BMC Research Notes

BioMed Central

Short Report

Blind search for post-translational modifications and amino acid substitutions using peptide mass fingerprints from two proteases

Harald Barsnes*[1], Svein-Ole Mikalsen[2] and Ingvar Eidhammer[1]

Address: [1]Department of Informatics, University of Bergen, PB 7803, N-5020 Bergen, Norway and [2]Institute for Cancer Research, The Radium Hospital, The National Hospital Medical Center, Montebello, N-0310 Oslo, Norway

Email: Harald Barsnes* - harald.barsnes@ii.uib.no; Svein-Ole Mikalsen - svein-ole.mikalsen@rr-research.no; Ingvar Eidhammer - ingvar.eidhammer@ii.uib.no

* Corresponding author

Published: 19 December 2008

BMC Research Notes 2008, **1**:130 doi:10.1186/1756-0500-1-130

Received: 17 June 2008
Accepted: 19 December 2008

This article is available from: http://www.biomedcentral.com/1756-0500/1/130

Abstract

Background: Mass spectrometric analysis of peptides is an essential part of protein identification and characterization, the latter meaning the identification of modifications and amino acid substitutions. There are two main approaches for characterization: (i) using a predefined set of possible modifications and substitutions or (ii) performing a blind search. The first option is straightforward, but can not detect modifications or substitutions outside the predefined set. A blind search does not have this limitation, and therefore has the potential of detecting both known and unknown modifications and substitutions. Combining the peptide mass fingerprints from two proteases result in overlapping sequence coverage of the protein, thereby offering alternative views of the protein and a novel way of indicating post-translational modifications and amino acid substitutions.

Results: We have developed an algorithm and a software tool, MassShiftFinder, that performs a blind search using peptide mass fingerprints from two proteases with different cleavage specificities. The algorithm is based on equal mass shifts for overlapping peptides from the two proteases used, and can indicate both post-translational modifications and amino acid substitutions. In most cases it is possible to suggest a restricted area within the overlapping peptides where the mass shift can occur. The program is available at http://www.bioinfo.no/software/massShiftFinder.

Conclusion: Without any prior assumptions on their presence the described algorithm is able to indicate post-translational modifications or amino acid substitutions in MALDI-TOF experiments on identified proteins, and can thereby direct the involved peptides to subsequent TOF-TOF analysis. The algorithm is designed for detailed and low-throughput characterization of single proteins.

Background

The detection and verification of post-translational modifications in proteins and peptides by mass spectrometry (MS) is a common technique in protein characterization. The protein is proteolytically cleaved into peptides and analyzed by MS. MALDI-TOF instruments generate a list of mass-over-charge ratios (m/z values), referred to as a peptide mass fingerprint (PMF), which is compared to theoretical PMFs of known proteins. Modifications can be included in the theoretical PMFs. However, including too few can result in undetected modifications, while selecting too many can result in wrongly suggested modifications. One option is to perform the search in two iterations, where first a few expected modifications are considered. Thereafter, unmatched peptides are submitted to a modification search, e.g., in FindMod [1] or MassSorter [2]. FindMod only considers 22 common modifications. Here we present an alternative approach using blind search, where PMF data from two proteases on two aliquots of a sample are used to indicate modifications and amino acid substitutions. If the same mass shift relative to the unmodified theoretical values is observed for both proteases, and the peptides are overlapping, the mass shift can correspond to a modification or a substitution. MacCoss et al. [3] used a similar reasoning, but only to verify a limited set of predefined modifications in LC-MS/MS experiments. Unrestricted search for modifications using LC-MS/MS data has been developed more recently [4,5].

Figure 1 shows two overlapping peptides, p_1 and p_2, generated by different proteases. Let p_1 be the most N-terminal peptide, and p_2 the most C-terminal peptide. The overlapping peptides define three areas: the overlapping area, *Y*; the area of p_1 not overlapping with p_2, *X* (N-terminal area); and the area of p_2 not overlapping with p_1, *Z* (C-terminal area). Together, these will be referred to as the covered area. Note that p_1 and p_2 may have the same start or end residue, or that one peptide can completely cover the other. The main idea for our method is that a modification or an amino acid substitution occurring in area *Y* can be detected as an equal mass shift in p_1 and p_2. Equal mass shifts occurring in *X* and *Z*, but not in *Y*, can also be detected. This means that the non-overlapping areas *X* and *Z* both contain the same modified amino acid, or different amino acids carrying an identical modification, e.g., phosphorylation on S and T.

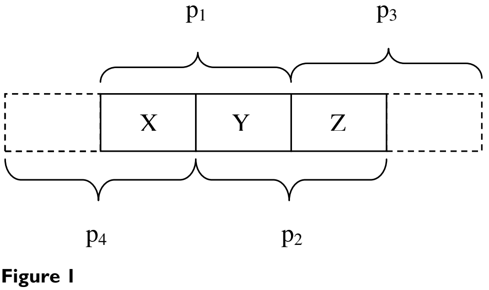


Figure 1
Overlapping peptides. The peptides p_1 and p_2 define different regions (*X*, *Y*, *Z*) of the covered area as explained in the text. The figure also indicates that if adjacent peptides p_3 and p_4 are found, they can be combined with p_2 and p_1, respectively, to strengthen the probability for found mass shifts in X or Z being real mass shifts.

Let (i) p_1 and p_2 be two overlapping theoretical peptides; (ii) t_1 be the theoretical mass of p_1, and t_2 be the theoretical mass of p_2; and (iii) e_1 be an experimental mass using protease A, and e_2 be an experimental mass using protease B. Suppose the following equation is observed: $e_1 - t_1 = e_2 - t_2 = \Delta m$. Δm is then either a real mass shift or an artifact.

Real mass shifts: e_1 corresponds to p_1, and e_2 corresponds to p_2. The mass shift can then (i) occur solely in *Y*; (ii) occur in both *X* and *Z*; or (iii) the mass shifts occur as a combination of the two former cases. In the first case, the mass shift can correspond to one or more modifications/substitutions, while in the other cases, two or more modifications are needed.

Artifact: at least one of the masses e_1 or e_2 does not correspond to p_1 or p_2 respectively. Artifacts are covered in more detail in the Discussion.

