Proteomic database search engine for two-dimensional partial covariance mass spectrometry

Taran Driver,[†] Rüdiger Pipkorn,[‡] Leszek J. Frasiński,[†] Jon P. Marangos,[†] Marina Edelson-Averbukh,[†] and Vitali Averbukh[*,†]

[†] *The Blackett Laboratory, Imperial College London, Prince Consort Road, London SW7 2AZ, United Kingdom*

[‡] *German Cancer Research Centre, Department of Translational Immunology, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany*

E-mail: v.averbukh@imperial.ac.uk

Abstract

We present a protein database search engine for the automatic identification of peptide and protein sequences using the recently introduced method of two-dimensional partial covariance mass spectrometry (2D-PC-MS). Since 2D-PC-MS measurement reveals correlations between fragments stemming from the same or consecutive decomposition processes, the first-of-its-kind 2D-PC-MS search engine is based entirely on the direct matching of the pairs of theoretical and the experimentally detected correlating fragments, rather than of individual fragment signals or their series. We demonstrate that the high structural specificity afforded by 2D-PC-MS fragment correlations enables our search engine to reliably identify the correct peptide sequence, even from a spectrum with a large proportion of contaminant signals. While for peptides the

2D-PC-MS correlation matching procedure is based on complementary and internal ion correlations, the identification of intact proteins is entirely based on the ability of 2D-PC-MS to spatially separate and resolve the experimental correlations between complementary fragment ions.

Introduction

Tandem mass spectrometry (MS/MS) is the most prevalent method for the identification and characterization of the amino acid sequence and covalent modifications of proteins. In a typical bottom-up MS/MS experiment, proteins under analysis are subjected to enzymatic digestion to cleave the long amino acid sequence into smaller peptide chains. These peptide chains are introduced into the gas phase, e.g. *via* electrospray ionization,[1] and then fragmented by one of a number of different available fragmentation methods [e.g. collision-induced dissociation (CID),[2] electron transfer/capture dissociation,[3,4] photodissociation[5]]. The mass-to-charge ratios (m/z) of the resultant fragments and their relative abundances are recorded as MS/MS spectrum, and this information is used to piece together the structure of the original molecule under analysis.

The requirement for high-throughput protein identification has driven a large research effort in the automated analysis of MS/MS spectra.[6] The different techniques for the automated analysis of MS/MS spectra can be broadly divided into three categories, where amino acid sequences and post-translational modifications are inferred from the experimentally measured spectra by: (i) comparing to large libraries of previously recorded experimental spectra of known sequences,[7] (ii) comparing to ‘theoretical’ spectra generated *in silico* from databases of known protein and/or genetic sequences, according to generalised peptide fragmentation rules[8–10] or (iii) so-called *de novo* sequencing, where a first-principles reconstruction is performed directly from the measured experimental spectrum without recourse to any database.[11,12]

A lack of suitably large spectral databases and strong dependence of fragmentation patterns on the activation method and experimental conditions renders option (i) challenging, while option (iii) is generally limited by the incomplete sequence information available from a standard MS/MS spectrum. As a result, the most common method of automated spectral interpretation is the comparison of experimentally measured spectra to theoretical spectra generated *in silico* from sequence databases. The software which performs this analysis is generally referred to as a 'database search engine'. The key step in a database search engine is the identification of the individual peptide molecules (and their post-translational modifications) from the MS/MS spectra. Using this information, complemented by e.g. knowledge of the digestion enzyme used and measurement of the m/z value of the intact peptide ion, the structure of the full protein chain which originally underwent enzymatic digestion is inferred.

Despite the notable success of database search engines,[13–15] the technology still suffers from a number of important setbacks. A highly variegated problem is the variability of peptide decomposition patterns, which are affected by the particular amino acid sequence, the presence of PTMs, peptide length, secondary structure, etc.[16–18] This frequently leads to considerable deviation of the experimental mass spectra from the theoretical ones, predicted by on the basis of simplified peptide fragmentation rules. Additionally, many of the peptide fragments predicted by these generalised fragmentation rules (e.g. b- and y-type ions for CID[10]) can appear at very low abundances, and so are often missed during the spectrum-to-structure matching procedure. This can result in incorrect assignments even from mass spectra displaying standard fragmentation patterns. Moreover, 1D m/z-based MS analysis is often compromised by false identifications caused by fragments being attributed to isobaric (within a given mass tolerance) and isomeric ions of incorrect structures.

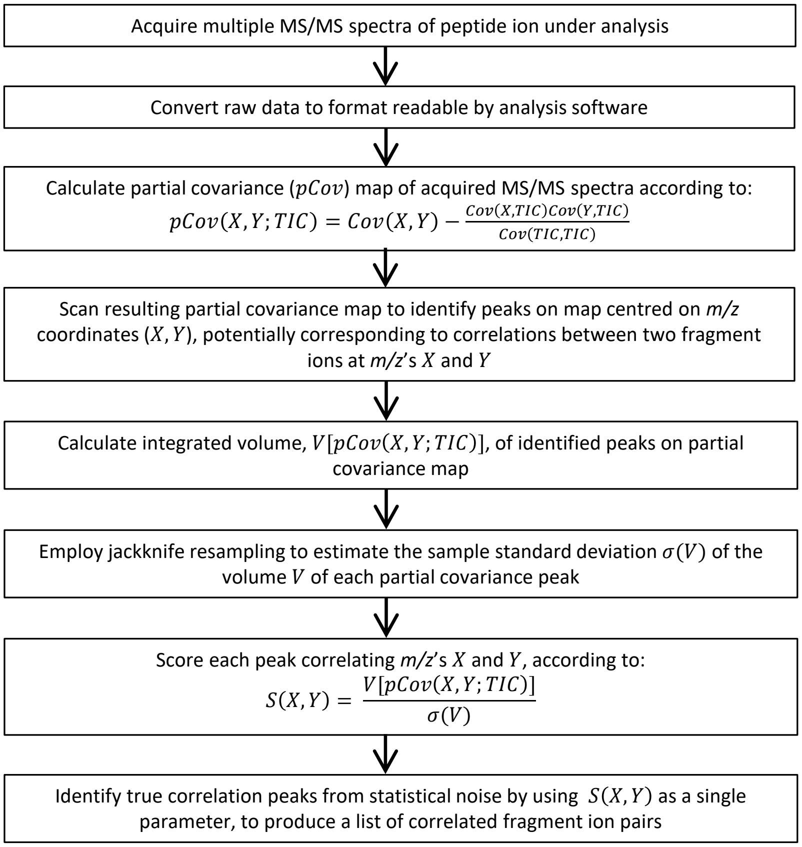
Two-Dimensional Partial Covariance Mass Spectrometry

