Journal of proteome research

Article

pubs.acs.org/jpr

Identification of Related Peptides through the Analysis of Fragment Ion Mass Shifts

Thomas Wilhelm*[,†] and Alexandra M. E. Jones[‡,§]

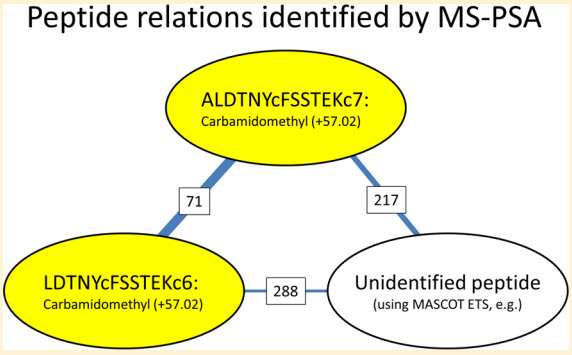
[†]Institute of Food Research, Norwich Research Park, Norwich NR4 7UA, United Kingdom

[‡]The Sainsbury Laboratory, Norwich Research Park, Norwich NR4 7UH, United Kingdom

[§]School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom

Supporting Information

ABSTRACT: Mass spectrometry (MS) has become the method of choice to identify and quantify proteins, typically by fragmenting peptides and inferring protein identification by reference to sequence databases. Well-established programs have largely solved the problem of identifying peptides in complex mixtures. However, to prevent the search space from becoming prohibitively large, most search engines need a list of expected modifications. Therefore, unexpected modifications limit both the identification of proteins and peptide-based quantification. We developed mass spectrometry–peak shift analysis (MS-PSA) to rapidly identify related spectra in large data sets without reference to databases or specified modifications. Peptide identifications from established tools, such as MASCOT or SEQUEST, may be propagated onto MS-PSA results. Modification of a peptide alters the mass of the precursor ion and some of the fragmentation ions. MS-PSA identifies characteristic fragmentation masses from MS/MS spectra. Related spectra are identified by pattern matching of unchanged and mass-shifted fragment ions. We illustrate the use of MS-PSA with simple and complex mixtures with both high and low mass accuracy data sets. MS-PSA is not limited to the analysis of peptides but can be used for the identification of related groups of spectra in any set of fragmentation patterns.



KEYWORDS: *mass spectrometry, MS/MS, data analysis, proteomics, protein modification*

INTRODUCTION

Currently, the most widely used method to identify proteins and their post-translational modifications is through the analysis of enzymatically derived peptides by tandem mass spectrometry (MS). Mass spectrometers coupled to liquid chromatography systems are capable of routinely analyzing complex mixtures of peptides and may generate tens of thousands of mass spectra per hour. Typically only a small fraction of the spectra acquired contribute to the identification of peptides. Part of this inefficiency is due to acquisition of spectra when peptides are not eluting, limitations of sensitivity, and sample contamination with nonpeptide molecules. However, many unidentified spectra remain that show characteristics of peptide fragmentation, such as regularly spaced ions.

Computational methods for assigning peptide identifications to MS/MS spectra can be grouped into two main categories: database similarity searches and de novo techniques, which directly reconstruct peptide sequences from spectra. Database searches can be further subdivided into two different approaches that use either theoretical spectra predicted from peptide sequences or empirical high-quality spectra (spectral libraries). Well-known search engines that use theoretical spectra are SEQUEST,[1] MASCOT,[2] X!Tandem,[3] and OMSSA,[4] among others. However, because of insufficient understanding of the factors that determine peptide fragmentation, most current search tools employ simplified fragmentation models, such as the uniform backbone dissociation model, leading to many unidentified or misidentified spectra.[5] Spectral library searches, introduced by Yates et al.,[6] became important with the availability of millions of confidently identified MS/MS spectra. MS library search tools include SpectraST,[7] NIST MSPepSearch (http://peptide.nist.gov/), BiblioSpec,[8] X!Hunter,[9] and ProMEX.[10]

Theoretically, de novo techniques should have advantages over database search algorithms because they can identify protein variants not contained in a database, but database search algorithms have historically outperformed de novo algorithms.[11] However, the development of improved resolution and mass accuracy methods such as high-energy collisional dissociation (HCD) has reignited research on de novo methods (pNovo[12]). Hybrid similarity search–de novo methods have also been suggested.[13,14]

Received: April 3, 2014
Published: July 24, 2014

A current challenge in proteomics is to rapidly identify modified peptides in complex mixtures. As powerful tandem MS becomes available to more researchers and many basic problems in identifying peptides have been resolved, the ability to identify peptides and infer protein sequences has become commonplace. The attention of the proteomics community is now focused on quantification using a variety of methods based on peak area of precursor ions, selective measurement of precursor-fragment transitions, fragmentation of all precursor ions (e.g., MS^E from Waters), or spectrum counting. All of these quantitative methods use peptide ions as measurable proxies for the protein(s) they represent, and the correct assignment of peptide sequence to a spectrum is critical. Unexpected modifications can limit identification of peptides and further confound quantification as unidentified modified forms of an identified peptide will reduce the measured signal.

Identification of peptide modifications is of tremendous importance not only for better identification of peptides but also in its own right. Post-translational modifications (PTMs) of proteins play a central role in regulating their cellular function. Proteins are also modified outside the cell, for example, during industrial food processing and gastrointestinal digestion. These modifications have important implications for human health, nutrition, and allergenic responses. Peptides may also be modified during sample preparation (i.e., extraction and purification of proteins) or miscleaved by proteolytic enzymes prior to analysis by MS. Here we use the term PTM as a proxy for *any* peptide modification regardless of its origin. Hundreds of different PTMs have been identified so far; the majority are simple ones such as phosphorylation,[15−17] and it is likely that many more remain to be discovered.[18]

Specific PTM identification techniques are often developed by adjusting general peptide sequence assignment tools. Database searches are widely used, generally achieved by a procedure first proposed by Yates et al.,[6] where users prespecify which modifications are expected. Specifying several modifications increases the search space exponentially and increases the potential for false discoveries (PTMFinder,[18] PeaksPTM[19]). MASCOT deals with this problem by a two-stage process to identify unspecified modifications, called "Error Tolerant Searches" (ETS): (i) spectra are used for a database search and (ii) peptides robustly identified in the first stage are used for a more relaxed search using all mass shifts that are specified in a modification database. The limitations of the ETS approach are (a) the protein must have at least one high scoring peptide in the first stage and (b) the mass of the modification must be specified to the search algorithm. These limitations represent a critical bottleneck in the interpretation of peptide mass spectra. Granholm et al.[20] identified PTMs as a problem leading to incorrectly assigned spectra. Other PTM finders use a sequence-tag-based search where de novo sequence tags are computed from MS/MS spectra and used to find matches in a sequence database.[21−24] However, this method to reduce the search space is largely limited to known protein sequences and modifications and can lead to misinterpretation if they occur inside the sequence tag. InsPect[25] (a database filtering method) aims to identify any PTM (i.e., is unrestricted in mass search) but needs the identification of unmodified peptides as a prerequisite. PeaksPTM[19] is a newer method outperforming InsPect and MASCOT in identification of known PTMs but is restricted to all known PTMs and needs identification of unmodified peptides as well. SeMoP[26] is another unrestricted PTM identifier that needs (unmodified) peptide identification first. PTM identification remains a focus of intense research.[27]

