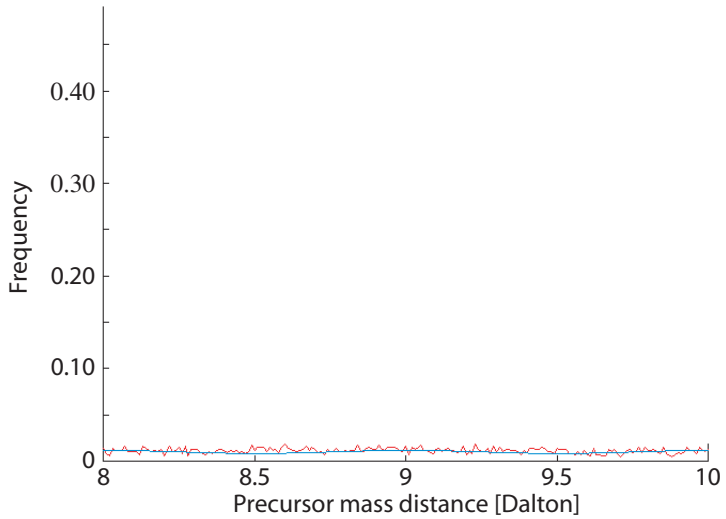

Fig. .4. Illustrating the importance of mass accuracy. Scales are equal to those of Figure .3. Red line: $MDH(\Delta m)$ obtained from an ICAT sample measured on an LCQ mass spectrometer [14] in the mass range from $\Delta m = 8$ to 10 Da. The plot shows that the LCQ precursor data does not yield any signal with Peptoscope due to low mass accuracy; the MDH fluctuates randomly around the expected random value of 0.01 (blue line). This is to be compared with the LTQ-FT results shown in Figure .3.

www.peptoscope.ms results for dataset4.mgf					
version	precursors	total pairs	range pairs	date	model σ
1.6	912	415416	52478	Mon Nov 27 19:01:20 CET 2006	0.055

mass distance fingerprint					chemical/ptm annotation	
mass [Da]	σ [10^{-4}]	intensity	number of "true" pairs	$\pm 2\sigma$ true positive [%]	ptm [unimod.org]	deviation [10^{-4}Da]
15.9947	14	1.1490	211.6	86	*Oxidation* *Deoxy* Ala↔Ser Phe↔Tyr	2
9.0297	11	2.2284	322.4	90	*Label:13C(9)*	5
17.0260	13	0.7370	126.0	82	*Gln→pyro-Glu* *Ammonia-loss*	5
14.0154	10	0.6644	87.4	80	*Methyl* Ala↔Gly Glu↔Asp Ile↔Val Thr↔Ser Val→Leu	2
6.9648	18	0.2660	63.0	71	?	
58.0052	12	0.4256	67.2	76	*Carboxymethyl* Asp↔Gly Glu↔Ala	3
25.0252	16	0.2949	62.1	68	?	
0.9834	12	0.3716	58.7	73	*Amidated* *Deamidated* Asp↔Asn Glu↔Gln	6
7.9957	4001	0.0010	52.6	0	?	
33.9610	6	0.5872	46.3	84	?	
30.0103	13	0.2970	50.8	68	*Pro→Pyrrolidinone* *Hydroxymethyl* Ser↔Gly Thr↔Ala	3
33.0217	11	0.3702	53.6	72	?	
28.9787	2335	0.0010	30.7	1	?	
1.0311	14	0.2088	38.5	64	*Lysaminoadipicsealde*	51
1.9698	30	0.0052	2.1	5	?	
18.0103	19	0.1504	37.6	53	*Dehydrated* *Glu→pyro-Glu*	3

Table .1

Peptoscope result for data set 4. In the *mass distance fingerprint* columns, the five MDF values (described in Section 3.1) are printed for each signal. Annotation of the MDF, together with mass shift from known chemical or biological modifications, is shown in the *chemical/ptm annotation* columns. As annotations the PSI-MS names are preferred and printed in italics whereas interim names are printed in normal font style. The optimal σ_R (Equation 1) for the analyzed data is given in the *model* σ field at top right.

www.peptoscope.ms results for dataset1.mgf					
version	precursors	total pairs	range pairs	date	model σ
1.6	4199	8813701	1791500	Mon Nov 27 20:43:01 CET 2006	0.105

mass distance fingerprint					chemical/ptm annotation	
mass [Da]	σ [10^{-4}]	intensity	number of "true" pairs	$\pm 2\sigma$ true positive [%]	ptm [unimod.org]	deviation [10^{-4}Da]
44.0259	15	2.8991	19528	93	*Ethanolyl*	3
88.0522	16	2.5187	18097	92	2*Ethanolyl	2
17.0260	15	2.1099	14212	92	*Gln→pyro-Glu Ammonia-loss*	
61.0523	15	1.7019	11464	91	++C2H7ON	5
71.0258	15	1.5243	10268	91	++71.02619	4
4.9554	9990	0.0014	6280	2	?	
26.9988	9098	0.0015	6128	2	?	
21.9815	14	0.8528	5361	88	*Cation:Na*	4
9.9731	7335	0.0015	4941	2	?	
83.0972	39	0.3566	6245	82	?	
34.0525	14	0.7470	4696	88	?	
48.9815	12	0.8472	4565	89	?	
22.0206	732	0.0040	1316	5	Asp↔His	114
39.0702	14	0.8325	5234	89	?	
93.0081	14	0.6592	4144	87	?	
98.0256	15	0.5389	3630	85	?	