Two-dimensional partial covariance mass spectrometry (2D-PC-MS) is a new kind of two-dimensional MS, based entirely on fragment-fragment correlations.[19,20] By mapping the fluctuations in fragment ion abundances across a series of repeated fragment mass spectra, 2D-PC-MS identifies fragment ions born in the same or consecutive decomposition reactions of the same parent molecular ion. This is in sharp contrast to the well familiar two-dimensional Fourier transform ion cyclotron resonance (2D FT-ICR) mass spectrometry, which identifies correlations between a *parent* ion and a fragment ion.[21] A significant fraction of the fragment-fragment correlations revealed by 2D-PC-MS have been shown to be much more structurally specific than the individual fragment ion signals of the 1D MS/MS.[19] The 2D spectral information accessible *via* 2D-PC-MS has been used to solve the long-standing problem of resolving diacetylated isomers of histone H3,[22] and its application to top-down measurements of intact proteins provides a straightforward means to deconvolve the overlapping fragment ion peaks of co-isolated intact proteins.[23]

2D-PC-MS is based on calculation of the self-correcting partial covariance between the signals of the fragment ions X and Y across a large number of individual fragment mass spectra (e.g. microscans of a linear ion trap mass spectrometer):

$$pCov(X,Y;\mathrm{TIC}) = Cov(X,Y) - \frac{Cov(X,\mathrm{TIC})Cov(Y,\mathrm{TIC})}{Cov(\mathrm{TIC},\mathrm{TIC})}\,, \qquad (1)$$

where $Cov(X,Y)$ is the simple covariance between the signals of X and Y, $Cov(X,Y) = \langle XY\rangle - \langle X\rangle\langle Y\rangle$, $\langle\ldots\rangle$ denoting the averaging over the individual fragment spectra. TIC is the total ion count within a given single fragment spectrum, i.e. the partial covariance parameter derived from the spectrum itself, see Ref.[19] for details. The fragment ions X and Y produced in the same or in the consecutive decomposition processes are characterised by positive peaks of the self-correcting partial covariance (1) on the 2D-PC-MS map, symmetrical with respect to the $m_X/z_X = m_Y/z_Y$ diagonal. If X and Y are two complementary fragments, their masses and charges are related to the mass, M and charge Z of the peptide ion by mass and charge conservation laws: $(m_X/z_X)\times z_X + (m_Y/z_Y)\times z_Y = M$, $z_X + z_Y = Z$. Therefore, 2D-PC-MS signals of correlations between complementary fragments lie on the straight lines with slopes



Acquire multiple MS/MS spectra of peptide ion under analysis

Convert raw data to format readable by analysis software

Calculate partial covariance ($pCov$) map of acquired MS/MS spectra according to:

$$pCov(X,Y;TIC) = Cov(X,Y) - \frac{Cov(X,TIC)Cov(Y,TIC)}{Cov(TIC,TIC)}$$

Scan resulting partial covariance map to identify peaks on map centred on *m/z* coordinates (X,Y), potentially corresponding to correlations between two fragment ions at *m/z*'s X and Y

Calculate integrated volume, $V[pCov(X,Y;TIC)]$, of identified peaks on partial covariance map

Employ jackknife resampling to estimate the sample standard deviation $\sigma(V)$ of the volume V of each partial covariance peak

Score each peak correlating *m/z*'s X and Y, according to:

$$S(X,Y) = \frac{V[pCov(X,Y;TIC)]}{\sigma(V)}$$

Identify true correlation peaks from statistical noise by using $S(X,Y)$ as a single parameter, to produce a list of correlated fragment ion pairs

Figure 1: Flowchart of the 2D-PC-MS: from individual MS/MS scans to self- correcting partial covariance map to identification and scoring of the fragment- fragment correlations on the partial covariance map.

$-z_1/z_2$, called primary mass conservation lines. Correlations between a terminal fragment and a complementary ion which underwent a neutral loss are offset from the primary mass conservation lines by the mass of the lost neutral molecule divided by the charge of its parent fragment. Finally, fragment correlations resulting from two peptide bond dissociations, for example those between a terminal and an internal ion form a manifold of scattered signals on the 2D-PC-MS map. These various types of fragment-fragment correlations revealed by 2D-PC-MS are described in detail in Ref.[19] Efficient discrimination between the true low-intensity 2D-PC-MS correlations (X,Y) and statistical noise stemming from the finite number of scans used for the self- correcting partial covariance in Eq. (1) is achieved using the peak score, $S(X,Y)$,

$$S(X,Y) = \frac{V[pCov(X,Y;\mathrm{TIC})]}{\sigma(V)}\,, \qquad (2)$$

where $V[pCov(X,Y;\mathrm{TIC})]$ is the 2D-PC-MS peak volume, and $\sigma(V)$ is the variance of the peak volume under jackknife resampling. The full flowchart of 2D-PC-MS is given in Fig. 1. Here we report the algorithm and the first results of the operation of 2D-PC-MS database search engine for sequencing of peptides and proteins.

Experimental

Materials and Peptide Synthesis

Water, acetonitrile and formic acid used for the MS analysis were of Optima LC-MS grade and were purchased from Fisher Scientific Ltd. Ammonium acetate was of LC-MS Chromasolv grade and manufactured by Fluka Analytical and triethylamine was of analytical standard and purchased from Sigma Aldrich Company Ltd. Ubuquitine, Myoglobine and Cytochrome c were purchased from Sigma Aldrich.

For the solid phase synthesis of all peptides the Fmoc (9-fluorenyl-methyloxycarbonyl) methodology[24] was employed, using a fully automated multiple synthesizer (Syro II from

Multi Syntech Germany). The peptide synthesis was carried out on preloaded Wang resins. Peptide chain assembly was performed by *in situ* activation of amino acid building blocks by 2-(1H-benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate.