Analyzing the similarities of measured spectra has been suggested as an alternative approach for unrestricted PTM identification. Falkner et al.[28] developed a specific dot product similarity score for clustering spectra. However, this overall spectra similarity score might become corrupted by the presence of noise peaks. Tsur et al.[29] developed a spectra alignment approach that is potentially more robust to noise and allows PTM identification within modification masses from −100 to +160 Da. Spectra alignment was later used in PTMFinder,[18] allowing the identification of one unrestricted PTM (±250 Da) per peptide, and in spectral network analysis (SNA[30]), yielding unrestricted PTM identification, but only for precursor ion charges 1 and 2.

Here we present a novel approach called mass spectrometry-peak shift analysis (MS-PSA) for PTM identification that is not restricted to any modification masses or ion charges and is complementary to current standard methods such as MASCOT and SEQUEST. MS-PSA is not based on database search or peptide identification but allows the identification of spectra that are related, for instance, due to one or more PTMs. MS-PSA is inspired by frequent pattern mining.[31] Such biclustering approaches[32,33] avoid the noise dependence of whole-spectra clustering. MS-PSA is straightforward and fast, essentially based on frequent peak shifts that match the precursor ion mass difference. It primarily provides potential PTM masses that can be confirmed by existing standard methods. MS-PSA improves the speed and sensitivity of peptide identifications and allows unrestricted PTM identification. We present the analysis of a simple and a complex proteomic data set to illustrate the use of MS-PSA.

MS-PSA identifies related spectra without reference to either a sequence database or a modification database. However, to aid interpretation, MS-PSA permits the adoption of peptide annotations from a database search of the user's choice. The parameters used by MS-PSA are readily adjustable for analysis of data from low (e.g., ion traps) or high mass accuracy instruments at both precursor and fragment levels (to permit the use of hybrid instruments such as the Orbitrap family). In particular, if the approximate location of a modification on a specific protein (i.e., the peptide) is suspected, MS-PSA can rapidly guide the user to informative spectra. We note that the application of the general approach used by MS-PSA is not limited to peptides and could be used for the analysis of any set of fragmentation patterns.

METHODS

MS-PSA takes as input MS/MS spectra in MASCOT generic format. If some spectra have already been identified (e.g., by SEQUEST or MASCOT), a corresponding list containing spectrum titles, peptide annotations, and modifications can be provided, but MS-PSA itself works without any annotation information. If annotation is available it assists with interpretation of results. MS-PSA outputs a ranked list of related spectra and a corresponding graph visualization.

MS-PSA Workflow

The main parameters are discussed, and default settings are given in brackets. Figure S1 in the Supporting Information shows the default setting of all parameters.

(1) Spectra preprocessing: Remove spectra with very low numbers of peaks (likely noise, default: spectra with <10 peaks

dx.doi.org/10.1021/pr500347e | J. Proteome Res. 2014, 13, 4002−4011

(a)

	A	B	C	D	E	F	G	H	I
1	MS-PSA ID	specs1	specs2	peakshift matching mass differ	# such peaks	#sp/length small cs	# common peaks	#cp/length small cs	quality sum
49	48	{"1535.--148-rm20_SFhi1.1	{"1687.5--258-	152	14	0.341	9	0.22	0.561
50	49	{"1536.5--211-rm20_SFhi1	{"1687.5--258-	152	18	0.581	6	0.194	0.774
51	50	{"1535.5--155-rm20_SFhi1	{"1752.--768",	217	5	0.132	16	0.421	0.553
52	51	{"1752.5--784", "1752.5--8	{"1535.--148-r	217	5	0.122	13	0.317	0.439
53	52	{"1557.5--165-rm20_SFhi1	{"1774.5--786'	217	5	0.132	10	0.263	0.395

(b)