Table .2
Peptoscope result for the MDF obtained for data set 1. The strongest signals are two PEG signals (called Ethanolyl in UniMod). The three modifications ++C2H7ON, ++71.02619, and 2*Ethanolyl are modifications that have been seen in our laboratory repeatedly, they are currently not listed in UniMod. The suggested C2H7ON has a mono-isotopic weight of 61.05276 Da.

www.peptoscope.ms results for dataset2.mgf						
version	precursors	total pairs	range pairs	date		model σ
1.6	819	334971	101945	Mon Nov 27 19:23:06 CET 2006		0.055
mass distance fingerprint					chemical/ptm annotation	
mass [Da]	σ [10^{-4}]	intensity	number of "true" pairs	$\pm 2\sigma$ true positive [%]	ptm [unimod.org]	deviation [10^{-4}Da]
15.9947	20	0.3750	191.7	73	*Oxidation* *Deoxy* Ala↔Ser Phe↔Tyr	4
1.0019	23	0.2216	130.2	62	*Dehydro*	59
27.9951	20	0.1836	93.8	59	*Formyl* *Pro→Pyrrolidone*	-2
43.9899	25	0.1303	83.2	55	*Carboxy* Ala↔Asp	-1
61.9560	757	0.0017	32.9	3	?	
60.9898	979	0.0011	27.5	1	?	
17.9535	178	0.0010	4.6	1	Ile↔Met Leu↔Met	29
5.9864	340	0.0040	34.7	3	?	
44.9928	774	0.0010	19.8	0	*Nitro*	-77
86.9955	11	0.1010	28.4	51	?	
17.9827	1605	0.0011	45.1	1	*Fluoro*	79
62.1187	550	0.0040	56.2	9	?	
44.0261	12	0.1447	44.4	52	*Ethanolyl*	1
49.9754	11	0.0571	16.1	40	?	
59.9860	451	0.0029	33.4	3	?	
45.9620	47	0.0127	15.3	15	?	

Table .3
Peptoscope result for the MDF obtained for data set 2.

www.peptoscope.ms results for dataset3.mgf					
version	precursors	total pairs	range pairs	date	model σ
1.6	16177	130839576	19093504	Mon Nov 27 11:31:03 CET 2006	0.105

mass distance fingerprint					chemical/ptm annotation	
mass [Da]	σ [10^{-4}]	intensity	number of "true" pairs	$\pm 2\sigma$ true positive [%]	ptm [unimod.org]	deviation [10^{-4}Da]
15.9945	24	0.1712	19665	70	*Oxidation* *Deoxy* Ala↔Ser Phe↔Tyr	4
14.0152	26	0.0726	9034	51	*Methyl* Ala↔Gly Glu↔Asp Ile↔Val Thr↔Ser Val→Leu	4
31.9899	23	0.0969	10667	59	*Dioxidation*	-1
1.9688	71	0.0137	4655	18	?	
58.0057	21	0.0580	5829	46	*Carboxymethyl* Asp↔Gly Glu↔Ala	-2
44.0263	21	0.0615	6181	47	*Ethanolyl*	-1
42.0107	23	0.0604	6649	47	*Acetyl*	-1
42.0131	1233	0.0011	6491	1	*Guanidinyl* *Amidino* *Arg→Orn*	87
3.9946	22	0.0546	5749	45	*Trp→Kynurenin* Pro↔Thr	3
30.0103	21	0.0621	6241	48	*Pro→Pyrrolidinone* *Hydroxymethyl* Ser↔Gly Thr↔Ala	3
26.0154	25	0.0467	5588	41	*Delta:H(2)C(2)* Pro↔Ala	2
46.0059	20	0.0517	4949	44	?	
17.0239	58	0.0283	7856	30	*Gln→pyro-Glu* *Ammonia-loss*	26
27.9911	67	0.0279	8947	30	*Formyl* *Pro→Pyrrolidone*	38
28.0315	24	0.0432	4962	39	*Dimethyl* *Delta:H(4)C(2)* *Ethyl* Ala↔Val	4
47.9778	101	0.0246	11891	29	*Trioxidation*	69

Table .4
Peptoscope result for the MDF obtained for data set 3. With more than 19 million mass distances within the MDH range of 0 to 100 Da, this is by far the largest data set presented in this study.

preliminary validation results for dataset1.mgf			
mass	MDF true pairs	MS/MS true pairs	ptm[unimod.org]
44.0259	19528	19089	*Ethanolyl*
88.0522	18096	14944	2*Ethanolyl
17.026	14212	2688	*Gln→pyro-Glu*
61.0523	11463	1330	++C2H7ON
71.0258	10267	291	++71.02619
4.9554	6280	-	Unknown
26.9988	6128	1428	Unknown
21.9815	5361	1517	*Cation:Na*
83.0972	6245	235	Unknown
39.0702	5233	725	Unknown
validation results for dataset2.mgf			
mass	MDF true pairs	MS/MS true pairs	ptm[unimod.org]
15.9945	191	308	*Oxidation*
1.00188	130	-	*Dehydro*
27.9951	93	121	*Formyl*
43.9899	83	63	*Carboxy*
validation results for dataset4.mgf			
mass	MDF true pairs	MS/MS true pairs	ptm[unimod.org]
15.9947	211	190	*Oxidation*
9.02967	322	241	*Label:13C(9)*
17.026	126	85	*Gln→pyro-Glu*
14.0154	87	73	*Methyl*
6.96475	63	25	Unknown
58.0052	67	18	*Carboxymethyl*
25.0252	62	25	Unknown
0.983416	58	-	*Deamidated*
7.9957	52	33	Unknown
30.0103	50	15	*Hydroxymethyl*

Table .5
Validation of MDF results with a method using MS/MS information. The detected number of pairs is in the same order of magnitude in most cases confirming the usefulness of the MDF. For compactness, the table shows only one UniMod abbreviation for each mass distance. The MS/MS algorithm cannot detect low mass shift modifications.