Journal of proteome research

Article

pubs.acs.org/jpr

A Turn-Key Approach for Large-Scale Identification of Complex Posttranslational Modifications

Jian Wang,[†] Veronica G. Anania,[‡] Jeff Knott,[§] John Rush,[§] Jennie R. Lill,[‡] Philip E. Bourne,[∥] and Nuno Bandeira*,[∥,⊥,¶]

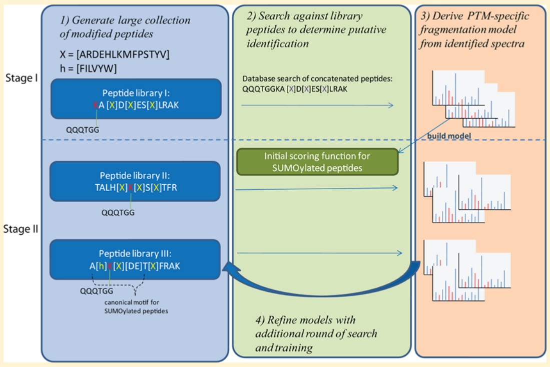
[†]Bioinformatics Program, [∥]Skaggs School of Pharmacy and Pharmaceutical Sciences, [⊥]Center for Computational Mass Spectrometry, and [¶]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093, United States

[‡]Protein Chemistry Department, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, United States

[§]Cell Signaling Technologies, Danvers, Massachusetts 01923, United States

S *Supporting Information*

ABSTRACT: The conjugation of complex post-translational modifications (PTMs) such as glycosylation and Small Ubiquitin-like Modification (SUMOylation) to a substrate protein can substantially change the resulting peptide fragmentation pattern compared to its unmodified counterpart, making current database search methods inappropriate for the identification of tandem mass (MS/MS) spectra from such modified peptides. Traditionally it has been difficult to develop new algorithms to identify these atypical peptides because of the lack of a large set of annotated spectra from which to learn the altered fragmentation pattern. Using SUMOylation as an example, we propose a novel approach to generate large MS/MS training data from modified peptides and derive an algorithm that learns properties of PTM-specific fragmentation from such training data. Benchmark tests on data sets of varying complexity show that our method is 80−300% more sensitive than current state-of-the-art approaches. The core concepts of our method are readily applicable to developing algorithms for the identifications of peptides with other complex PTMs.


1) Generate large collection of modified peptides
X = [ARDEHLKMFPSTYV]
h = [FILVYW]
Stage I
Peptide library I:
KA [X]D[X]ES[X]LRAK
QQQTGG
Peptide library II:
TALH[X]K[X]S[X]TFR
QQQTGG
Stage II
Peptide library III:
A[h]K[X][DE]T[X]FRAK
QQQTGG
canonical motif for SUMOylated peptides
2) Search against library peptides to determine putative identification
Database search of concatenated peptides:
QQQTGGKA [X]D[X]ES[X]LRAK
Initial scoring function for SUMOylated peptides
4) Refine models with additional round of search and training
3) Derive PTM-specific fragmentation model from identified spectra
build model

KEYWORDS: *small ubiquitin-like modification (SUMOylation), posttranslational modification (PTM), combinatorial peptide library, peptide fragmentation patterns, algorithms, database search method, linked peptides*

INTRODUCTION

In recent years the focus in proteomics has shifted from cataloging the "parts list" of gene products inside the cell toward understanding the structural and functional properties of proteins in a systematic and high-throughput manner.[1] A key step toward this goal is the comprehensive characterization of protein post-translational modifications (PTMs). These "decorations" on the protein surface have been shown to play crucial roles in determining a protein's activity state, localization, turnover rate, and interactions with other proteins.[2,3] Recent advances in mass spectrometry (MS) and enrichment protocols that selectively capture peptides with specific PTMs have enabled the detection of many PTMs on a large scale, thus providing scientists with a global view on various PTMs and their interplay at a systems level.[4−8] However, such success has mostly been limited to PTMs that result from the addition of a relatively simple chemical group to one or more amino acid residues in the proteins. Common examples include acetylation, deamidation, phosphorylation, and oxidation. These modifications can be readily identified with tandem mass spectrometry (MS/MS) by considering characteristic shifts in peptide precursor mass as well as in modified fragment ion masses. However, more complex PTMs, such as glycosylation,[9] Small Ubiquitin-like Modification (SUMOylation),[10] PUPylation,[11] and ADPribosylation,[12] present a more difficult problem because the PTMs themselves are large and complex molecules rather than simple chemical moieties, creating unusual "branched" structures for the modified peptides. As a result, these modified peptides display rather different fragmentation pattern than their unmodified counterparts, and thus new experimental and computational methods are needed for the analysis of peptides with complex PTMs.