Findings

Algorithm

The following algorithm detects equal mass shifts in overlapping peptides:

1. Let E_1 be the peptide mass list from an experiment using protease A, and E_2 be the peptide mass list from an experiments using protease B.

2. Let T_1 be the list of theoretical peptide masses resulting from an in silico digestion using protease A, and T_2 be the list of theoretical peptide masses resulting from an in silico digestion using protease B.

3. Remove from E_1 all peaks corresponding to unmodified peptides in T_1 and all peaks corresponding to autolytic peaks from protease A.

4. Repeat Step 3 with mass lists E_2 and T_2 from protease B.

5. Compare each mass $e_i \in E_1$ to each mass $t_j \in T_1$, and each mass $e_k \in E_2$ to each mass $t_m \in T_2$. Store the mass shifts $(e_i - t_j)$ for all *i* and *j* and the mass shifts $(e_k - t_m)$ for all *k* and *m*, in two lists M_1 and M_2, which now contain all possible mass shifts between corresponding experimental and theoretical data.

6. Let p_j and p_m be the theoretical peptides corresponding to t_j and t_m respectively. Compare M_1 and M_2 and find all pairs such that:

a. $|(e_i - t_j) - (e_k - t_m)| = \omega$ and (ω is the mass shift accuracy)

b. $|(e_i - t_j)| > \varepsilon$ and $|(e_k - t_m)| > \varepsilon$ and (ε is the mass shift threshold)

c. p_j and p_m overlap

The output is a list of overlapping peptides from E_1 and E_2 with equal mass shifts. The reason for the mass shifts, i.e., modification(s) or substitution(s), has to be positioned in the covered area, and have a mass equal to the detected mass shift. The list should be cross-checked against a database of known modifications and substitutions (e.g., UniMod [6,7]), and/or the included peptides can be tested in additional experiments, i.e., by MALDI-TOF-TOF, verifying or rejecting the proposed modification or substitution.

Implementation

The described algorithm is implemented in Java [8] and available as a software tool, MassShiftFinder, at http://www.bioinfo.no/software/massShiftFinder.

The main input to MassShiftFinder is the protein sequence and the experimental masses from two PMF experiments on the same protein using different proteases. Before running the algorithm it is recommended to remove all identified peptides from the PMFs, e.g., by using MassSorter [2]. Unmodified peptides, autolytic protease peaks and known noise/contaminating peaks (e.g., keratin) can be filtered within adjustable accuracy limits in the program. Using filters limits the number of unnecessary mass shift comparisons (see additional file 1, Fig. 1 (TheoreticalExamples.pdf)).

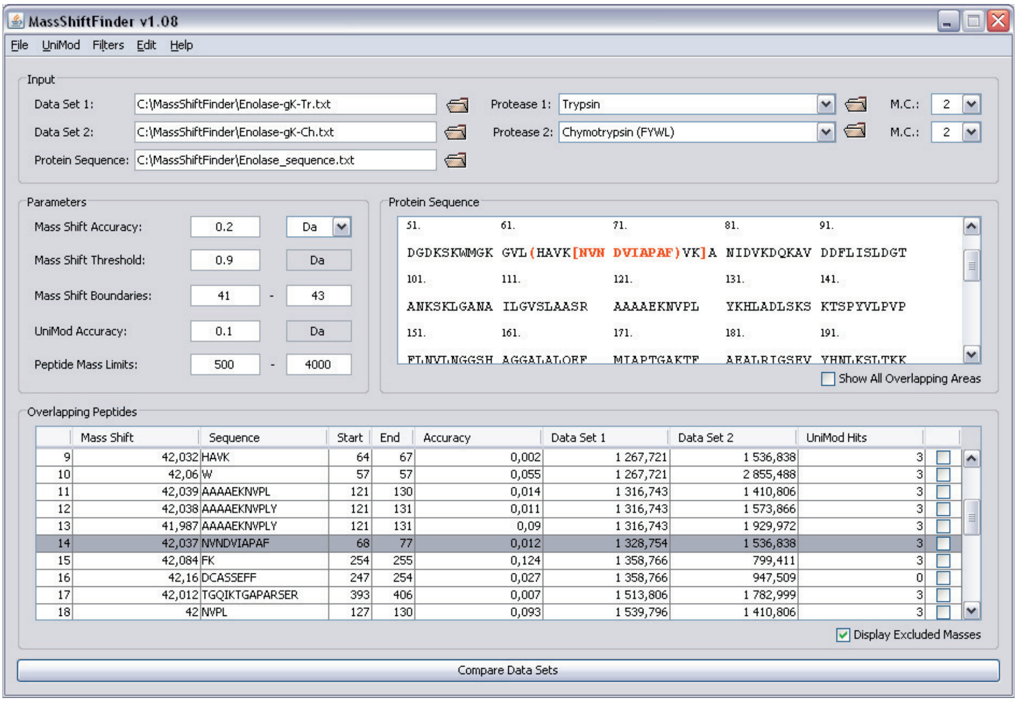
In order to reduce search space and increase the possibility of detecting real mass shifts, the following parameters should be set to reasonable values. (i) Mass Shift Threshold, where mass shifts below this threshold are excluded to avoid spurious comparisons among very small mass shifts. We would in general recommend setting this value to 0.9 to achieve the inclusion of deamidations. (ii) Mass Shift Boundaries, determine the search limits for a mass shift being a modification or substitution. It can be set to a more limited mass range, e.g., 79–81 Da to search for phosphorylations. (iii) Mass Shift Accuracy, where equal mass shifts are recognized when the difference between two mass shifts are within this accuracy (in Da or ppm). We would in general recommend setting this parameter at 0.2 Da when 25 ppm accuracy limit is used for the experimental peptides, and to decrease it if the instrument is more exact. Note that this parameter refers to inaccuracy of the potential modification as calculated from the comparison of experimental data and the theoretical peptide sequence.