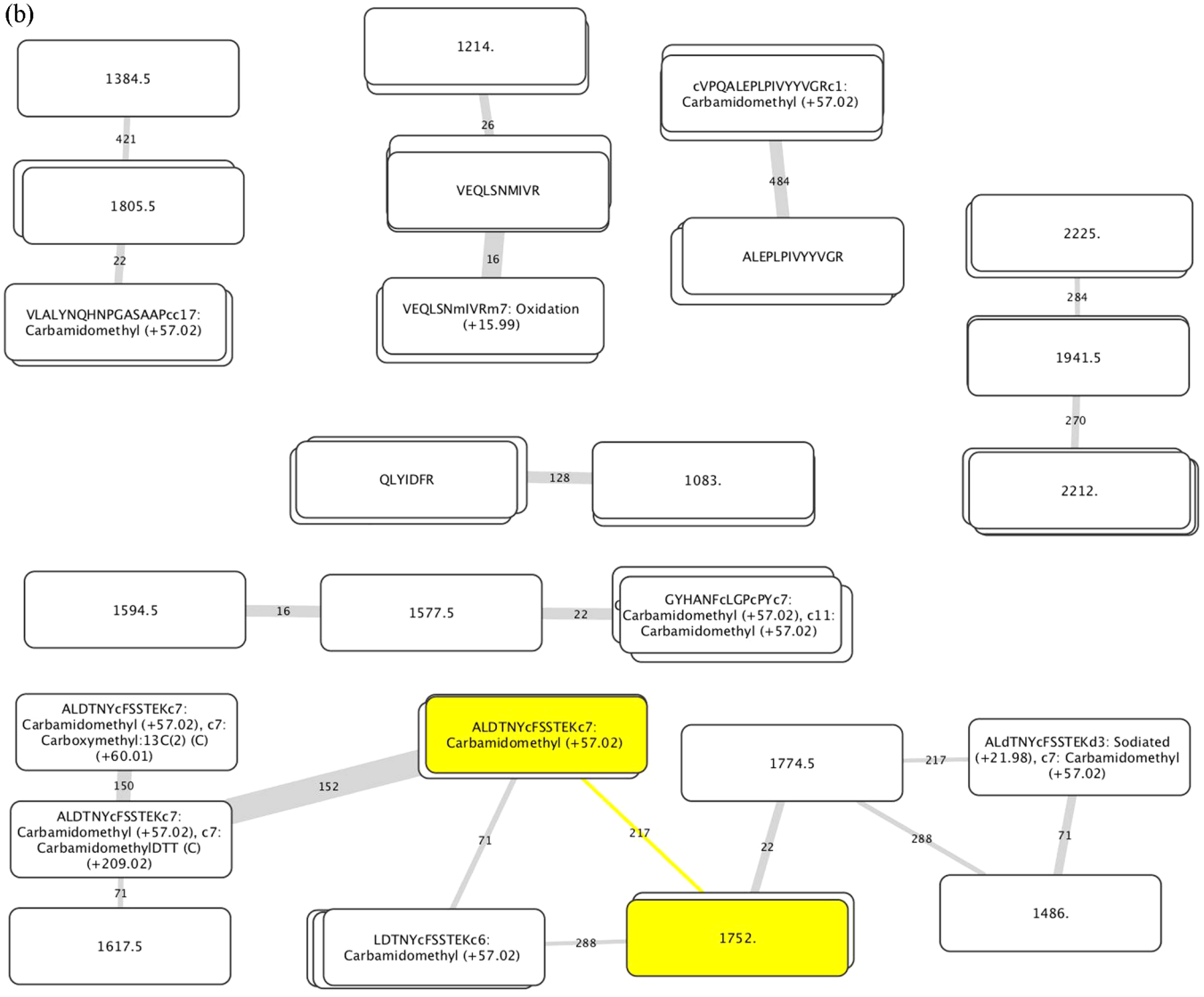


Figure 1. (a) Example of the main result table generated by MS-PSA. Column B contains the most representative spectra of spectra group 1 (typically two spectra if *csf* = 2). A spectrum is represented by its name, containing the precursor mass (first number) and the spectrum number in the mgf file (second number). Column C contains the names of the spectra representing group 2. Groups of columns A and B are related by the mass difference in column D, with the number of such peaks in column E. Columns F and H contain parameters measuring the quality of the suggested spectrum relation. Column I contains the corresponding sum. For the sake of clarity, the following columns are omitted in the Figure: Column M contains the shifted peaks (peak shift list (cf. Methods) including mass and frequency), and column N contains the corresponding maximum frequency. Column O indicates if the most frequent shift matches the precursor mass difference (difference between columns E and N = 0). Columns J–L provide information for merged peaks. The final columns provide the titles of MS2 spectra for groups 1 and 2. (b) MS-PSA network graph for the small low-resolution example, visualized using Cytoscape. Nodes represent groups from the first two columns of the main results table and are labeled with the peptide sequences propagated from Mascot or with the precursor mass if not identified by Mascot. Edges joining related spectra groups are labeled with the corresponding mass shift. The width of the edge indicates confidence using column I of the main results table (panel a). Overlaid nodes indicate multiple groups with the same mass difference, a feature created by the mass tolerance specified in the specific parameters used.

dx.doi.org/10.1021/pr500347e | *J. Proteome Res.* 2014, 13, 4002–4011

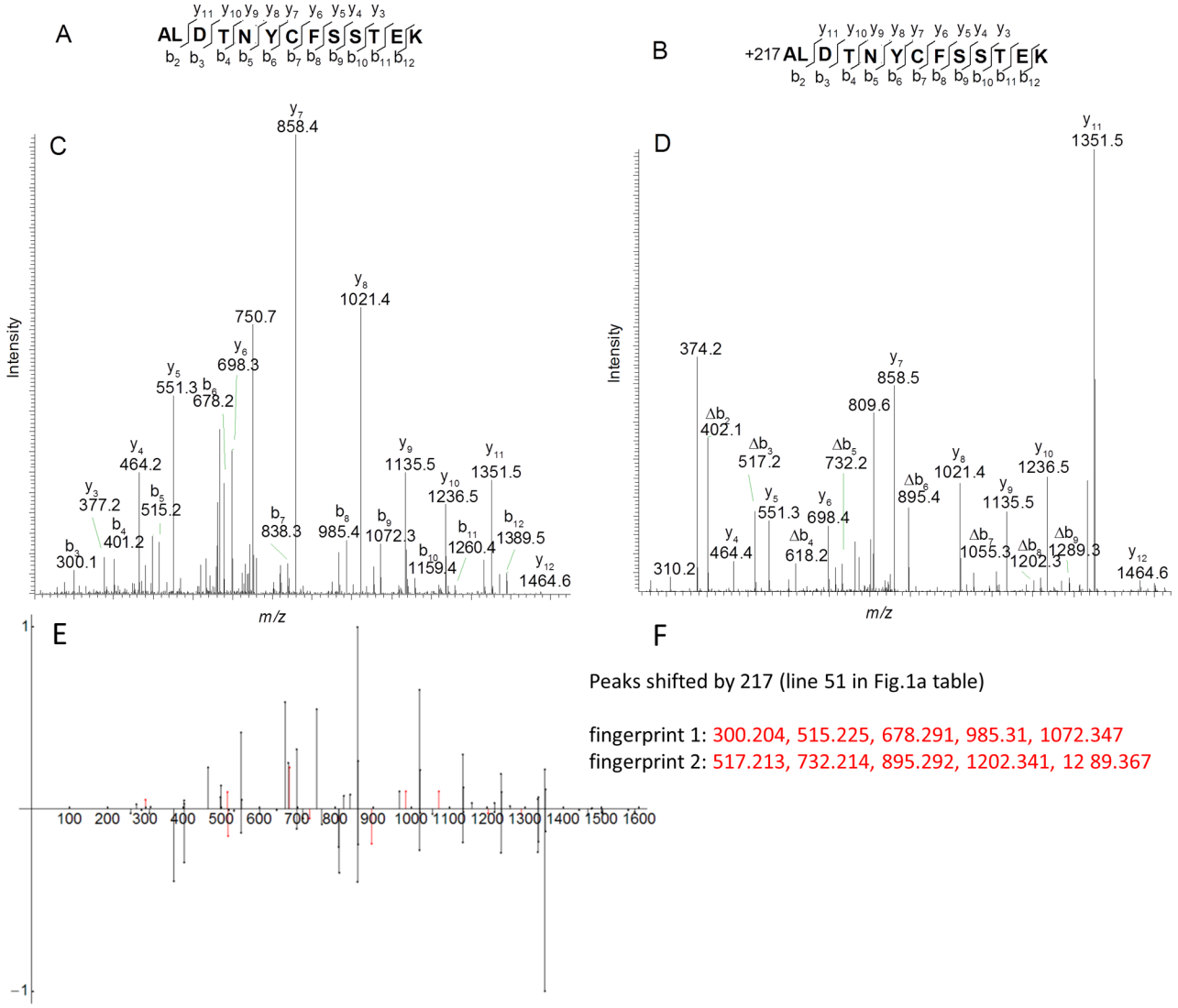


Figure 2. (A–D) Previously published fragmentation and the corresponding annotated spectra of the peptide ALDTNYCFSSTEK with and without the +217 modification. The peaks marked with a triangle are the ones that shift with the modification. The shifted peaks identified by MS-PSA are shown as mirrored spectra (E) and the fingerprint sets with shifted peaks are marked in red (F).

removed), round precursor ion masses, and MS/MS peak positions according to the resolution of the MS machine used (default: precursor ion mass to 0.5 Da and fragment mass to 0.5 m/z). Noise filtering: Keep only the $t1$ most intense peaks per fragment-mass-difference $t2$ (default: $t1$ = 5, $t2$ = 100 m/z).