We propose an automated approach, *Specialize* (Spectra of complex-PTModified peptides identification tool), to derive new algorithms for any type of modified peptide fragmentation using synthetic peptide libraries. While previous studies have mostly used synthetic peptide libraries to evaluate peptide identification algorithms,[13−15] Specialize learns the PTM-specific fragmentation patterns from the peptide libraries and is able to generalize to peptides that are not in the libraries across different biological samples. In addition, the cost of

Received: April 19, 2013
Published: January 20, 2014

generating the peptide libraries is kept very low by using combinatorial peptide synthesis. A training data with thousands of unique peptides was generated using only three peptide synthesis experiments as opposed to synthesizing each peptide separately. To illustrate the concept, we focus on one specific example of complex PTMs (SUMOylation) and use it to demonstrate the feasibility and practicality of our approach. Small Ubiquitin-like Modifiers (SUMO) are small proteins of around 100 amino acids that reversibly attach to substrate proteins to modify their functions. SUMOylation has been shown to be involved in many cellular pathways such as cellular trafficking, cell cycle, and DNA repair and replication.[16] It is also implicated in several neurodegenerative diseases such as Alzheimer's disease and Huntington disease.[17,18] Similar to ubiquitination, SUMOylation is regulated by a series of enzymatic reactions involving SUMO-activating enzymes, conjugating enzymes, and SUMO E3 ligases that covalently attach SUMO to substrate proteins via an iso-peptide bond between the C-terminus of SUMO and a specific lysine residue on the substrate protein. Previously it was thought that SUMOylation occurred within a strict consensus motif [XK(D/E)],[19] but more recently it has been shown that several motifs including an "inverted" consensus motif, a hydrophobic patch motif, and a phosphorylation dictated motif exist as common localizers of SUMOylation.[20] It has also been observed in many cases that SUMOylation can occur on lysine residues not located within any predetermined motif, hence the increased need for unbiased methods to detect SUMOylated lysine residues. Upon enzymatic digestion of SUMOylated proteins, peptides that contain the SUMO conjugation site will be covalently linked to a C-terminal remnant or tag of SUMO, resulting in 'y-shaped' linked peptides (see Figure 1a). Unbiased detection of this type of linked peptide is the most informative as it provides direct evidence that a protein is a SUMO substrate as well as revealing the specific amino acid site of SUMOylation. However these branch-linked peptides present several challenges when being analyzed by MS/MS methods to detect SUMOylated lysine residues.

First, the attachment of a SUMO tag to the substrate peptide inhibits tryptic digestion due to steric hindrance and therefore inaccessibility of the enzyme to the SUMO-conjugated lysine residue, generating peptides with internal lysine that are unusual in unconjugated tryptic peptides. In addition, the SUMO tag resulting from tryptic digestion is relatively large, ranging from 20 to 30 residues depending on the isoforms of SUMO. As a result the SUMO tag tends to dominate the MS/MS spectrum and make it difficult to identify the substrate peptide using current MS/MS methods. In order to address these issues, previous studies have generated SUMO mutants by inserting a lysine or arginine at specific positions along the SUMO C-termini tail so that shorter SUMO C-termini tags (4−6 residues) can be generated upon trypsin digestion.[21−25] On the other hand, alternative enzymes such as chymotrypsin, GluC, and LysC can also be used to generate shorter SUMO tags (4−12 residues) attached to the substrate peptides,[26] making them more suitable for MS/MS analysis.