FmocThr(PO(OBzl)OH)-OH and Fmoc-Tyr(PO(OBzl)OH)-OH were purchased from Merck Millipore GmbH. Fmoc-Tyr(Nitro)-OH, Fmoc-Arg-(Me)-OH, Fmoc-Arg(sMe2)-OH and Fmoc-Lys(Ac)-OH were purchased from Bachem AG. The synthesized peptides were purified by preparative HPLC on a Kromasil (AkzoNobel/Sweden) 100–10C 10 μm 120 Å reverse phase column (30 × 250 mm) using an eluent of 0.1% trifluoroacetic acid in water (A) and 80% acetonitrile in water (B). The peptides were eluted with a successive linear gradient of 25and lyophilized. The purified peptides were characterized with analytical HPLC and MS (Thermo Finnigan, LCQ).

MS analysis and data processing

We have implemented 2D-PC-MS on a Thermo Fisher Scientific LTQ XL instrument. The instrument required no modification. The MS/MS scans were performed at a scan rate of 125 000 Da/s, with AGC MS$^{\mathrm{n}}$ target values of 10^2 for single peptides and 3×10^2 for intact proteins. The self-correcting partial covariance maps have been built according to Eq. (1) using 10000 microscans. For the 2D-PC-MS measurements, the synthetic peptides were dissolved in 50% acetonitrile / 2% formic acid in water to produce concentrations of 100 fM to ≈ 1 μM.

For the top-down measurements protein samples were prepared to a concentration of $\approx 1\,\mu$l in a solution of 1% formic acid and 50% acetonitrile in water. The samples were infused into the mass spectrometer via a Harvard Apparatus 11 Plus Single Syringe Pump coupled to a Nanospray II Ion Source (Thermo Fisher Scientific) at a flow rate of 3–5 μl/min (1 μl/min for protein samples) and spray voltage of 1.8–2.2 kV in positive ion mode, and a flow rate of 1 μl/min using no auxiliary desolvation gas. The temperature of the ion transfer capillary was held constant at 200 °C. The parent ions of interest were fragmented by collision-induced dissociation at normalized collision energies of 35% (for peptides)and 70% (for proteins), with activation time of 30 ms and Mathieu q-value of 0.25.

The 2D-PC-MS data processing (see Fig. 1) was carried out by in-house computer code written in Python (2.7) using numerical routines from the NumPy (http://www.numpy.org/) and SciPy (http://www.scipy.org/) libraries. The software reads in the MS/MS raw data in text file format and calculates the TIC partial covariance (pCov) between each pair of m/z channels in the tandem mass spectra using Eq. (1). Another Python code was written for processing the resulting 2D-PC-MS maps to produce the scored lists of fragment ion correlations. This code first determined the features of a 2D-PC-MS map potentially corresponding to true correlation peaks, according to the height of their apices, followed by the calculation of the 2D-PC-MS correlation score using Eq. (2).

For standard 1D MS/MS measurement of the mixture of palindromic sequences, $\approx 1\mu$M to $\approx 10\mu$M peptide solutions in 50% acetonitrile / 2% formic acid were used, at 16 666 Da/s scan rate to increase the mass resolution and 5000 microscan averaging.

All components of the 2D-PC-MS search engine (e.g. *in silico* digestion & modification, *in silico* fragmentation, fragment correlation matching) were written from scratch in the Python (2.7) programming language, making use of the NumPy and SciPy libraries.

2D-PC-MS search engine

The vast majority of the state-of-the-art 1D MS database search engines[25] are based on scoring the candidate (database) sequences by matching their predicted (theoretical) fragment spectra to the results of the MS/MS measurement. This sequence scoring procedure may be executed in two stages, for example by applying the cross-correlation method to the a number of top scoring sequences ranked by the direct comparison between the experimental and the theoretical spectra (as is done in the SEQUEST algorithm[26]). Within all such algorithms, the fundamental unit of similarity between the theoretical and the measured MS/MS spectrum is a matched fragment ion signal, i.e. matching of a single experimental m/z value to a single theoretical m/z value. Therefore, the success of all such search engines ultimately rests on the selectivity of the fragment ion matching process. Modern mass spectrometry has made enormous advances in improving the selectivity of the fragment ion matching through increasing the precision and the resolution of the m/z measurement of the fragment ions.[27,28] Nevertheless, a large fraction, $\sim 60\%$, of MS/MS spectra (a significant proportion of those being high quality) still remains unassigned to the correct peptide and protein sequences.[29]

2D-PC-MS offers a different approach to improving the selectivity of spectrum to sequence matching, the one where the fundamental unit of comparison is no longer a single fragment but rather a pair of correlating fragments. As we have demonstrated by numerical simulations,[19] the selectivity provided by the 2D-PC-MS fragment-fragment correlations involving internal fragments in peptides of the lengths typical of tryptic digests is dramatically higher than can be available from any single fragment signals, even in the theoretical limit of an infinitely precise m/z measurement. The proportion of such ultra-high specificity signals among all the fragment-fragment correlations measured in 2D-PC-MS is expected to depend on the peptide length, peptide charge state and the activation technique. For example, their proportion observed in the triply charged peptides of $\approx 1.0 \div 1.6$ kDa mass studied in Ref.[19] upon CID is about 25%. This significant share of the ultra-high specificity signals suggests that a database search engine based on matching 2D-PC-MS fragment-fragment correlations instead of 1D m/z values of individual fragment ions has the potential to provide highly accurate peptide sequence matches.

Moreover, 2D-PC-MS maps readily expose various types of fragment-fragment correlations (between complementary terminal fragments, between a terminal and an internal product ion, between the neutral loss products of terminal/internal ions) by separating them geometrically on the 2D map. For example, the pairs of complementary terminal fragment ions can be readily identified as lying on the primary mass conservation lines purely on the basis of the mass and the charge state of the peptide ion, without any knowledge or assumption about the peptide sequence. This is in contrast to 1D MS, where any m/z signal can a priori belong to any type of fragment (terminal, terminal with a neutral loss, internal or internal with a neutral loss). This useful feature of 2D-PC-MS makes it especially straightforward to give different weights to matching the correlations involving different types of fragments, e.g. giving more weight to matching the ultra-high specificity internal-terminal correlations against those between the complementary terminal fragments.