An example of output is shown in Figure 2. By selecting a row, the overlapping peptides are indicated in the protein sequence. The detected mass shifts are searched against a local version of the UniMod database. To reduce the amount of incorrect UniMod explanations, this search can be restricted by choosing the allowed modification types, e.g., amino acid substitutions, post-translational modifications, etc. Up to two modifications per peptide are supported. Note that changing the settings for the UniMod search only affects the number of suggested explanations for each mass shift, not the number of mass shifts. Unexplained mass shifts may correspond to unknown modifications or more than two modifications per peptide. An example showing detection of modifications in an artificial dataset is found in additional file 1 (TheoreticalExamples.pdf).

Experimental Example

We compared connexin43 (Cx43) [9] from three species. The experimental peak lists of Cx43 from Syrian hamster, Chinese hamster and rat were collected in MassSorter [2] using the Syrian hamster sequence as basis of comparison [10]. After removing autolytic protease peaks, peaks from the contaminating antibody and peaks in common with Syrian hamster, the remaining peaks were inserted into MassShiftFinder using the following parameters: Filter Accuracy and Unmodified Peptide Accuracy, 50 ppm (found under Edit/Preferences); Mass Shift Accuracy, 0.2 Da; Mass Shift Threshold, 0.9 Da; Mass Shift Boundaries, -200 to 200 Da; UniMod Accuracy, 0.1 Da; Missed Cleavages, 1; and including only amino acid substitutions in the search.

For Chinese hamster, MassShiftFinder pointed out a potential substitution within the area 347-IAAGHELQPL-356 with a mass shift of 17.96 Da. This would correspond to a substitution from I or L to M. The rat data also indicated a potential substitution in the same sequence with a mass shift of -14.02 Da. This could correspond to a substitution from A to G, E to D, or I or L to V. The Chinese hamster and rat peptides with m/z 1748.91 and m/z 1716.84 (corresponding to mass shifts of 17.95 Da and -14.02 Da relative to the Syrian hamster peptide with m/z 1730.96) were targeted for TOF-TOF analysis (Fig. 3). The only possible substitution in Chinese hamster that is consistent with all data is a change in position 347 from I (Syrian hamster) to M (Chinese hamster). For rat, both I347 to V and A348 to G are consistent with these data. The former is the correct alternative. This example shows that our approach can be used to narrow the range of possibilities when detecting amino acid substitutions. For more examples and details, see additional file 2 (ExperimentalExamples.pdf).



Data Set 1: C:\MassShiftFinder\Enolase-gK-Tr.txt

Data Set 2: C:\MassShiftFinder\Enolase-gK-Ch.txt

Protein Sequence: C:\MassShiftFinder\Enolase_sequence.txt

Protease 1: Trypsin M.C.: 2

Protease 2: Chymotrypsin (FYWL) M.C.: 2

Mass Shift Accuracy: 0.2 Da

Mass Shift Threshold: 0.9 Da

Mass Shift Boundaries: 41 - 43

UniMod Accuracy: 0.1 Da

Peptide Mass Limits: 500 - 4000

```
51.        61.          71.          81.        91.
DGDKSKWMGK GVL(HAVK[NVN DVIAPAF)VK]A NIDVKDQKAV DDFLISLDGT
101.       111.         121.         131.       141.
ANKSKLGANA ILGVSLAASR   AAAAEKNVPL   YKHLADLSKS KTSPYVLPVP
151.       161.         171.         181.       191.
FLNVLNGGSH AGGALALQEF   MIAPTGAKTF   AEALRIGSEV YHNLKSLTKK
```

□ Show All Overlapping Areas

	Mass Shift	Sequence	Start	End	Accuracy	Data Set 1	Data Set 2	UniMod Hits	
9	42,032	HAVK	64	67	0,002	1 267,721	1 536,838	3	□
10	42,06	W	57	57	0,055	1 267,721	2 855,488	3	□
11	42,039	AAAAEKNVPL	121	130	0,014	1 316,743	1 410,806	3	□
12	42,038	AAAAEKNVPLY	121	131	0,011	1 316,743	1 573,866	3	□
13	41,987	AAAAEKNVPLY	121	131	0,09	1 316,743	1 929,972	3	□
14	42,037	NVNDVIAPAF	68	77	0,012	1 328,754	1 536,838	3	□
15	42,084	FK	254	255	0,124	1 358,766	799,411	3	□
16	42,16	DCASSEFF	247	254	0,027	1 358,766	947,509	0	□
17	42,012	TGQIKTGAPARSER	393	406	0,007	1 513,806	1 782,999	3	□
18	42	NVPL	127	130	0,093	1 539,796	1 410,806	3	□

☑ Display Excluded Masses

Figure 2
Screenshot of the MassShiftFinder main window. The data are taken from an experiment on guanidinated enolase (see additional file 2: ExperimentalExamples.pdf). The mass shift boundaries are set to restrict the detected mass shift mainly to guanidinations (mass shift of 42 Da). The highlighted row (14) indicates an "X-Z" mass shift (see Figure 1). Rows 11–13 show a mass shift occurring in the Y-area, with one tryptic peptide (1316.7) paring up with three chymotryptic peptides. The right-most column (UniMod hits) for most of the rows indicates that there are three modifications (acetylation, tri-methylation and guanidination) that are consistent with a mass shift of 42 for the indicated peptides. Row 16 has 0 UniMod hits because the calculated mass shift is more than 0.1 Da from the three mentioned modifications; furthermore, this mass shift cannot be due to guanidinated K as only one of the peptides contains a K.

Discussion

The algorithm depends on good experimental sequence coverage and overlapping peptides. Sequence coverage mainly depends on the amino acid sequence, the sample amount, the protease used, and purity. An analysis of human proteins in SwissProt suggests that approximately 70–90% of the proteins have a theoretical coverage between 50 and 100%, regardless of whether trypsin, chymotrypsin or gluC was used (see additional file 3: SupplementaryMaterials.pdf). The experimental sequence coverage is usually lower than the theoretical upper limit, but a considerable degree of experimental overlap would generally be expected.

A detected mass shift (Δm) can either be real, i.e., resulting from a modification/substitution, or an artifact. Although unknown modifications still can be found [4], it is more likely that a mass shift is due to a known modification. Following the parsimony principle, it seems reasonable to first assume that a mass shift is caused by a single known modification. Accepted modified peptides can then be removed before subsequent searches are performed with less restricted parameters, e.g., allowing two modifications per peptide.