Even as we circumvent these hurdles and manage to generate SUMOylated peptides with favorable properties to be analyzed by tandem mass spectrometry, it remains a challenge to interpret the resulting MS/MS spectra because almost all mainstream database search algorithms are designed for MS/MS spectra from *linear*, *unlinked* peptides. In contrast, an MS/MS spectrum from a SUMOylated peptide contains a mixture of fragment ions from both the substrate peptide and the SUMO tag. In addition, the linkage of two peptides together results in fragmentation patterns that are different from those of common linear peptides. While there have been several attempts to address these issues, none of them captured the specific fragmentation pattern of SUMOylated peptides due to the lack of appropriate training data.[27,28] Here, we propose a novel experimental and computational hybrid procedure to reliably generate large MS/MS reference data for SUMOylated peptides, which are then used to derive a database search algorithm capturing the PTM-specific fragmentation patterns of SUMOylated peptides.

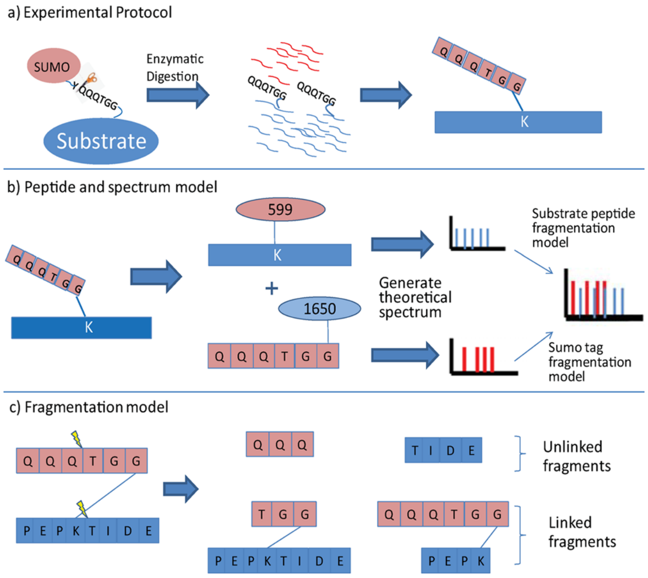


Figure 1. Conceptual model of SUMOylated peptides. (a) Small Ubiquitin-like Modifiers (SUMO) are small proteins that reversibly attach to substrate proteins to regulate their functions. Upon enzymatic digestion of SUMO-conjugated proteins, peptides that contain the SUMO conjugation site in the substrate protein have a SUMO C-terminus remnant (or SUMO tag) covalently attached to the lysine residue, resulting in 'y-shaped' peptides. Here we use QQQTGG as an example of SUMO-tag, which is the last six amino acid residues at the C-terminus of Human SUMO2 protein. (b) SUMOylated peptides are modeled as two peptides: a substrate peptide carrying a modification of mass +599 Da (the mass of the SUMO tag) at the lysine residue and a peptide with sequence QQQTGG carrying a modification at the C-terminus with mass equal to that of substrate peptide (which is assumed to be 1650 Da for illustration purposes). Theoretical fragments from a SUMOylated peptide are represented as two sets of fragment ions, one set from the substrate peptide and another set from the SUMO tag peptide. This way, different scoring models can be used for the substrate peptide and the SUMO tag to account for their distinct fragmentation patterns. (c) Fragment ions from SUMOylated peptides are divided into two categories: linked-fragments and unlinked fragments. Linked fragments are from peptide fragment ions that are covalently linked to a second peptide. Linked and unlinked fragments have different fragmentation statistics, and thus different scoring models are used to score each type of fragment.