Here we demonstrate both peptide and intact protein identification through constructing and applying a prototypical 2D-PC-MS search engine which includes no additional steps beyond straightforward evaluation of the weighted overlap between the measured and the theoretical fragment-fragment correlations. A schematic of the prototypical 2D-PC-MS search engine is shown in Fig. 2. The search engine accepts as input a measured precursor m/z and charge, a list of ranked fragment-fragment correlations derived from a 2D-PC-MS spectrum by peak scoring,[19] a series of user-specified parameters and a database of protein amino acid sequences. The engine first performs enzymatic digestion of the database protein sequences *in silico*, with the additional possibility to perform a non-specific digest (cleavage at all possible sites considered). Expected post-translational modifications can be specified as either fixed (every occurrence of the specified residue features the modification) or variable (all possible combinations are considered with the specified residue either modified or not). For each structure under analysis, digestion products corresponding to possible sequence matches are identified based on a measured parent m/z value and user-defined fractional parent m/z tolerance. Computational overhead is limited for large digests by performing m/z-based precursor selection on-the-fly during the digestion step.

Each of the candidate peptide sequences selected on the basis of the molecular ion m/z are subjected to 2D-PC-MS correlation matching against the measured 2D-PC-MS spectrum represented as a list of $(m_X/z_X, m_Y/z_Y)$ fragment-fragment correlations and their scores calculated according to Eq. (2). *In silico* fragmentation is performed according to experimentally derived CID fragmentation rules from our own 2D-PC-MS measurements[19] – the three following categories of fragment ion are considered for comparison with the experimental correlation signals:

(i) complementary terminal b-ion & y-ion correlations;

(ii) complementary terminal b-ion & y-ion correlations involving a neutral loss of either H_2O or NH_3 from one or both of the correlated fragment ions, either with or without CO loss from the b-ion (producing an a-ion);

(iii) terminal b-ion & internal b-ion and internal b-ion & terminal y-ion correlations, including those followed by a neutral loss of either H_2O or NH_3 from one or both of the correlated fragment ions, either with or without CO loss from any of the b-type fragment ions (producing an a-type ion).

For categories (i) and (ii), all fragment ion correlations where the charge state sums to the measured precursor charge state Z or lower are considered. For correlations in category (iii), all fragment ion correlations where the charge state sums to $Z-1$ are considered, except when $Z = 2$ in which case all fragment ion correlations where the charge state sums to Z are considered. Limiting the possible charge states for type (iii) correlations reflects the fact that for precursor ions with $Z \geq 3$, type (iii) correlations are rarely observed without being accompanied by the loss of at least one charge. We have also found this to increase selectivity of the 2D search engine. Although CO loss is limited to only occur from a b-type ion (either terminal b-ion or internal b-ion), loss of both H_2O and NH_3 is independent of the type or amino acid sequence of the fragment ion. This is a departure from the algorithms of many 1D database search engines, which restrict these small molecule neutral losses to only the most likely residues to improve search specificity. For the 2D-PC-MS prototype search engine, we have found such a restriction to be unjustified as it causes structure-specific correlation signals to be missed.

The experimentally measured 2D-PC-MS correlations are selected to be included into the sequence matching procedure according to their scores calculated using Eq. (2). The correlation score represents a robust, global 2D-PC-MS peak selection measure that, within our matching algorithm, serves both for peak detection and for sequence matching.

Specifically, for each candidate peptide sequence, falling within the specified m/z tolerance from the measured molecular ion, the peptide/protein score, PS, is calculated according to:

$$PS = \sum_{i=1}^{N} \tilde{S}(X_i, Y_i) \times P(X_i, Y_i) \times W(X_i, Y_i)\,, \tag{3}$$

where the sum runs over the N top scoring 2D-PC-MS peaks correlating fragment ions X_i and Y_i (optimal N has been empirically determined to be 50 for triply charged peptide ions and 40 for doubly charged peptide ions), $\tilde{S}(X_i, Y_i)$ is the normalised correlation score of Eq. (2), $\sum_i \tilde{S}(X_i, Y_i) = 1$, $P(X_i, Y_i)$ is either 1 or 0, depending on whether two fragment ions with m/z ratios within a given m/z tolerance from $m_{X_i}/z_{X_i}, m_{Y_i}/z_{Y_i}$ can $[P(X_i, Y_i) = 1]$ or cannot $[P(X_i, Y_i) = 0]$ be produced by the candidate peptide sequence, and finally $W(X_i, Y_i)$ is a weight that depends on the category, (i), (ii) or (iii), to which the fragments of the candidate sequence matching X_i and Y_i within the m/z tolerance belong. The weights $W(X_i, Y_i)$ have been empirically optimised across a series of measurements of synthetic peptide sequences (see Table I of Ref.[19]), resulting in: $W(X_i, Y_i) = 0.8$ for type (i) correlations, $W(X_i, Y_i) = 0$ for type (ii) correlations and $W(X_i, Y_i) = 1.0$ for type (iii) correlations. The highest score for type iii) correlations is in agreement with the higher calculated estimated false positive rate for matching these correlations.[19] The zero weight of type (ii) correlations is a result of the fact that in 2D-PC-MS, the appearance of a b-ion & y-ion correlation with at least one ion suffering neutral loss of a small molecule (H_2O, NH_3, CO) without detection of the corresponding intact b-ion & y-ion correlation is vanishingly rare. In principle, the experimentally measured correlations could be pre-separated according to their geometric position on the 2D-PC-MS map. This would enable the immediate identification of e.g. complementary b-ion & y-ion correlations and so negates the requirement to test these correlations against theoretical type (ii) or (iii) correlations. In practice, the precursor m/z

filtering already nullifies the danger of any such avoidable inter-category matching so any pre-separation would be for computational efficiency only and is not performed here.

Search Engine Performance

Peptide Sequences

The search engine was first tested on a series of peptide ions investigated in Ref.[19] which contains also the details of the experimental 2D-PC-MS procedure. All searches were performed over a non-specific digest of the UniProtKB/Swiss-Prot database,[30] at precursor m/z tolerance 5 ppm and fragment ion tolerance 0.8 Da. For the search engine runs, the precursor ion mass was manually adjusted in order to simulate the standard MS/MS workflow with a high (5 ppm) mass accuracy precursor ion measurement.

Fig. 3 illustrates the performance of the candidate sequence scoring of Eq. (3) for peptide identification. The correct sequence is identified by virtue of its outstanding 2D-PC-MS peptide score, which is almost twice as large as that of the two next highest-scoring sequences. The matched experimental signals are illustrated in the scatter plots. The greatest differentiation is found in the internal ions: the correct sequence matches with almost three times as many internal ions than the closest false match. Given the large number of peptide sequences tested (note the logarithmic scale on the y- axis), this illustrates the remarkable sensitivity of internal ions correlations for peptide identification. Note also the prevalence of such correlations (17 are identified for the this length 13 sequence).