The tendency for artifacts is augmented by the clustering of peptide masses [11-15] and the fact that most modification masses also are close to integers. In the mass range

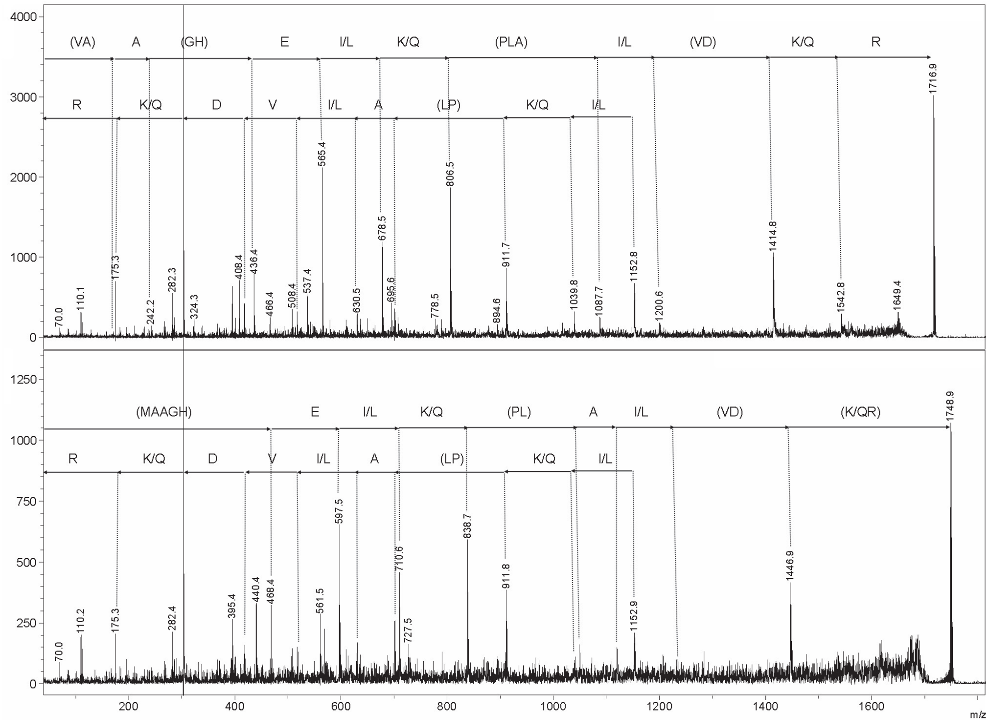


Figure 3

TOF-TOF data of the Cx43 peaks at m/z 1716.84 and m/z 1748.91. The peptides 1716.84 (upper) and 1748.91 (lower) are from rat and Chinese hamster Cx43, respectively. Note that the y_2-ion at m/z 303.2 had higher intensity than all other ions. In both panels, the upper sequence is read from the b-ions, and the lower sequence is read from the y-ions. Further note that also the a_5- to a_8-ions can be distinguished in the upper panel, and the a_5- and a_6-ions in the lower panel. See text for more explanation.

from 1 to 100 Da, approximately 75% of all integers have one or several modifications/substitutions with a mass close to it [6,7]. This means that at any random, but near-integer, distance from the true peptide m/z value, there is a considerable chance that one or several modifications will fit to this integer value. Furthermore, any positive near-integer value between 2 and 100 can be achieved by a combination of two modifications in the peptide. Thus, as the number of non-identified peptides increases, the likelihood of finding artifacts also increases.

In characterization high sequence coverage is desired, and one might therefore use as many peaks as possible, including low intensity peaks that would not have been used for identification purposes. Such peaks are more influenced by random noise, and are in general expected to have lower accuracy than high intensity peaks. Proteolytic cleavage specificity and efficiency are also not perfect. Thus, several factors will contribute to artifacts. A main strategy is to remove all peptides that can be identified with reasonable confidence before the initial mass shift comparison is performed. It is also recommended to search for peptides with unexpected cleavages or many missed cleavages by using MassSorter [2], FindPept [16] or similar tools, especially if an "unreliable" protease (like chymotrypsin) has been used.

Our primary objective with the algorithm is to promote the detailed low-throughput characterization of single proteins by indicating peptides that may contain modifications or substitutions. This can help in selecting peaks to target in fragmentation experiments. Furthermore, it is well known that a number of peptides are difficult to fragment in (LC-)MS/MS experiments. If an accurate instrument is used (e.g., Orbitrap or Q-TOF), it would be possible to extract suggestions for modifications from the survey scans, which could be the basis of alternative experiments (the use of other proteases, introduced chemical modifications, site-directed mutations in recombinant proteins, etc.).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HB did the programming, in silico analyses, contributed ideas and was the main author of the manuscript. SOM made the initial description of the method, performed the mass spectrometry experiments, and participated in writing the manuscript. IE supervised the programming work, contributed ideas, and participated in writing the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Theoretical Examples*. A PDF file containing examples showing detection of modifications in an artificial dataset.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-1-130-S1.pdf]

Additional file 2

Experimental Examples*. A PDF file containing details on MS experiments performed to show the proposed usage of the algorithm and software tool.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-1-130-S2.pdf]

Additional file 3

Supplementary Material*. A PDF file containing additional information regarding a theoretical analysis of the degree of coverage and overlap between the proteases trypsin, chymotrypsin and gluC using 19,852 human proteins.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-1-130-S3.pdf]

Acknowledgements

Dr. Burkhard Fleckenstein and Marit Jørgensen (Institute of Immunology, The National Hospital, Oslo) are thanked for making the MALDI-TOF-TOF available to us and for their assistance in use of the instrument.