■ RESULTS

Fragmentation Pattern of SUMOylated Peptides

In order to obtain a large MS/MS data set with identified SUMOylated peptides, we designed and synthesized three combinatorial peptide libraries, each with a SUMO C-terminus

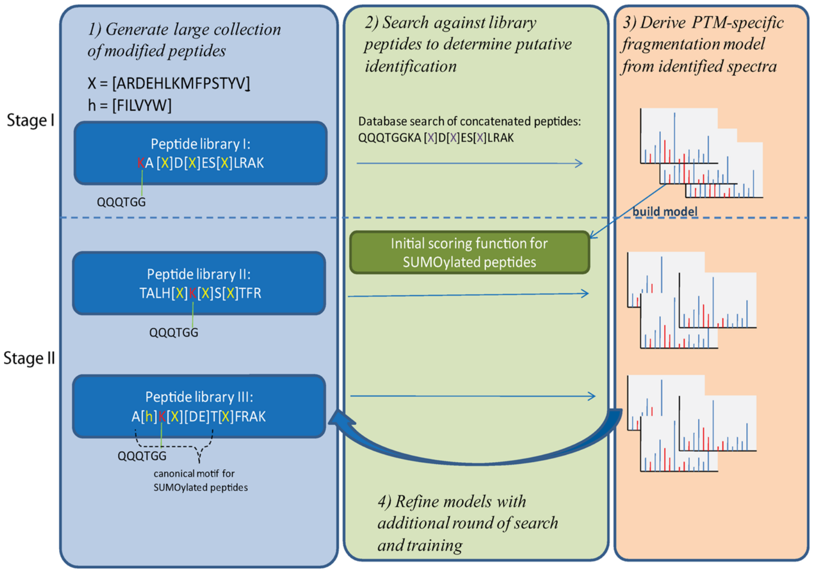

Stage I
Stage II
1) Generate large collection of modified peptides
X = [ARDEHLKMFPSTYV]
h = [FILVYW]
Peptide library I:
KA [X]D[X]ES[X]LRAK
QQQTGG
Peptide library II:
TALH[X]K[X]S[X]TFR
QQQTGG
Peptide library III:
A[h]K[X][DE]T[X]FRAK
QQQTGG
canonical motif for SUMOylated peptides
2) Search against library peptides to determine putative identification
Database search of concatenated peptides:
QQQTGGKA [X]D[X]ES[X]LRAK
Initial scoring function for SUMOylated peptides
4) Refine models with additional round of search and training
3) Derive PTM-specific fragmentation model from identified spectra
build model

Figure 2. Generating training data using combinatorial synthetic peptide libraries. We designed and synthesized three combinatorial peptide libraries, each with a SUMO tag (QQQTGG) attached via a lysine residue at a different position along the library peptide. The sequence pattern for each library is shown on the left. The symbols *X* and *h* stand for variable positions where multiple amino acid residues are possible. MS/MS spectra from peptide libraries were identified using a two-step search strategy. First for library I, since the SUMO tag is attached to the library peptide at the first residue, this is essentially equivalent to the substrate peptide having a prefix extension of QQQTGG. Thus, we can identify MS/MS spectra from library I by searching a database where the sequence QQQTGG is concatenated to the N-terminus of every possible peptide sequence in library I. This initial set of identified MS/MS spectra from SUMOylated peptides was used to build a SUMO-specific database search tool to identify MS/MS spectra from libraries II and III, which are more a realistic representation of SUMOylated peptides. Identified spectra from library II and III were then incorporated into the training data to build an even better scoring model. This refined method was used to search the spectra from all three libraries to obtain a final set of MS/MS spectra from SUMOylated peptides.