(2) Group spectra into groups with the same precursor mass.

(3) Analyze spectra of a given group for additional group splitting. Two different peptides might have the same precursor mass but are still distinguishable by different peak patterns. In general, significantly different peak patterns can be found by appropriate (bi)clustering. Dot products have been used for clustering of MS data.[5,27,34] Biclustering methods such as DASS[32,33] group spectra with similar subsets of peaks together, such that noise peaks tend to be less distracting. MS-PSA has implemented a local optimization clustering method using specific significance tests. It can easily be adjusted by several parameters. Clustering increases the number of spectra groups to consider. A graph output of split spectra groups helps manual revision of the clustering.

(4) Remove spectra that might not belong to the group considered. For each spectra group, the *nmfp* most frequent peaks are identified. Only spectra containing at least *mofp* ≤ *nmfp* such peaks are considered for further analyses, and only groups containing at least *mnspg* spectra are further analyzed.

(5) Identification of one "fingerprint" per spectra group: A "fingerprint" is the largest closed set (cs) with frequency ≥ *csf* (default: *csf* = 2). A cs is a subset of elements (in our context peak positions) that occurs more frequently in host sets (all spectra in a group) than any of its supersets.[32,33] This represents an information-rich (many peaks) but still reliable (found in at least *csf* spectra) set of peaks representing the group. We implemented *csf* ϵ {1,1.5,2}, with the following specificities: If *csf* = 1 and different cs have the same maximum size, then the cs corresponding to peaks with higher intensity is chosen; if *csf* = 2 and different cs have the same maximum size, then the union of the cs is taken if this is supported by at least one spectrum. If *csf* = 1.5, first the most frequent peaks in the whole spectra group are calculated. All peaks with frequency ≥ *peakfreq* are considered. The fingerprint is the largest intersection of these peaks with any spectrum in the group. If *csf* = 1.5 and *peakfreq* = 2, the corresponding fingerprint is typically larger than for *csf* = 2 and smaller than for *csf* = 1; the

dx.doi.org/10.1021/pr500347e | J. Proteome Res. 2014, 13, 4002–4011

confidence of the peaks is in between, too. If different *csf* = 1.5 fingerprints have the same maximum size, the one corresponding to peaks with higher intensity is chosen.

(6) The core of MS-PSA: For all pairs of spectra groups potential peak shifts are identified. If the fingerprints of two groups have at least *cp* peaks in common and the number of common peaks divided by the size of the smaller fingerprint is at least *qual2*, the following is done: All positive differences in fragment peak positions from the fingerprint of the lighter precursor ion to the fingerprint of the heavier one are calculated. (Positive difference means a peak position of the heavier precursor ion must represent a higher mass than a peak position of the lighter one to be considered for the difference.) All differences are collected and sorted by frequency. Two spectra groups (peptides) are considered to be potentially related (by one modification, e.g., PTM) if one of the peak differences with a frequency of at least *pwss* (peaks with same shift) and a quality of at least *qual1* (*qual1* = *pwss*/size of smaller fingerprint) is deviating less than *maxdev* from the corresponding precursor ion mass differences (default: *cp* = 6, *pwss* = 5, *qual1* = 0.1, *qual2* = 0.15, *maxdev* = 1 (Da, should be adjusted according to MS instrument accuracy)). The list containing the *pwss* frequent peak differences is called a peak-shift list. It can also be used for the identification of more than one potential PTM: If the sum of two frequent peak shifts is (nearly) identical to the precursor ion mass difference it could be due to two modifications. Identification of higher numbers of modifications is technically possible but not considered so far. The parameter setting the maximum number of peptide modifications to consider is *nummods* (default: 2). Note that the core of MS-PSA does not consider peak intensities but only matches peak positions.

MS-PSA outputs two lists of related spectra groups: "Few" contains only the spectra containing the fingerprints of the two spectra groups (i.e., the most reliable spectra for the considered spectrum relation), and "more" contains all spectra of the two groups. "Few" is best used when one is hunting for new modifications; this table identifies the most representative spectra assigned to the groups. The output "more" is appropriate for spectrum counting or when checking spectra within peak areas assigned to peptide identifications. Sometimes peak differences deviate only by small amounts from each other, so MS-PSA also provides merged peak shifts. Two peak shifts are merged if they deviate by ≤ 1. Note that a manageable number of spectra groups and peak numbers can always be obtained by appropriate adjustment of the parameters previously discussed. Accordingly, MS-PSA can handle MS/MS data of any size and mass accuracy.

The main result of MS-PSA analysis is the "few" list of related spectra (Figure 1a) that can be ranked by the size of the corresponding precursor ion mass difference (i.e., mass of the potential modification) or by different quality parameters, such as the confidence scores *qual1/2* and the corresponding sum. In addition, MS-PSA provides a graphical output of the spectral relations (Supporting Information, Figure S2 and example data MS-PSA_example folder). Note that the main results table ("few") contains all information to visualize the spectra relations graph using Cytoscape (Figure 1b); to simplify node annotation, we provide an additional annotation table. MS-PSA also provides spectra visualization, highlighting fingerprint peaks, and peaks shifted due to the potential PTM to enable easier interpretation of results (Figure 2E).

The complete Mathematica code is freely available as MS-PSA.nb (contained in the Supporting Information MS-PSA_example folder). If a user of MS-PSA.nb puts the two files spectra.mgf and Mascot-annotations.txt (contained in the Supporting Information) into the folder "MS-PSA_example" and this folder is just in C, then the user can run MS-PSA.nb without any modification of the code and gets exactly all results that are in the folder "2014_3_10_11_57" (contained in the Supporting Information MS-PSA_example folder).