References

1. **FindMod** [http://ca.expasy.org/tools/findmod/]
2. Barsnes H, Mikalsen SO, Eidhammer I: **MassSorter: a tool for administrating and analyzing data from mass spectrometry experiments on proteins with known amino acid sequences.** *BMC bioinformatics* 2006, **7**:42.
3. MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, *et al.*: **Shotgun identification of protein modifications from protein complexes and lens tissue.** *Proc Natl Acad Sci USA* 2002, **99**:7900-7905.
4. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA: **Identification of post-translational modifications by blind search of mass spectra.** *Nature Biotechnology* 2005, **23**:1562-1567.
5. Tanner S, Pevzner PA, Bafna V: **Unrestrictive identification of post-translational modifications through peptide mass spectrometry.** *Nat Protoc* 2006, **1**:67-72.
6. **UniMod** [http://www.unimod.org/]
7. Creasy DM, Cottrell JS: **UniMod: Protein modifications for mass spectrometry.** *Proteomics* 2004, **4**:1534-1536.
8. **Java** [http://www.java.com]
9. Cruciani V, Mikalsen SO: **Evolutionary selection pressure and family relationships among connexin genes.** *Biol Chem* 2007, **388**:253-264.
10. Cruciani V, Heintz KM, Husøy T, Hovig E, Warren DJ, Mikalsen SO: **The detection of hamster connexins: a comparison of expression profiles with wild-type mouse and the cancer-prone Min mouse.** *Cell Commun Adhes* 2004, **11**:155-171.
11. Gay S, Binz PA, Hochstrasser DF, Appel RD: **Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra.** *Proteomics* 2002, **2**:1374-1391.
12. Wolski WE, Farrow M, Emde AK, Lehrach H, Lalowski M, Reinert K: **Analytical model of peptide mass cluster centres with applications.** *Proteome Sci* 2006, **4**:18.
13. Mann M: **Useful tables of possible and probable peptide masses.** *43rd ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, GA* .
14. Wool A, Smilansky Z: **Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting.** *Proteomics* 2002, **2**:1365-1373.
15. Barsnes H, Eidhammer I, Cruciani V, Mikalsen SO: **Protease-dependent fractional mass and peptide properties.** *Eur J Mass Spectrom (Chichester, Eng)* 2008, **14(5)**:311-317.
16. **FindPept** [http://au.expasy.org/tools/findpept.html]

Theoretical Examples

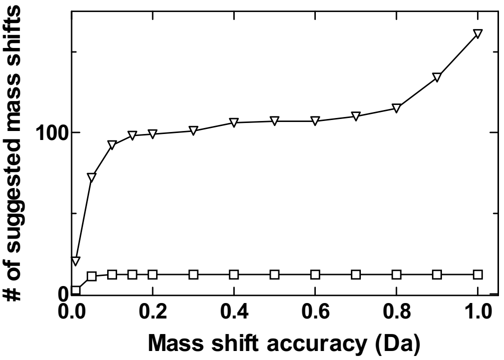
The tumor suppressor protein p53 (P04637) was used as a theoretical exercise for MassShiftFinder. Random inaccuracies (within ±25 ppm) were introduced in in-silico generated tryptic and chymotryptic peptides of p53. Some modifications and/or substitutions were then introduced in the peptides. The datasets were generated in Bergen and then sent to Oslo. The only information given was the identity of the protein and the protease used for each of the datasets. Information about the introduced modifications was not provided. Thus, the information available was rather similar to an experimental situation.

The data were first analyzed in MassSorter, and peptides fitting with theoretically unmodified peptides, peptides containing oxidized methionine and N-terminal pyroglutamic acid generated from glutamine were removed. In each of the test situations, five to seven peptides from each of the "digests" showed no obvious matches, and these peptides were transferred to MassShiftFinder. This dataset is referred to as filtered data.

We will here give some examples from one of the tests. Nine different combinations of peptides indicated a mass shift of -58.0 Da centered around the region 271-EVR-273. The mass shift immediately suggested a substitution, of which E271 to A would easily explain these data. However, also the double substitution E to G together with V to I/L would give the same mass shift. The former situation seemed the biologically most plausible, and was correct.

Another mass shift of 239.9 Da was detected for three pairs of overlapping peptides in the region 94-SSSVPSQK-101. This mass shift did not correspond to any single modification available in UniMod, and also not the sum of two modifications. Thus, this corresponded either to an unknown modification or more than two modifications. The region contained four S residues, and the mass shift was very close to three phosphorylations. Experimentally, this possibility could easily be tested by looking for unfocused post-source decay fragments in MALDI-TOF or neutral loss by MS/MS, or treating the sample with alkaline phosphatase.

The described example was also used to generate graphs describing how the number of suggested mass shifts changed as a function of the Mass Shift Accuracy, which was set at 0.2 Da during these exercises (TheoreticalExamples Fig. 1). The number of suggested mass shifts increased rapidly up to 0.1 (filtered data) or 0.2 Da (unfiltered data, where only m/z values corresponding to unmodified peptides were removed), and thereafter it plateaued until approx. 0.9 Da. Other theoretical and experimental examples followed similarly shaped curves. TheoreticalExamples Fig. 1 also indicates the advantage of removing peptides that can be identified with reasonable confidence before the analysis in MassShiftFinder.



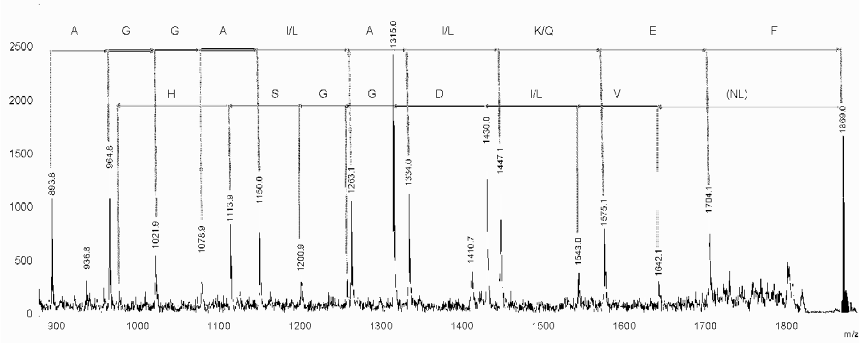
TheoreticalExamples Figure 1: Number of suggested hits by MassShiftFinder using artificially created digests and modifications in human p53. The generation of the datasets is described in the text above, and they were analyzed using the following settings: Peptide Accuracy, 25 ppm; Missed Cleavages, trypsin 1, chymotrypsin 2; Mass Shift Accuracy, variable (see figure); Mass Shift Threshold, 0.9 Da; Mass Shift Boundaries, -200 to 400, UniMod Accuracy, 0.1 Da, Peptide Mass Limits, 500 to 4000. Squares: peptides corresponding to unmodified p53 peptides, and peptides with oxidized methionines or with N-terminal pyroglutamic acid were removed before the analysis (filtered data). Triangles: only unmodified p53 peptides were removed before analysis (unfiltered data).