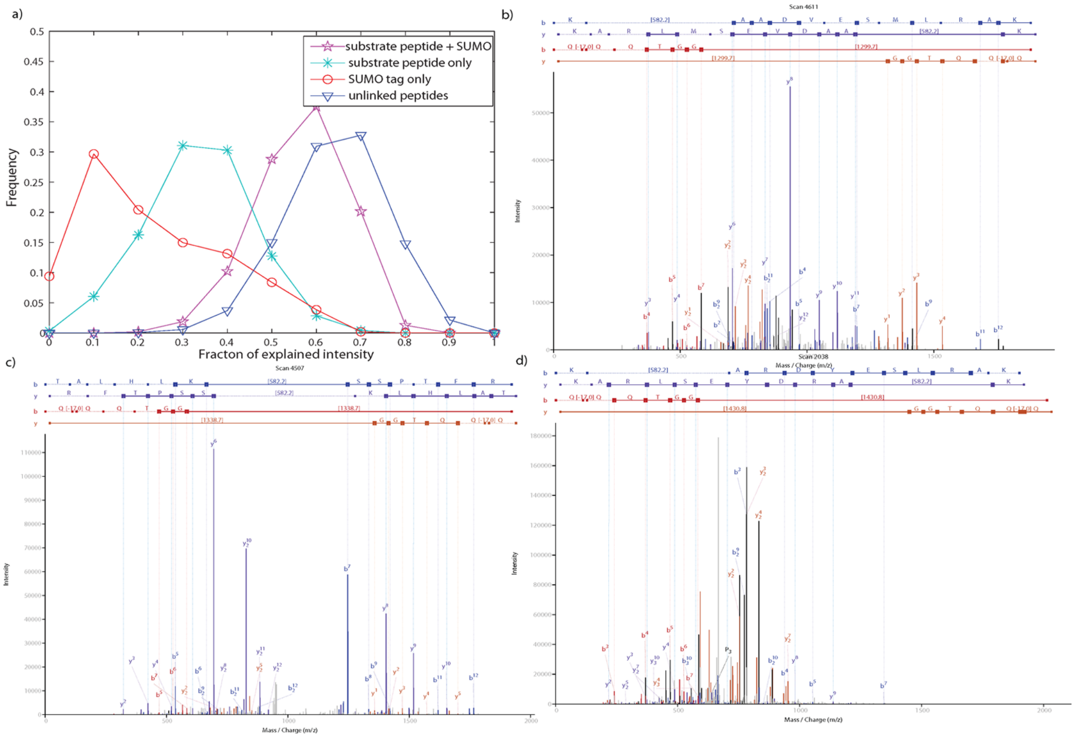


Figure 3. Contribution of ion intensity from SUMO tag. We computed the fraction of ion intensity corresponding to fragment ions from the SUMO tag in our training data. As shown in panel a, the fragment ions from the SUMO tag contribute a significant fraction (10−60%) of total intensity in the MS/MS spectra (red line). The fractions of total ion intensity from SUMO tag peptides are compared to those from substrate peptides (cyan), substrate peptide and SUMO tag combined (magenta), and linear, unlinked peptides (blue). (b−d) Examples of identified MS/MS spectra from SUMOylated peptides from the synthetic peptide libraries. Peaks explained by substrate peptides are colored blue, while peaks explained by SUMO tag are colored red. Peaks corresponding to neutral-losses or explained by both substrate peptide and SUMO tag are colored in black. (d) An example in which peaks from the SUMO tag dominate the observed MS/MS spectra.

Table 1. Identified Spectra from SUMOylated Peptides and Unique SUMOylated Peptides Identified[a]

Identified spectra from SUMOylated peptides				
	library I	library II	library III	Yeast
InsPecT	743 (6.1%)	1826 (16.4%)	531 (4.4%)	22658 (29.7%)
Specialize	2320 (19.0%)	4967 (44.7%)	2929 (24.2%)	N/A
no. of MS/MS spectra	12202	11113	12177	76177

Unique SUMOylated peptides identified			
	library I	library II	library III
InsPecT	543	941	385
Specialize	1018	1404	1070

[a] MS/MS spectra from each synthetic peptide library were analyzed by Specialize and InsPecT, and the number of identified spectra and unique peptides from SUMOylated peptides at 1% FDR are shown. Numbers inside the parentheses indicate the identification rate, which is the percentage of total number of spectra that are identified. The Yeast data set represents a typical proteomic experiment designed for linear, unlinked peptides. In comparison InsPecT's identification rate is much lower for SUMOylated peptides as compared to unlinked peptides. On the other hand, Specialize's identification rate for SUMOylated peptides is comparable to InsPecT's identification rate for unlinked peptides.