Details of Test Set Data

We demonstrate the application of MS-PSA using two MS/MS example data sets (based on collision-induced dissociation fragmentation):

(1) A single modified protein digest from Traka et al.[35] containing 2211 spectra collected on a LTQ (Thermo Scientific): Tandem mass spectra were extracted by Xcalibur (Thermo Scientific). Charge-state deconvolution and deisotoping were not performed. All MS/MS samples were analyzed using MASCOT (Matrix Science, London, U.K.) to search the SwissProt/TREMBL database (selected for *Homo sapiens*, 108 143 entries) specifying the digestion enzyme trypsin and with a fragment ion mass tolerance of 0.80 Da and a precursor ion tolerance of 1.5 Da. Iodoacetamide derivative of cysteine was specified as a fixed modification and error tolerant modifications permitted. A second Mascot search included the custom sulphoraphane (SF) modification of +217 as a variable modification. Mascot results are compiled in Scaffold (version Scaffold_4.3.0, Proteome Software, Portland, OR) and presented in the Supporting Information. A free Scaffold viewer is available from Proteome Software. However, note that no additional software like Mascot or a Scaffold viewer is required to run MS-PSA.

(2) A standard protein mix from Klimek et al.[36] (specifically "mix7 Orbitrap"): We downloaded the 10 raw files and made a single .mgf file using Progenesis software (Non Linear Dynamics), provided in the Supporting Information in MASCOT generic format. The data were generated on a high mass accuracy instrument (Orbitrap); therefore, a precursor mass tolerance of 0.01 was allowed for the precursor ions. The standard sample contained 18 proteins and a number of known contaminants. MS/MS spectra were searched using Swiss Prot/TREMBL (17 035 495 entries) with the digestion enzyme trypsin and a fragment ion mass tolerance of 0.80 Da and a precursor ion tolerance of 10.0 ppm. Iodoacetamide derivative of cysteine was specified in Mascot as a fixed modification. Oxidation of methionine was specified in Mascot as a variable modification. An error tolerant search was performed with these same parameters, and results are available in a scaffold file in the Supporting Information.

Networks were visualized using Cytoscape.[37]

RESULTS

Analysis of a Small Low-Resolution Data Set

To test the ability of MS-PSA to assist in the identification of unspecified modifications, we reanalyzed a previously published data set.[35] It contains a novel modification that was extremely difficult to identify for the initial publication. The sample consists of purified TGFβ treated with an isothiocyanate, sulforaphane. During alkylation for tryptic digestion, an unexpected secondary chemical reaction altered the expected mass shift from +177 to +217. The Traka data set contains 2211 MS/MS spectra, 529 of which were assigned to 14 unique

peptides with an error tolerant MASCOT search (counting mis-cleavages as unique hits and including the +217 modification).

Using the default MS-PSA parameters given in Figure S1 in the Supporting Information, analysis with MS-PSA gave 83 spectral relations containing 111 unique spectra in the main results ("few") table and 699 in the "more" table (Supporting Information, Table S1 example data MS-PSA_example folder). We propagated the MASCOT ETS peptide annotations for identified spectra on to these relations. The majority of the mass shifts represent common modifications: oxidation (+16), salt adducts (+22), or formylation (+26). Among the remaining mass shifts is the +217 that Traka et al.[35] assigned to the further modification of sulforaphane with iodoacetamide. This peptide is identified as ALDTNYCFSSTEK by the propagated MASCOT annotation. The analysis with MS-PSA found the same spectra as Traka et al. and additionally found a second group that had a salt adduct (Figure 1). When +217 was added to the MASCOT search parameters, these spectra were also identified by MASCOT.

To facilitate the interpretation of the spectral groups, we can visualize the main results table as a network (e.g., using Cytoscape), where each spectral group can be represented as a node and spectra relations (rows in Figure 1a) as edges. Edges can be labeled with mass shifts (as in Figure 1b) or other values from the main results table. Each row of the results table creates an edge between two nodes. For convenience, MS-PSA provides graph pictures for the main results table and for all spectra of related groups (shown in the example data set and Figure S2 in the Supporting Information). Additionally, MS-PSA generates a table to facilitate the labeling of nodes in Cytoscape with MASCOT annotations or the precursor mass (if no MASCOT annotation available). Thus, the 83 spectra relations could be reduced to seven networks (Figure 1b). Each node corresponds to one spectra group. Note that the same peptide might correspond to several spectra groups with slightly different parent ion masses. We grouped parent ion masses to 0.5 Da in this analysis; therefore, multiple overlaid nodes can be seen with spectra outside this mass tolerance (due to C13 isotopes or other factors). Such redundant nodes are shown as stacks in Figure 1b. The graphs in Figure 1b clearly link related spectra, and the mass shifts associated with oxidation (16) and sodium adducts (22) are apparent. Miscleavage and large truncations are also obvious, such as the peptides CVPQALEPIVYYVGR and its nontryptic cleavage product ALEPIVYYVGR, linked by 484.

If the peptide of interest is known, the graphs can quickly guide users to related spectra. We illustrate the use of the graph in Figure 1b, where we highlight the spectral relations associated with ALDTNYCFSSTEK, the peptide modified by +217. This graph clearly identifies a loss of the N-terminal alanine residue (−71), the +217, and its salt adduct (+22). Thus, this feature can assist when multiple modifications are present by narrowing the likely sums that could account for the mass shift of more than one modification.

To assist manual verification that the two spectral groups are indeed related, MS-PSA returns the fragment masses that were used as closed sets in each group as well as the corresponding shifted peaks. This information is returned to the user in text files, where the line number is the same as the ID in the main results table. MS-PSA also provides a visualization of all spectra contained in the main results table (provided in a folder called "spectra-pics", cf. Supporting Information, MS-PSA_example folder) as well as the fingerprint peaks of the two related spectra groups in one figure (in a folder called "peak-shift-pics" in the example data set and illustrated in Figure 2E). When combined with an annotated spectrum, this makes a powerful tool to manually verify the MS-PSA assigned relationship. We illustrate this feature by comparing the annotated spectra for the peptide ALDTNYCFSSTEK with and without the +217 modification (published in Traka et al.[35]) with the peak shifts identified by MS-PSA (Figure 2). Because the modification to ALDTNYCFSSTEK is at the N-terminus of the peptide, all of the b ions shift, while the y ions are common. All of the peaks that MS-PSA identifies as shifted by 217 are b ions; the y ion series remains unchanged.

Analysis of a Large High-Resolution Data Set

To test the performance of MS-PSA to identify unspecified modifications in a more complex mixture, we reanalyzed a standard data set from Klimek et al.[36] We performed two types of MASCOT search (simple specified modifications and error tolerant). These two searches enabled us to ask two questions: (A) Could MS-PSA identify the same modified−unmodified peptide pairs as MASCOT? (B) How many additional spectra could be assigned to related groups through identification of additional modifications? It should be noted that Data set 2 was acquired with higher mass accuracy of the precursor ion, and this was accommodated in MS-PSA by narrowing the precursor mass tolerance in the parameters. The mass accuracy of the fragment spectra remains the same as in the previous data set.