Fig. 3 and Fig. 4a-e demonstrate the performance of the 2D-PC-MS search engine for six individual peptide sequences of different charge and modification states. The user-specified potential modifications are provided in the caption. The results show that the simple weighted overlap peptide scoring of the 2D-PC-MS search engine is straightforwardly able to identify the correct sequence from a large pool of candidate peptide sequences (note the logarithmic scale of the y-axes in Fig. 4).

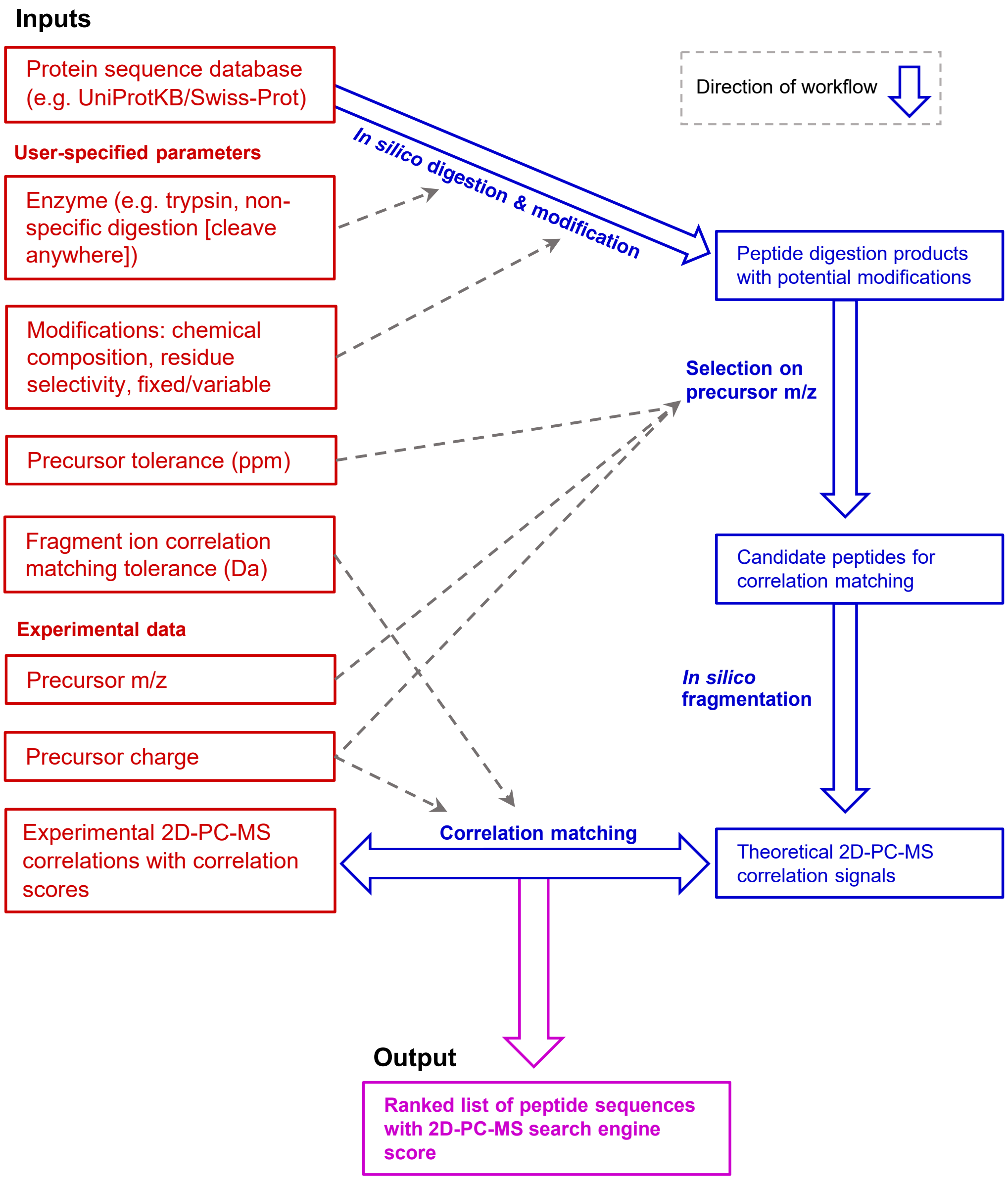

Inputs
Protein sequence database (e.g. UniProtKB/Swiss-Prot)
User-specified parameters
Enzyme (e.g. trypsin, non-specific digestion [cleave anywhere])
Modifications: chemical composition, residue selectivity, fixed/variable
Precursor tolerance (ppm)
Fragment ion correlation matching tolerance (Da)
Experimental data
Precursor m/z
Precursor charge
Experimental 2D-PC-MS correlations with correlation scores
Direction of workflow
In silico digestion & modification
Peptide digestion products with potential modifications
Selection on precursor m/z
Candidate peptides for correlation matching
In silico fragmentation
Theoretical 2D-PC-MS correlation signals
Correlation matching
Output
Ranked list of peptide sequences with 2D-PC-MS search engine score

Figure 2: Schematic for the prototypical 2D-PC-MS search engine based on fragment-fragment correlations.