Experimental Examples

Modifications in enolase: Tryptic and chymotryptic peptides from unmodified enolase were analyzed by MassShiftFinder. The same parameters were used as in the main text, except that mass shift boundaries were set to -20 to 200, missed cleavages were increased to 2 and post-translational and artefactual modifications (but no substitutions) were included in the search. For unmodified enolase, the areas 33-SIVPSGASTGVHEALEMR-50 and 369-GVMSHR-375 were both indicated to contain a modification with a mass shift of 16.0 Da. This would correspond to an oxidized M in both peptide sequences. A sodium adduct (mass shift of 22 Da relative to MH^+) was indicated for the enolase C-terminal peptides (the tryptic peptide 416-437 and chymotryptic peptide 421-437). A pair of m/z 2968.5 (tryptic) and 1745.9 (chymotryptic) was suggested as (K)289-RYPIVSIEDPFAEDDWEAWSHFK-311 and (W)308-SHFKTAGIQIVADDL-323 with a mass shift of -16 Da, corresponding to a β-elimination followed by reduction. This seemed rather unlikely, and the chymotryptic peptide, which had the more intense peak, was targeted for fragmentation. The fragment spectrum was consistent with the sequence 412-QLLRIEEELGDNAVF-426, and was thus cleaved C-terminal to N411. This serves as an example of an erroneous suggestion caused by an unexpected chymotryptic cleavage. (The tryptic digest of (unmodified) enolase had a coverage of 60%, the chymotryptic digest 62%.)

Modifications in ovalbumin: Again, several suggestions for oxidized methionines were obtained for positions 196, 210, 211, 222, 227 and/or 239. Another pair of peptides suggested potassium adducts occurring in the tryptic peptide 187-AFKDEDTQAMPFR-199 (m/z 1593.7) and the chymotryptic peptide 199-RVTEQESKPVQMMYQIGLF-217 (m/z 2322.2). Additionally suggested mass shifts will not be discussed as fragmentation was not productive, or the peaks were of very low intensity and fragmentation was not attempted. (The tryptic digest of ovalbumin had a coverage of 52%, the chymotryptic digest 48%.)

Guanidination of enolase: To test how MassShiftFinder would handle a massive amount of modifications, enolase was guanidinated. Enolase contains in total 14 arginines and 37 lysines. The majority of tryptic peptides (and numerous chymotryptic peptides) should therefore be modified by guanidination. Chemical modifications were included in the settings, but otherwise the same parameters as previously described were used. MassShiftFinder indicated a total of 183 tryptic-chymotryptic pairs that would give equal mass shifts when mapped to the enolase sequence. Among these, 48 were consistent with one or two guanidinations, i.e., mass shifts of 42 or 84 Da. Twenty-five of the pairs had K in the *Y* area, while 17 had K in both the *X* and *Z* (but not *Y*) area, and 6 had K in *X*, *Y* and *Z*. In several cases, more than one hit pointed to the same position(s). The pairs indicative of guanidination were removed. Further analysis of the remaining peaks using MassSorter [1] indicated that some of the remaining peptides were due to non-overlapping guanidinated peptides. These peptides were removed manually and the analysis was repeated, now giving a total of 32 hits, with 10 of the hits corresponding to one UniMod [2, 3] registered modification per peptide. Two of the UniMod hits could correspond to oxidations (mass shifts of 16 or 32 Da) of W57, M58, W273 and M287. Another hit with a mass shift of 100.1 Da was consistent with two guanidinations of Ks and one oxidation of M in each of the peptides 32-RSIVPSGASTGVHEALEMRDGDKSKW-57 (chymotryptic) and 57-WMGKGVLHAVK-67 (tryptic). Several of the peptides with no UniMod hits were subjected to TOF-TOF analysis, but in general these peaks were of low intensity. Only one of the fragmentations was productive, a chymotryptic peak of m/z 1868.9 that paired with several tryptic peptides, giving mass shifts ranging from -19 to 175 Da, and pointing to several areas in enolase. The spectrum was resolved and corresponded to (F)152-LNVLN*GGSHAGGALALQEF-170, where N* (N156) had been deamidated, resulting in D, see ExperimentalExamples Figure 1. The non-modified peptide (at m/z 1867.9) was found in the enolase samples that had not been guanidinated. The NG combination is known to be particularly sensitive to deamidation [4, 5].



ExperimentalExamples Figure 1: Deamidation of Asn156 to Asp in guanidinated enolase. Partial fragmentation spectrum of the chymotryptic peptide, 152-LNVLNGGSHAGGALALQEF-170, from guanidinated enolase. The upper sequence is read from b-ions, and the lower from y-ions. The y-ions show the deamidation of N156 to D, changing the m/z value of the peptide from 1867.9 to 1868.9.

Methods

Materials: Chicken ovalbumin, enolase from Saccharomyces cerevisiae, anti-Cx43 antibody (C6219) and bovine chymotrypsin were bought from Sigma. Porcine trypsin was from Promega. Zip-Tips were obtained from Millipore.

Immunoprecipitation of connexin43: Connexin43 (Cx43) was immunoprecipitated from primary Syrian hamster embryo cells, Chinese hamster V79 cells and rat NRK52E cells by a polyclonal anti-Cx43 antibody. The separation of the immunoprecipitated proteins on gels, silver staining, band excision, destaining and dehydration with acetonitrile were performed essentially according to Gharahdaghi et al. [6]. The gel pieces were treated with trypsin or chymotrypsin overnight at 37 °C. The peptides were extracted by acetonitrile [6], dried down, and further desalted and purified by μC18-ZipTips. The samples were analyzed in a Bruker Ultraflex MALDI-TOF-TOF.

Preparation of enolase and ovalbumin samples: Enolase and ovalbumin were run on gels and the gel bands excised and further treated as described above, and digested with either trypsin or chymotrypsin. Enolase gel bands were also modified by guanidination before proteolytic treatment as described [7].