tag (QQQTGG) attached via a lysine residue at different position along the peptide (see Figure 2). The peptide libraries are designed with the goal of promoting sequence diversity while also representing a realistic model of endogenous SUMOylated peptides. For example, in library III the known consensus motif for SUMOylation is incorporated into the sequence pattern. The synthetic SUMOylated peptide libraries were analyzed using an LTQ-Orbitrap mass spectrometer, and MS/MS spectra were identified using our proposed two-step search strategy that takes advantage of the special design of the library peptides (see Figure 2). A total of 10216 MS/MS spectra from SUMOylated peptides were identified, corresponding to 3492 unique peptides. To our knowledge this is the largest mass spectral data set for SUMOylated peptides known to date. From this training data, we studied the PTM-specific fragmentation pattern of SUMOylated peptides. First the prominence of the SUMO fragment ions presented in the MS/MS spectra was assessed by the fraction of total ion intensity corresponding to SUMO tag fragments. As shown in Figure 3, SUMO fragment ions can contribute a large fraction of the total intensity in MS/MS spectra, ranging from 10% to 60% of total intensity, with an average of 20%. To put these statistics in context, we also show the fractions of total intensity from linked substrate peptides (light blue) and from common, unlinked peptides (dark blue line). Since the SUMO tag represents a significant fraction of total ion intensity in the MS/MS spectra, our new database search method, Specialize, considers all possible fragment ions from *both* the SUMO tag and the substrate peptide when matching a SUMOylated peptide against a query spectrum rather than simply treating it as a peptide with a big mass offset at lysine residue as is presently modeled in current database search methods (see Figure 1b). Moreover, we use a separate scoring model for the substrate peptide and SUMO tag to account for their difference in fragmentation statistics (see Supplementary Figure 1a).

In addition to generating extra fragment ions, the conjugation of a SUMO tag to a substrate peptide changes its physicochemical properties and thus changes its fragmentation pattern in MS/MS spectra. Conceptually, fragment ions from SUMOylated peptides can be divided into two categories: linked-fragments and unlinked fragments (see Figure 1c). Linked fragment ions are from peptide fragments that remain covalently linked to a second peptide. Assuming there is no double-fragmentation, for substrate peptides, these are fragments that are linked to the SUMO tag; for the SUMO-tag peptide, these are fragments that are linked to the substrate peptide. In general, unlinked fragments result in fragmentation patterns similar to those of common, unlinked peptides (see Supplementary Figure 1b), while linked fragments result in fragmentation patterns substantially different from those of unlinked peptides (see Supplementary Figure 1c). In particular, multiply charged fragments are more prominent (i.e., fragment ions have more intense peaks). This makes intuitive sense because linked fragments are covalently attached to a second peptide that contains an additional N and C-terminus that are also available to capture additional charges. Specialize accounts for these characteristics by introducing different ion models for linked and unlinked fragments (e.g., linked b-ions vs unlinked b-ions). Therefore, during training of the Peptide-Spectrum-Match scoring function, separate probabilistic models are used for linked and unlinked fragments.

Identifying SUMOylated Peptides in Combinatorial Peptide Library

To benchmark Specialize, we searched MS/MS spectra from the three synthetic peptide libraries using a standard database search tool, InsPecT,[29] with variable lysine modifications +599.266 Da (for SUMO tag QQQTGG) and +582.239 Da (for SUMO tag with a pyro-Q modification). To avoid overfitting, we split the data evenly into two subsets and trained Specialize on one subset and tested it on the other subset (i.e., 2-fold cross validation[30]). As shown in Table 1, Specialize identified 3−7 times more MS/MS spectra from SUMOylated peptides than InsPecT at 1% FDR. To provide some perspective, we also ran InsPecT on a Yeast data set[31] representing a typical proteomics experiment designed for linear, unlinked peptides. Out of the 76,177 MS/MS spectra in the Yeast data set, InsPecT identified 22,658 spectra corresponding to an identification rate of 29.7%. However in the three synthetic libraries of SUMOylated peptides, the identification rate for InsPecT drops to 3.2−16.4% (see Table 1), substantially lower than that of the Yeast data set even though the combinatorial peptide libraries are much less complex samples than the Yeast lysate. This supports the observations that attachment of SUMOylation tags to substrate peptides indeed changes peptide fragmentation patterns in a way that limits the ability of current database search tools to identify MS/MS spectra from SUMOylated peptides. In contrast, using Specialize's scoring models that capture SUMO-specific fragmentation characteristics, the identification rate in the SUMOylated peptide libraries was increased to 19−