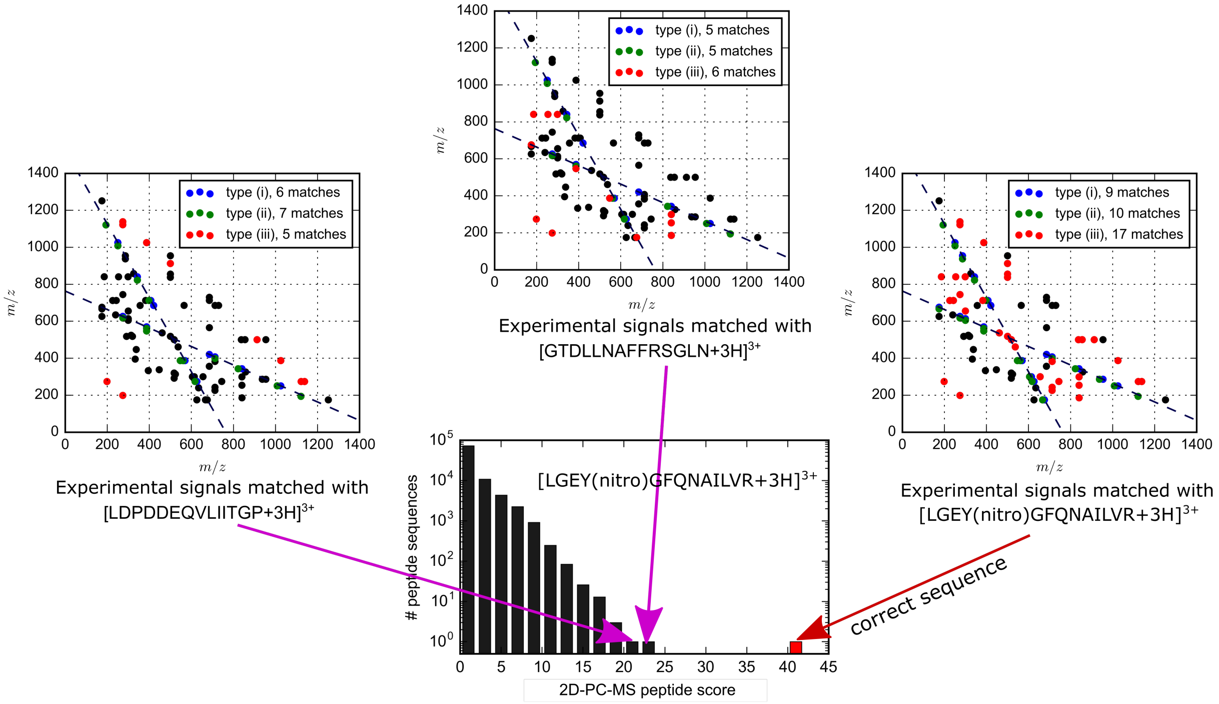


Figure 3: Illustration of 2D-PC-MS search engine performance for the triply charged peptide ion [LGEY(nitro)GFQNAILVR+3H]$^{3+}$. The measured 2D-PC-MS correlation signals are reproduced across each of the three scatter plots. The blue dashed lines indicates the mass conservation lines along which the complementary terminal fragment ion correlations [type (i) correlations] lie. Each scatter plot shows the experimental signals which matched to the expected 2D-PC-MS signals of the given database sequence (see legends). The correct sequence is identified by its outstanding 2D-PC-MS peptide score, Eq. (3), produced by matching a substantially larger number of correlation signals than any other database sequence.

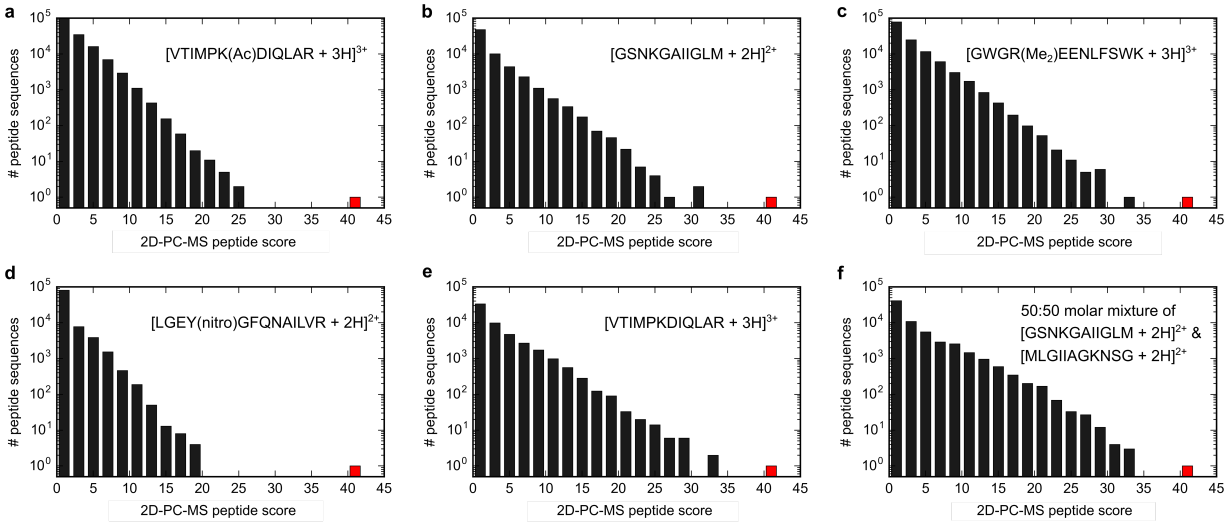


Figure 4: 2D-PC-MS search engine results for a selection of different peptide sequences with different charge and modification states, including the scenario of abundant contamination with an unnatural contaminant ion. Note the logarithmic scale on the y-axes of all plots, showing the high sequence selectivity of the search engine. The 2D-PC-MS peptide score is calculated according to Eq. (3). Modifications specified were: **a**: lysine acetylation (variable), **b**: no modification, **c** arginine dimethylation (variable), **d**: tyrosine nitration (variable), **e**: no modification, **f**: no modification . In panel **d** the isomeric database sequences LGEY(nitro)GFQNALLVR and LGEY(nitro)GFQNALIVR, which were trivially awarded the same 2D-PC-MS search engine score, are not plotted on the histogram.

We also investigated the performance of the 2D-PC-MS search engine in the challenging situation of abundant contaminant fragment ion signals, which can significantly disrupt the performance of 1D database search engines. The contaminant signals often arise as a result of the co-isolation and co- fragmentation of contaminant isobaric or isomeric species, which is estimated to occur for up to 50% of MS/MS spectra in a standard MS/MS run.[31,32] To compare the performance of the 2D-PC-MS search engine and the state-of-the-art 1D MS database search, the amyloid beta peptide GSNKGAIIGLM was prepared in an equimolar mixture with its unnatural (therefore not part of any sequence database) palindromic isomer MLGIIAGKNSG. Fig. 4f demonstrates the performance of the 2D-PC-MS search engine in the case of abundant unnatural contaminant ions. For this measurement we also tested the Mascot search engine,[33] using identical search parameters. Whilst the 2D-PC-MS search engine correctly identifies the natural peptide sequence despite the contaminant ion presence, Mascot fails to do so, producing three dissimilar sequences with scores higher than the correct one, see Supplementary Information.