References

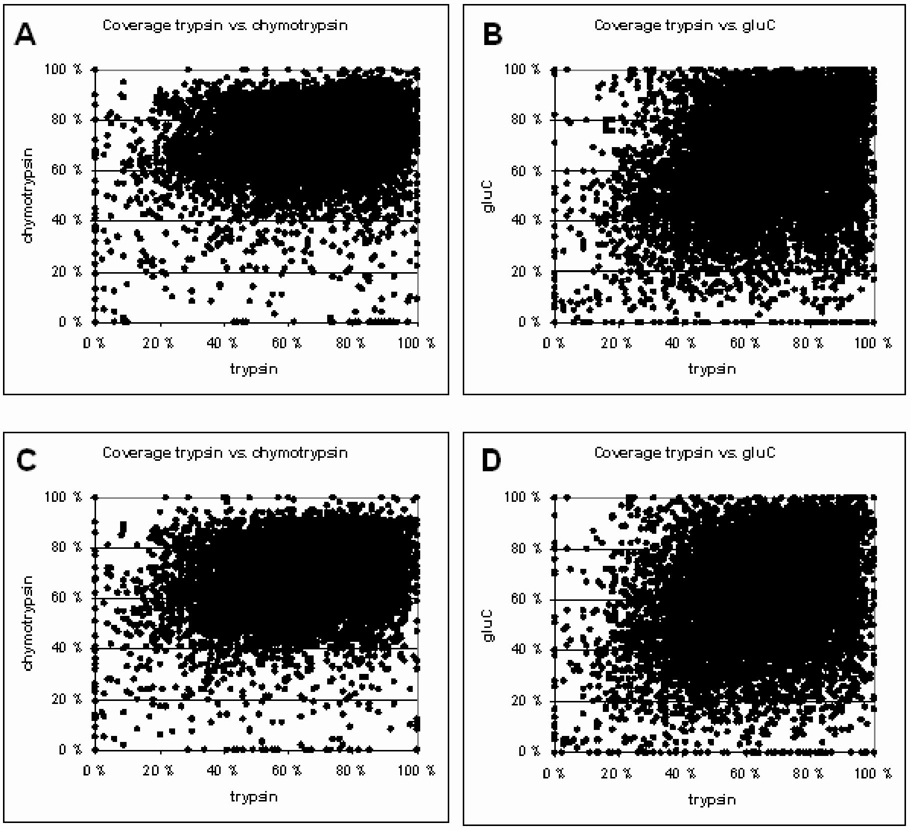
1. Barsnes H, Mikalsen SO, Eidhammer I: MassSorter: a tool for administrating and analyzing data from mass spectrometry experiments on proteins with known amino acid sequences. *BMC bioinformatics* 2006, 7:42.
2. UniMod [http://www.unimod.org/]
3. Creasy DM, Cottrell JS: UniMod: Protein modifications for mass spectrometry. *Proteomics* 2004, 4:1534 - 1536.
4. Krokhin OV, Antonovici M, Ens W, Wilkins JA, Standing KG: Deamidation of -Asn-Gly- sequences during sample preparation for proteomics: Consequences for MALDI and HPLC-MALDI analysis. *Anal Chem* 2006, 78:6645-6650.
5. Robinson AB, Rudd CJ: Deamidation of glutaminyl and asparaginyl residues in peptides and proteins. *Curr Top Cell Regul* 1974, 8:247-295.
6. Gharahdaghi F, Weinberg CR, Meager DA, Imai BS, Mische SM: Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: A method for the removal of silver ions to enhance sensitivity. *Electrophoresis* 1999, 20:601-605.
7. Lindstad RI, Sylte I, Mikalsen SO, Seglen PO, Berg E, Winberg JO: Pancreatic trypsin activates human promatrix metalloproteinase-2. *J Mol Biol* 2005, 350:682-698.

In Silico Analysis Of Overlapping Peptides

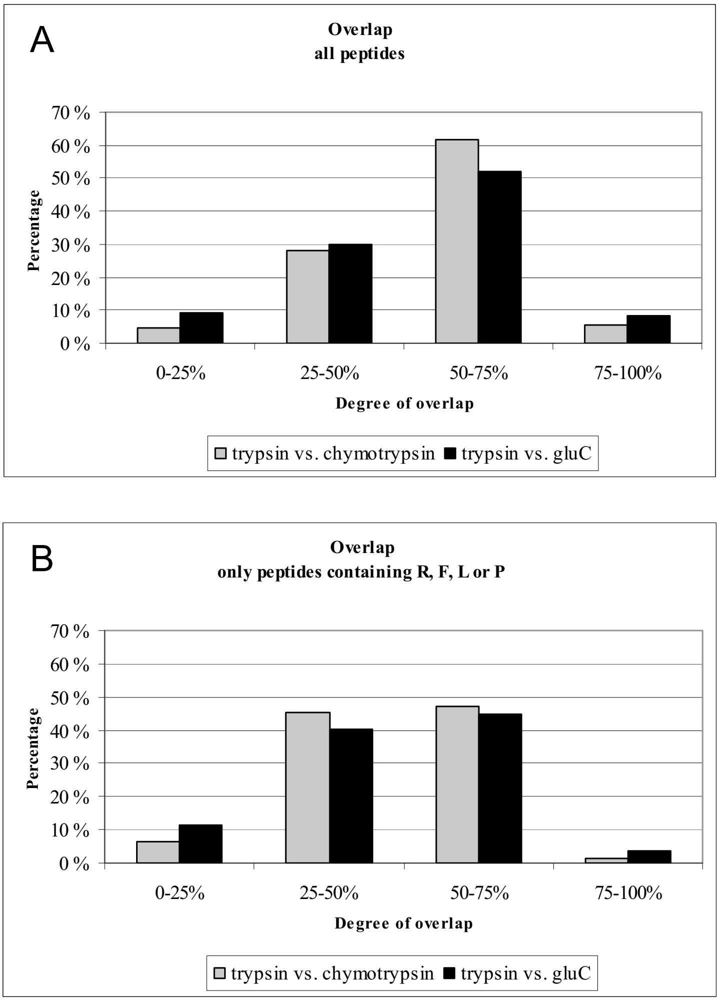
The described algorithm depends on the comparison of the digests from two proteases, and that the resulting peptides overlap. In turn, the degree of overlap depends mainly on the sequence coverage achieved. By applying the cleavage rules of a protease, along with peptide mass boundaries for mass spectrometry data and a maximum number of allowed missed cleavages, a theoretical upper boundary for sequence coverage can be determined for all protein-protease pairs. We therefore performed an analysis of all human proteins in the Swiss-Prot [1] database (Release 54, 19852 proteins) to investigate how different proteases theoretically affected the sequence coverage. All SwissProt proteins were in silico digested by trypsin (cleaves after R and K, unless followed by P), chymotrypsin (cleaves after F, Y, W and L, unless followed by P) and gluC (cleaves after E and D, unless followed by P). A maximum of one missed cleavage was allowed (two missed cleavages were also tested, but had little impact on the results). To emulate mass spectrometry conditions only peptides between 500 and 3500 Da were used.