Data set 2 contains 35 014 MS/MS spectra, including many from precursor ions with charge states greater than four. A simple MASCOT search (permitting only variable oxidation of methionine residues and fixed carbamidomethyl of cysteine residues as modifications) assigned peptide identifications to 1889 spectra. Repeating the MASCOT search with error-tolerant parameters assigned identifications to a further 1040 spectra. Altogether, 2929 spectra were assigned to 427 peptides by MASCOT (Table S2 in the Supporting Information; MASCOT parameters are specified in the Methods section).

To test the performance of MS-PSA, we created a reference set of related peptides from these two MASCOT searches. Peptides were defined as related if (i) the same peptide was present in the MASCOT results differing by modification with a mass of greater than 15 Da or if (ii) peptide sequences were related by mis-cleavage or nontryptic cleavage at either end of a peptide sequence only. This definition of peptide relations excluded internal overlap of sequences or modification and miscleavage with no intermediary forms. We call this subset "PAIRED" and use it to answer question A: to determine if MS-PSA could find the same mass shifts as indicated by MASCOT results. The PAIRED set contains 310 peptide relations that can be linked into 99 graphs of related peptides in a network view of the relations (Table S3 in the Supporting Information). This set covered 2102 spectra, the majority of all spectra (2929) assigned to peptide sequences by MASCOT.

Analysis of the full 35k data set with MS-PSA created 705 spectra relations (parameters used are given in the Supporting Information), covering 705 unique spectra in the "few" and 4814 spectra in the "more" table. Several of these 705 spectra relations correspond to the same peptide relation, so overall MS-PSA identified 602 different peptide relations that form 153 graphs of related spectra in a network view (Table S4 in the Supporting Information). The overlap between the 310 MASCOT relations and the 602 MS-PSA relations is 56. The

dx.doi.org/10.1021/pr500347e | J. Proteome Res. 2014, 13, 4002−4011

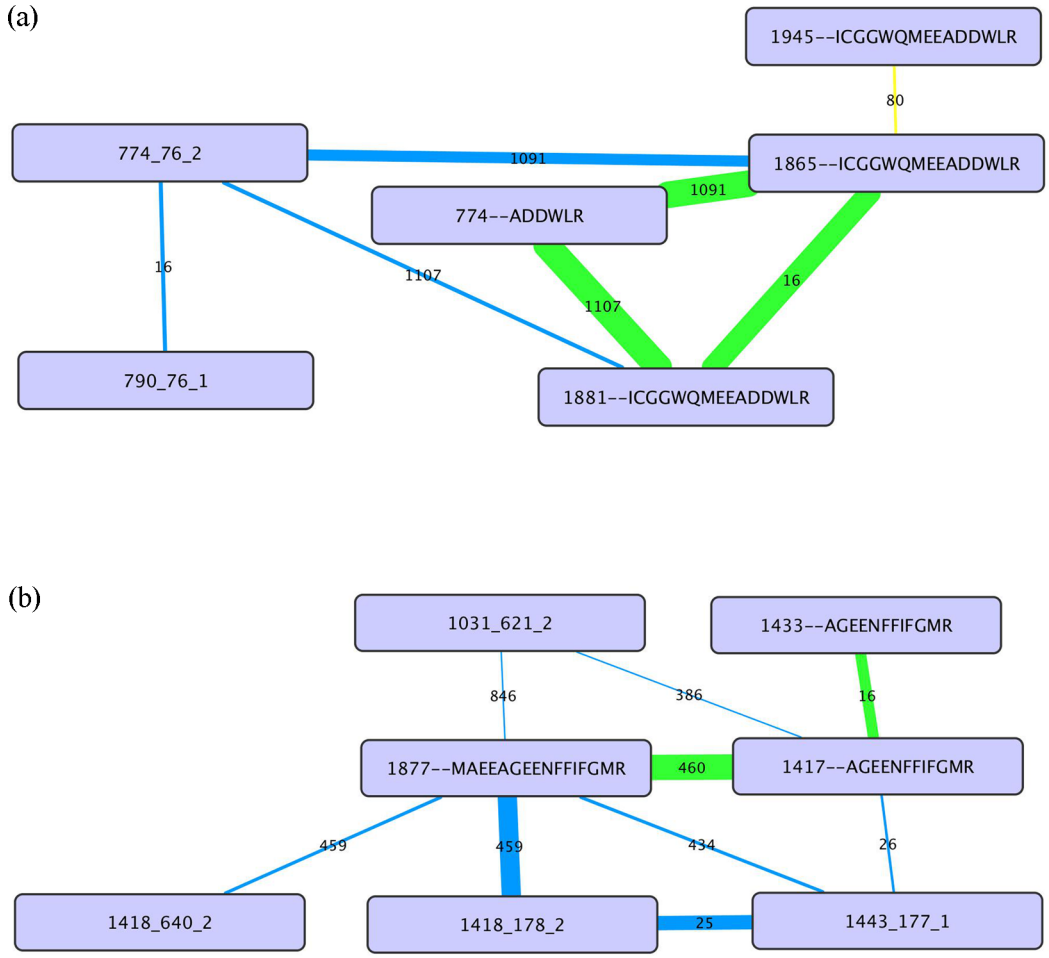


Figure 3. Examples from the large high-resolution data set. Nodes represent peptides/spectra (labeled mass_peptide sequence, or when no Mascot annotation is available mass_MS-PSA line_MS-PSA group). Edges represent relations between peptides/spectra and are labeled with the mass shift identified. Green edges indicate that both MASCOT and MS-PSA identified the same relation; yellow is Mascot alone and blue is MS-PSA alone. The thickness of the blue and green lines represents the MS-PSA confidence score.

statistical significance of this overlap is beyond the working precision of standard calculation software, $p < 10^{-300}$ (background of $\sim(30k)^2$ relations). Note that relaxing, for instance, the quality parameters of MS-PSA leads to identification of many more potential spectra relations, finally covering all 310 PAIRED relations. As with the first test data set, commonly identified modifications included oxidation, sodium adducts, and formylation and miscleavages.

There are 36 additional MS-PSA peptide relations to which MASCOT had also assigned peptide identifications in both groups, but these peptide pairs had not been included in the PAIRED test set (according to the definitions given above). In all of these cases MASCOT assigned the same or very similar peptides to both groups. Examination of these showed that amino acid substitution or variation in the assigned peptides was common or that the sequence overlap was internal so these did not fit our selection criteria. However, this illustrates the strength of MS-PSA in reliably grouping spectra that should be assigned to the same peptide or modifications thereof.