Top Down Analysis

Top down mass spectrometry refers to the direct fragmentation of intact protein ions.[34] 2D-PC-MS presents a number of benefits for top down mass spectrometry.[23] For example, the 2D information available from 2D-PC-MS enables the direct identification of pairs of complementary fragment ions (e.g., b/y fragments in CID measurements), which can be readily distinguished from the multitude of other fragment ion signals produced *via* secondary fragmentations and other non-trivial fragmentation pathways. As discussed above, these complementary pairs are identified as falling along the mass conservation lines, a set of straight lines on the 2D-PC-MS map which are fully defined by the mass and charge state of the parent ion. Peculiarly, the gradient of the line along which a particular signal lies provides the charge state of both the correlated fragment ions,[23] without the requirement for isotopic envelope resolution. Moreover, within top-down 2D-PC-MS it also turns out to be possible to separate fragment-fragment correlations resulting from the overlapping MS/MS spectra obtained by co-isolation and co-fragmentation of multiple parent protein ions.[23]

The utility of directly identifying pairs of complementary sequence ions becomes stark in the case of the intact protein molecules. The number of internal ions possible for a sequence of length n, and therefore also the number of possible internal-terminal ion correlations, scales as n^2 (assuming a minimum length of two residues for b-type ions, the number of the possible internal fragments is $(n-2)(n-3)/2)$. The higher charge states encountered in top-down mass spectrometry further inflate the number of internal ions that could be produced by a single intact protein ion. On the other hand, the number of the complementary fragment correlations scales only linearly with the sequence length. The unfavourable scaling of the number of internal ions and their correlations reduces the sequence specificity of the internal ion correlations in 2D-PC-MS with the sequence length, making the complementary terminal fragment correlations the primary basis for the top-down 2D-PC-MS sequence identification.

In Ref.,[23] we performed top-down measurements of the intact proteins ubiquitin^{10+}, cytochrome C^{12+}, and myoglobin^{13+} under CID using a linear ion trap mass spectrometer.[35] Here we provide the 2D-PC-MS search engine with the scored fragment-fragment correlations extracted from these 2D-PC-MS measurements. In light of the reduced sequence specificity of internal ion correlations for the long sequences, only the complementary fragment ion correlations are provided to the 2D-PC-MS search engine. Fig. 5 shows the results of the 2D-PC-MS search engine run for these top down measurements. The 2D-PC-MS search engine successfully identifies all three intact protein ions directly from the top-down 2D-PC-MS linear ion trap measurement. The correct protein sequence is assigned an outstanding protein score among the large number of candidate protein sequences tested.

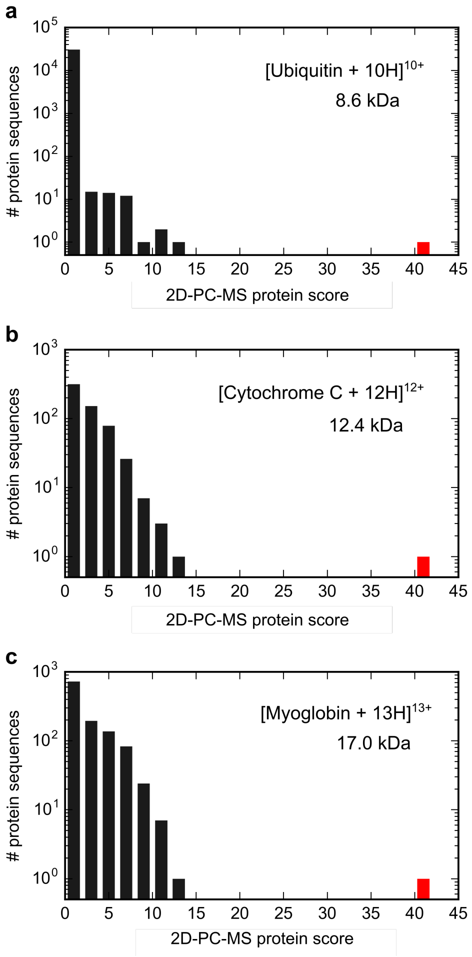


Figure 5: Performance of the 2D-PC-MS search engine for intact protein molecules measured in a linear ion trap. Note the logarithmic scale on the y-axes, demonstrating the large number of sequences subjected to fragment ion matching. The correct sequence is identified by virtue of an outstanding 2D-PC-MS protein score, Eq. (3).

Conclusions

In conclusion, we have developed a prototypical proteomic database search engine, which identifies an experimentally measured peptide or protein sequence from a database search based on the matching of 2D-PC-MS fragment correlation signals. The search engine reliably identifies the correct sequence using a simple single-stage matching algorithm both for peptides and for proteins in the mass range of $\sim 1 \div 20$ kDa, and is able to outperform the state-of-the-art 1D database search engine in the case of spectral contamination. The high specificity of 2D-PC-MS signals allows for robust sequence identification even in the presence of abundant contaminant fragment ions peaks. By exploiting the ability of 2D-PC-MS to identify experimentally complementary sequence ion pairs, we extend the search engine to the analysis of intact proteins.

Acknowledgement

M.E.A. acknowledges support of the Wellcome Trust through the research fellowship Grant No. WT100093MA. Support is provided by EPSRC program Grants No. EP/I032517/1 and EPSRC/DSTL MURI No. EP/N018680/1, Pathways to Impact Grant No. EP/K503733/1, and ERC ASTEX Project No. 290467.

References

(1) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246*, 64–71.

(2) Shukla, A. K.; Futrell, J. H. Tandem mass spectrometry: Dissociation of ions by collisional activation. *Journal of Mass Spectrometry* **2000**, *35*, 1069–1090.

(3) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and

protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 9528–33.

(4) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society* **1998**, *120*, 3265–3266.

(5) Brodbelt, J. S. Photodissociation mass spectrometry: New tools for characterization of biological molecules. *Chemical Society Reviews* **2014**, *43*, 2757–2783.

(6) Nesvizhskii, A.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **2007**, *4*, 787–797.

(7) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655–667.

(8) Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nature Methods* **2004**, *1*, 195–202.

(9) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences* **1986**, *83*, 6233–6237.

(10) Wysocki, V. H.; Resing, K. A.; Zhang, Q.; Cheng, G. Mass spectrometry of peptides and proteins. *Methods* **2005**, *35*, 211–222.

(11) Ma, B.; Johnson, R. De Novo Sequencing and Homology Searching. *Molecular & Cellular Proteomics* **2012**, *11*.

(12) Medzihradszky, K. F.; Chalkley, R. J. Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrometry Reviews* **2015**, *34*, 43–63.

(13) Eng, J. K.; Mccormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* **1994**, *5*, 976–989.