The ionization and intensity of peptide peaks detected in MALDI instruments partly depend on the presence of certain amino acids, especially R [2], but also F, L and P [3]. Only peptides containing at least one of these amino acids were included in the analysis. By this restriction, 8.3% (trypsin) to 9.3% (gluC) of the peptides were removed, which on the average corresponded to 3 to 4 peptides per protein, see Supplementary Table 1. The theoretical sequence coverages for the single proteases are shown in Supplementary Table 2, and pairwise comparisons of the sequence coverages are plotted in Supplementary Figure 1. A further comparison of the theoretical sequence coverages for the three proteases, showed that chymotrypsin had higher (theoretical) coverage than trypsin in 49.7% of the proteins, and gluC in 44.6% of the cases, see Supplementary Table 3.

The same datasets were used to analyze how much the coverage for different proteases overlaps, see Supplementary Figure 2. For both protease pairs, trypsin vs. chymotrypsin and trypsin vs. gluC, around 50% or more of the proteins had a theoretical overlap higher than 50%.



Supplementary Figure 1: Comparison of coverage degrees for 19,852 human proteins (Swiss-Prot November 22nd 2007) theoretically digested by trypsin, chymotrypsin and gluC. For **A** and **B** all peptides are used, while in **C** and **D** only peptides containing at least one R, F, L or P are included. Lower mass limit: 500, upper mass limit: 3500, maximum missed cleavages: 1. Chymotrypsin was used with the specificity FYWL.



Supplementary Figure 2: An overview of the degree of overlap for 19,852 human proteins theoretically digested by trypsin, chymotrypsin and gluC. The degree of overlap is calculated as the percentage of the total sequence covered by both of the proteases. It is divided into four groups, and the number of proteins in each group is counted. In **A** all peptides are used, while in **B** only peptides containing at least one R, F, L, or P are included. Lower mass limit, 500; upper mass limit, 3500; maximum missed cleavages, 1; chymotrypsin cleaves after F, Y, W and L.

	Trypsin	**Trypsin**	**Chymo-trypsin FYWL**	**Chymo-trypsin FYWL**	**Chymo-trypsin FYW**	**Chymo-trypsin FYW**	**GluC**	**GluC**
Coverage degree	**All**	**RFLP**	**All**	**RFLP**	**All**	**RFLP**	**All**	**RFLP**
0-25%	1.6 %	1.8 %	0.9 %	1.1 %	4.9 %	5.5 %	2.5 %	3.0 %
25-50%	8.4 %	10.1 %	2.0 %	4.0 %	19.6 %	22.2 %	10.4 %	12.3 %
50-75%	40.0 %	50.5 %	38.5 %	61.0 %	44.5 %	47.2 %	39.9 %	50.5 %
75-100%	50.0 %	37.6 %	58.6 %	34.0 %	31.0 %	25.1 %	47.1 %	34.2 %

Supplementary Table 1: Theoretical coverage of 19,852 human proteins. Lower mass limit, 500; upper mass limit, 3500; maximum missed cleavages, 1. The percentage of proteins within each coverage group is given as all peptides (columns marked "All"), or only peptides containing at least one R, L, F, or P (columns marked "RFLP"). As an example, 50.0% of the proteins have a coverage degree between 75 to 100% when digested with trypsin, and this decreases to 37.6% if only peptides containing at least one R, L, F, or P are included.

	Average #peptides per protein	**Total #peptides all proteins**
Trypsin All	33.8	668266
Trypsin RFLP	31.0	612652
Chymotrypsin All	43.2	853327
Chymotrypsin RFLP	39.1	773232
GluC All	34.2	672133
GluC RFLP	30.9	607306

Supplementary Table 2: An overview of the number of of peptides with and without the constraint that all peptides have to include at least one R, F, L or P. Lower mass limit, 500; upper mass limit, 3500; maximum missed cleavages, 1; chymotrypsin cleaves after F, Y, W and L.

Trypsin (T) vs. chymotrypsin (C) all peptides

Average difference:	-2.9 %
Coverage T >= coverage C	46.2 %
Coverage T < coverage C	53.8 %

Trypsin (T) vs. gluC (G) all peptides

Average difference:	1.9 %
Coverage T >= coverage G	55.1 %
Coverage T < coverage G	44.9 %

Trypsin (T) vs. chymotrypsin (C) RFLP*

Average difference:	-1.2 %
Coverage T >= coverage C	50.4 %
Coverage T < coverage C	49.6 %

Trypsin (T) vs. gluC (G) RFLP*

Average difference:	2.0 %
Coverage T >= coverage G	55.4 %
Coverage T < coverage G	44.6 %

Supplementary Table 3: Comparison of the overall coverage degrees for 19852 human proteins theoretically digested by trypsin, chymotrypsin and gluC. Lower mass limit: 500, upper mass limit: 3500, maximum missed cleavages: 1. Chymotrypsin is used with the specificity FYWL.
*Only peptides containing R, F, L or P were included.

References

1. Swiss-Prot [http://au.expasy.org/sprot/]
2. Krause E, Wenschuh H, Jungblut PR: The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal Chem* 1999, 71:4160-4165.
3. Baumgart S, Lindner Y, Kühne R, Oberemm A, Wenschuh H, Krause E: The contributions of specific amino acid side chains to signal intensities of peptides in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* 2004, 18:863-868.