Combing the networks of peptide relations in the MASCOT PAIRED results and the MS-PSA results gave a total of 181 related peptide graphs. 38 peptide graphs have at least one relationship in both the PAIRED test set and the set identified by MS-PSA. The 56 overlapping relations are shown in green. (Figure S3 in the Supporting Information).

While the peptide relation graphs can appear to be very complicated, it should be noted that these represent structured groups of thousands of spectra. To illustrate the utility of network relations, we consider some specific examples in Figure 3. The peptide ICGGWQMEEADDWLR has an oxidized form, an oxidized with formylkynurenine (+31.99) form (giving a mass shift of +80), and a tryptic mis-cleavage (ADDWLR) in the MASCOT results and PAIRED test set. Analysis with MS-PSA identified the relation between the oxidized and miscleaved forms (green lines Figure 3a) but did not assign the +80 relation (yellow line Figure 3A). In addition, MS-PSA identified other groups of spectra for the truncated form and likely oxidation thereof (blue lines Figure 3A). The peptide MAEEAGEENFFIFGMR also has a tryptic mis-cleavage (to AGEENFFIFGMR) that is also observed as an oxidized form. MS-PSA identified these relations and additionally other spectra for the mis-cleaved form and two other groups of spectra that are related to both MAEEAGEENFFIFGMR and AGEENFFIFGMR by mass shifts of 846, 434, and 386 and 26 (respectively, see Figure 3b). These dual assignments by MS-PSA make the relationships highly credible. Moreover, the MS-PSA confidence score, represented by the thickness of blue and green lines, also helps to identify related spectra.

dx.doi.org/10.1021/pr500347e | J. Proteome Res. 2014, 13, 4002−4011

DISCUSSION

According to Bandeira et al.,[30] the MacCoss group[38] was the first to realize the potential of spectral pairs for the identification of PTMs. The unrestricted PTM identification approaches ModifiComb[39] and SNA[30] are based on this idea. However, SNA works only for precursor ion charges 1 and 2, and ModifiComb needs the identification of unmodified peptides as a prerequisite and requires high mass accuracy data. Both methods also provide the modification site. ModifiComb and SNA work with potential b and y ion tags. ModifiComb calculates a number of common cleavage sites present in two peptides as the central parameter to quantify the peptide relatedness. SNA quantifies spectra similarity by an advanced spectra alignment algorithm. Both ModifiComb and SNA work with complete spectra, whereas MS-PSA is based on the representative fingerprints of spectra groups, making it robust to noise. ModifiComb additionally uses retention times as a second dimension, whereas Bandeira et al.[30] argue that "retention time analysis imposes the constraint that both spectra must come from the same sample", so they do not consider it. We agree with Banderia et al. and do not use the retention times of the spectra in MS-PSA analysis. MS-PSA is exclusively built on the spectral pair paradigm. It has no restrictions to charge states or PTM masses nor to identification of unmodified peptides. MS-PSA uses a completely different approach to ModifiComb and SNA; it provides a complementary method for PTM identification. Potential PTMs that score highly in MS-PSA can additionally be analyzed by existing technology such as MASCOT or SEQUEST by adding the modification mass identified by MS-PSA.

Retention time information is often used to reduce the complexity of the analysis algorithm. MS-PSA is working efficiently without using retention time; data of any reasonable size can be analyzed. The core of the MS-PSA algorithm (step 6 of the workflow, cf. Methods) analyzes all $n(n-1)/2$ fingerprint spectra pairs of the n spectra groups considered; the corresponding time complexity is $O(n^2)$. The analyzed large data set contains $n \approx 1000$ spectra groups (using the parameters given). The complete MS-PSA analysis takes ~1 h on a standard PC.

MS-PSA was developed as a tool complementary to existing technology to facilitate the identification of related spectra, likely representing unspecified modifications in large data sets. Of course, our new approach has room for further improvement. Working with only the measured spectra and no databases MS-PSA needs at least two related spectra for identification of potential PTMs, for instance, an unmodified peptide and the same with one or more modifications or the same peptide with one and two modifications. Providing theoretical spectra would enable the identification of potential PTMs in cases where just one modified form is present in the spectral data set. MS-PSA provides several quality parameters for the relation between spectra, for instance, the number of shifted peaks matching the precursor ion mass difference and the relative number (normalized by the number of peaks of the smaller fingerprint). False discovery rates (FDRs) can be calculated by exploiting a decoy (i.e., appropriately shuffled) database that should contain more spectra to enable conservative estimates; FDR = m/n, m = no. identified PTMs according to the given threshold in the decoy database, and n = no. in measured data.[19,20] All parameters of MS-PSA need to be set by hand in advance; a typical working default set is provided (cf. Methods, Figure S1 in the Supporting Information). However, at least some parameters could also be set automatically, depending on the data available at each step (e.g., *cp* and *pwss*, cf. Methods). The modification site of the potential PTM could be determined by the spectral products analysis method described by Bandeira et al.[30] For samples with reproducible liquid chromatography conditions, the addition of retention time could improve results. Interestingly, a method DeltAMT was suggested for the identification of peptide modifications, which works exclusively with precursor masses and retention times. (A corresponding bivariate Gaussian mixture model discriminates modification-related spectral pairs from random ones.)[40,41] DeltAMT is fast and complementary to MS-PSA, so a combination of these approaches could provide a highly efficient method.

Finally, the MS-PSA approach may be further developed for peptide quantification. There is considerable scope for improvement of spectrum counting methods by identifying all spectra associated with a peptide. MS-PSA may also contribute to precursor area methods as an unidentified modified version of the measured peptide will subtract from the intensity of the unmodified form.

ASSOCIATED CONTENT

Supporting Information

Parameters used for the analysis of the two test sets. Figure S1: Default parameters of MS-PSA. Figure S2: Network output from MS-PSA. Figure S3: Overview of the overlapping PAIRED and MASCOT peptide graphs identified in the large test set. Two test data sets from Traka et al.[35] and Klimek et al.[36] are provided in mascot generic format (.mgf) files. MS-PSA.nb contains the complete Mathematica code. This runs without any modification needed (cf. Methods) to analyze the small test set from Traka et al.[35] (spectra.mgf and Mascot-annotations.txt). The large test data from Klimek et al.[36] is contained in a .zip file. Supplemental tables S1–S4. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Thomas.Wilhelm@ifr.ac.uk. Tel: +44 1603 255313. Fax: ~507723.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

T.W. gratefully acknowledges the support of the Biotechnology and Biological Sciences Research Council (BBSRC). This research was funded by the BBSRC Institute Strategic Programme Grant BB/J004529/1. A.M.E.J. gratefully acknowledges support from the Gatsby Foundation and the School of Life Sciences, University of Warwick.