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

(15) Cottrell, J. S. Protein identification using MS/MS data. *Journal of Proteomics* **2011**, *74*, 1842–1851.

(16) Tsaprailis, G.; Nair, H.; Somogyi, r.; Wysocki, V. H.; Zhong, W.; Futrell, J. H.; Summerfield, S. G.; Gaskell, S. J. Influence of secondary structure on the fragmentation of protonated peptides. *Journal of the American Chemical Society* **1999**, *121*, 5142–5154.

(17) Paizs, B.; Suhal, S. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews* **2005**, *24*, 508–548.

(18) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical Chemistry* **2005**, *77*, 5800–5813.

(19) Driver, T.; Cooper, B.; Ayers, R.; Pipkorn, R.; Patchkovskii, S.; Averbukh, V.; Klug, D. R.; Marangos, J. P.; Frasinski, L. J.; Edelson-averbukh, M. Two-Dimensional Partial-Covariance Mass Spectrometry of Large Molecules Based on Fragment Correlations. *Physical Review X* **2020**, *10*, 41004.

(20) Miller, J. L. Biomolecule mass spectrometry enters a new dimension. *Physics Today* **2020**,

(21) Pfändler, P.; Bodenhausen, G.; Rapin, J.; Houriet, R.; Gäumann, T. Two-Dimensional Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Chem. Phys. Lett.* **1987**, *138*, 195–200.

(22) Driver, T.; Pipkorn, R.; Averbukh, V.; Frasinski, L. J.; Marangos, J. P.; Edelson-Averbukh, M. Breaking the Histone Code with Two-Dimensional Partial Covariance Mass Spectrometry. *ChemRxiv:12743669* **2020**,

(23) Driver, T.; Averbukh, V.; Frasinski, L. J.; Marangos, J. P.; Edelson-Averbukh, M. Two-dimensional partial covariance mass spectrometry for the top-down analysis of intact proteins. *arXiv:2004.11949* **2020**, 1–30.

(24) Merrifield, R. B. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *Journal of the American Chemical Society* **1963**, *85*, 2149–2154.

(25) Verheggen, K.; Ræder, H.; Frode S. Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews* **2017**, *39*, 292–306.

(26) Eng, J.; McCormack, A. L.; Yates III, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(27) Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* **2011**, *10*, M111.011015.

(28) Weisbrod, C. R.; Hoopmann, M. R.; Senko, M. W.; Bruce, J. E. Performance evaluation of a dual linear ion trap-Fourier transform ion cyclotron resonance mass spectrometer for proteomics research. *Journal of Proteomics* **2013**, *88*, 109–119.

(29) Pathan, M.; Samuel, M.; Keerthikumar, S.; Mathivanan, S. Unassigned MS/MS Spectra: Who Am I? *Methods Mol Biol* **2017**, *1549*, 67–74.

(30) Bateman, A. et al. UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **2017**, *45*, D158–D169.

(31) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N.; Old, W. Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomic studies. *Journal of Proteome Research* **2010**, *9*, 4152–4160.

(32) Luethy, R.; Kessner, D. E.; Katz, J. E.; MacLean, B.; Grothe, R.; Kani, K.; Faça, V.; Pitteri, S.; Hanash, S.; Agus, D. B.; Mallick, P. Precursor-ion mass re-estimation improves peptide identification on hybrid instruments. *Journal of Proteome Research* **2008**, *7*, 4031–4039.

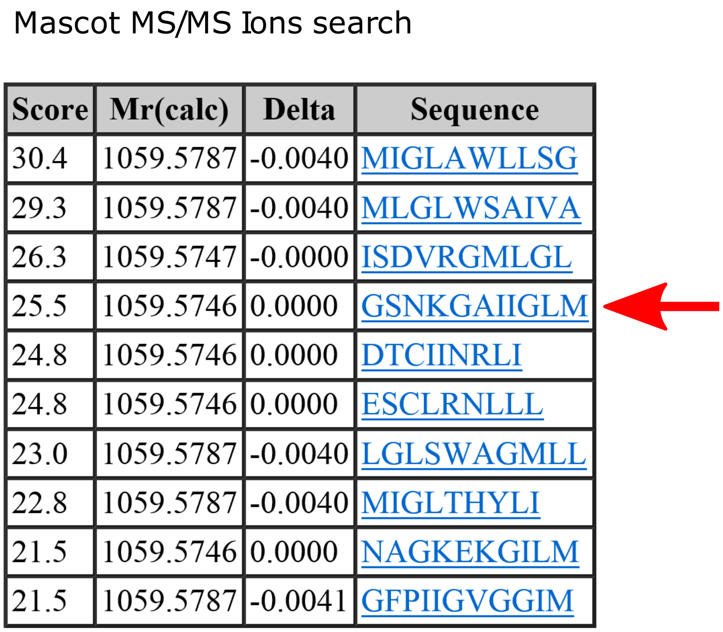
(33) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(34) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in top-down proteomics and the analysis of proteoforms. *Annual Review of Analytical Chemistry* **2016**, *9*, 499–519.

(35) Douglas, D. J.; Frank, A. J.; Mao, D. Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews* **2005**, *24*, 1–29.

Supporting Information Available

The Mascot search for the 50/50 mixture of the palindromic peptides, GSNKGAIIGLM and MLGIIAGKNSG was performed using a non-selective digest of the UniProtKB/Swiss-Prot database, with precursor m/z tolerance 5 ppm and fragment ion m/z tolerance 0.8 Da. No modifications were specified. The correct database sequence was ranked fourth by the Mascot search, see Fig. 6.



Mascot MS/MS Ions search

Score	Mr(calc)	Delta	Sequence
30.4	1059.5787	-0.0040	MIGLAWLLSG
29.3	1059.5787	-0.0040	MLGLWSAIVA
26.3	1059.5747	-0.0000	ISDVRGMLGL
25.5	1059.5746	0.0000	GSNKGAIIGLM
24.8	1059.5746	0.0000	DTCIINRLI
24.8	1059.5746	0.0000	ESCLRNLLL
23.0	1059.5787	-0.0040	LGLSWAGMLL
22.8	1059.5787	-0.0040	MIGLTHYLI
21.5	1059.5746	0.0000	NAGKEKGILM
21.5	1059.5787	-0.0041	GFPIIGVGGIM

Figure 6: Results of the Mascot database search (peptide score) for the 50/50 mixture of the palindromic peptides, GSNKGAIIGLM and MLGIIAGKNSG.