REFERENCES

(1) 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. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.

(2) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence data-

bases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(3) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol. Cell Proteomics* **2008**, *7* (5), 962−970.
(4) Tharakan, R.; Martens, L.; Van Eyk, J. E.; Graham, D. R. OMSSAGUI: An open-source user interface component to configure and run the OMSSA search engine. *Proteomics* **2008**, *8* (12), 2376−2378.
(5) Ye, D.; Fu, Y.; Sun, R. X.; Wang, H. P.; Yuan, Z. F.; Chi, H.; He, S. M. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **2010**, *26* (12), i399−i406.
(6) Yates, J. R., 3rd; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* **1998**, *70* (17), 3557−3565.
(7) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **2008**, *5* (10), 873−875.
(8) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **2006**, *78* (16), 5678−5684.
(9) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.* **2006**, *5*, 1843−1849.
(10) Hummel, J.; Niemann, M.; Wienkoop, S.; Schulze, W.; Steinhauser, D.; Selbig, J.; Walther, D.; Weckwerth, W. ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinf.* **2007**, *8*, 216.
(11) Ma, B.; Johnson, R. De novo sequencing and homology searching. *Mol. Cell.Proteomics* **2012**, *11* (2), O111 014902.
(12) Chi, H.; Sun, R. X.; Yang, B.; Song, C. Q.; Wang, L. H.; Liu, C.; Fu, Y.; Yuan, Z. F.; Wang, H. P.; He, S. M.; Dong, M. Q. pNovo: de novo peptide sequencing and identification using HCD spectra. *J. Proteome Res.* **2010**, *9* (5), 2713−27.
(13) Jeong, K.; Kim, S.; Bandeira, N.; Pevzner, P. A. Gapped spectral dictionaries and their applications for database searches of tandem mass spectra. *Mol. Cell Proteomics* **2011**, *10* (6), M110 002220.
(14) Bern, M.; Cai, Y.; Goldberg, D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **2007**, *79* (4), 1393−1400.
(15) Khoury, G. A.; Baliban, R.C.; Floudas, C.A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep* **2011**, *1*, 90.
(16) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4* (6), 1534−1536.
(17) Farriol-Mathis, N.; Garavelli, J. S.; Boeckmann, B.; Duvaud, S.; Gasteiger, E.; Gateau, A.; Veuthey, A. L.; Bairoch, A. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* **2004**, *4* (6), 1537−1550.
(18) Tanner, S.; Payne, S. H.; Dasari, S.; Shen, Z.; Wilmarth, P. A.; David, L. L.; Loomis, W. F.; Briggs, S. P.; Bafna, V. Accurate annotation of peptide modifications through unrestrictive database search. *J. Proteome Res.* **2008**, *7* (1), 170−181.
(19) Han, X.; He, L.; Xin, L.; Shan, B.; Ma, B. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *J. Proteome Res.* **2011**, *10* (7), 2930−2936.
(20) Granholm, V.; Kall, L. Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* **2011**, *11* (6), 1086−1093.
(21) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390−4399.
(22) Tabb, D. L.; Saraf, A.; Yates, J. R., 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **2003**, *75* (23), 6415−6421.
(23) Searle, B. C.; Dasari, S.; Turner, M.; Reddy, A. P.; Choi, D.; Wilmarth, P. A.; McCormack, A. L.; David, L. L.; Nagalla, S. R. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal. Chem.* **2004**, *76* (8), 2220−2230.
(24) Han, Y.; Ma, B.; Zhang, K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J. Bioinform. Comput. Biol.* **2005**, *3* (3), 697−716.
(25) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626−4639.
(26) Baumgartner, C.; Rejtar, T.; Kullolli, M.; Akella, L. M.; Karger, B. L. SeMoP: a new computational strategy for the unrestricted search for modified peptides using LC-MS/MS data. *J. Proteome Res.* **2008**, *7* (9), 4199−4208.
(27) Hoopmann, M. R.; Moritz, R. L. Current algorithmic solutions for peptide-based proteomics data generation and identification. *Curr. Opin. Biotechnol.* **2013**, *24* (1), 31−38.
(28) Falkner, J. A.; Falkner, J. W.; Yocum, A. K.; Andrews, P. C. A spectral clustering approach to MS/MS identification of post-translational modifications. *J. Proteome Res.* **2008**, *7* (11), 4614−4622.
(29) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **2005**, *23* (12), 1562−1567.
(30) Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (15), 6140−6145.
(31) Agrawal, R., Srikant, R. *Fast Algorithms for Mining Association Rules*, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994; IBM Almaden Research Center: San Jose, CA, 1994.
(32) Hollunder, J.; Friedel, M.; Beyer, A.; Workman, C. T.; Wilhelm, T. DASS: efficient discovery and p-value calculation of substructures in unordered data. *Bioinformatics* **2007**, *23* (1), 77−83.
(33) Hollunder, J.; Friedel, M.; Kuiper, M.; Wilhelm, T. DASS-GUI: a user interface for identification and analysis of significant patterns in non-sequential data. *Bioinformatics* **2010**, *26* (7), 987−989.
(34) Wang, J.; Perez-Santiago, J.; Katz, J. E.; Mallick, P.; Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* **2010**, *9* (7), 1476−1485.
(35) Traka, M.; Gasper, A. V.; Melchini, A.; Bacon, J. R.; Needs, P. W.; Frost, V.; Chantry, A.; Jones, A. M.; Ortori, C. A.; Barrett, D. A.; Ball, R. Y.; Mills, R. D.; Mithen, R. F. Broccoli consumption interacts with GSTM1 to perturb oncogenic signalling pathways in the prostate. *PLoS One* **2008**, *3* (7), e2568.
(36) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J. Proteome Res.* **2008**, *7* (1), 96−103.
(37) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13* (11), 2498−2504.
(38) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., 3rd Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (12), 7900−7905.
(39) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell Proteomics* **2006**, *5* (5), 935−948.
(40) Fu, Y.; Jia, W.; Lu, Z.; Wang, H.; Yuan, Z.; Chi, H.; Li, Y.; Xiu, L.; Wang, W.; Liu, C.; Wang, L.; Sun, R.; Gao, W.; Qian, X.; He, S.-M. Efficient discovery of abundant post-translational modifications and spectral pairs using peptids mass and retention time differences. *BMC Bioinf.* **2009**, *10* (SupplI), S50.

(41) Fu, Y.; Xiu, L.-Y.; Jia, W.; Ye, D.; Sun, R.-X.; Qian, X.-H.; He, S.-M. DeltAMT: A statistical algorithm for fast detection of protein modifications from LC-MS/MS data. *Mol. Cell Proteomics* **2011**, *10*, M110.000455.