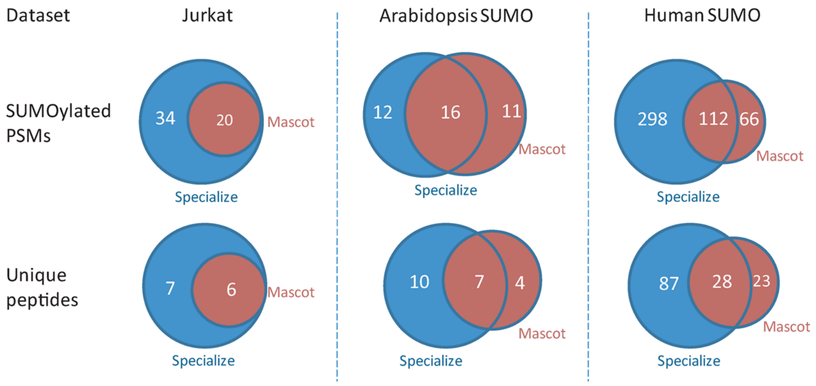


Figure 4. Comparison of identification of SUMOylated peptides between Specialize and Mascot. The ability of Specialize to identify SUMOylated peptides was tested on three data sets. The Jurkat data set contains a set of 20 synthetic SUMOylated peptides from the human MCL1 protein spiked into a background of Jurkat human cell lysate. The *Arabidopsis* and Human SUMO data sets were obtained from two previous proteomic studies on SUMO site identification. The numbers of MS/MS spectra from SUMOylated peptides as well as the numbers of unique SUMOylated peptides identified by Specialize are compared with those identified by Mascot at 5% FDR. As shown in the figure, Specialize is able to improve the identification of SUMOylated peptides by 54−125%.

44.7%, which is comparable to InsPecT's identification rate for linear, unlinked peptides.

Identifying SUMOylated Peptides from Cell Lysate

In order to demonstrate Specialize's ability to process biological samples, we synthesized 20 peptides from the human myeloid cell leukemia protein (MCL1_Human) with a SUMO tag QQQTGG attached to a lysine residue. Since MCL-1 carries canonical SUMOylation motifs, these synthetic peptides were used as a model for endogenous SUMOylated peptides. The samples were analyzed using an LTQ-Orbitrap mass spectrometer, and a total of 207 MS/MS spectra from SUMOylated peptides were identified by Specialize. This corresponds to 18 out of the 20 SUMOylated peptides synthesized. The remaining two peptides that Specialize was unable to identify because they are very short (3 and 5 residues long), reflecting the general limitation of database search methods in identifying short peptides (short peptides tend to have relatively few fragment ions in MS/MS spectra).

To test the identification of SUMOylated peptides in complex samples, the synthetic SUMOylated peptides from MCL1 were spiked into a Jurkat human cell lysate background and analyzed by MS/MS (Jurkat data set). From this data set Specialize was able to identify 13 unique MCL1 SUMOylated peptides, while Mascot was able to identify 6 unique SUMOylated peptides at 5% FDR. To estimate an upper bound for the number of SUMOylated peptides that could have been identified from the set of acquired spectra, identified MS/MS spectra from the pure MCL-1 data set described above were used to build a spectral library of MCL1 SUMOylated peptides. This spectral library was then used to search the Jurkat data to determine a list of possible SUMOylated peptides that were selected by the instrument for MS/MS analysis out of all acquired spectra. Spectral library search identified a total of 16 unique MCL1 SUMOylated peptides in the Jurkat lysate sample, which indicates that Specialize has a sensitivity of approximately 13/16 ≈ 80%. We noted that this is not related to the sensitivity of mass spectrometry to detect SUMOylated peptides but rather reflects the ability of computational tools to identify SUMOylated peptides given that such MS/MS spectra have already been acquired.

Specialize was also evaluated on two large-scale proteomic experiments from two previous studies on SUMOylation in Human[20] and *Arabidopsis*.[32] Most SUMOylation studies to date have focused on identifying potential substrate proteins. While these studies often identified many potential substrate proteins after immunoprecipitation, the number of SUMOylated peptides identified is usually rather small, underscoring the current challenges in distinguishing true SUMO substrates from immunoprecipitation artifacts.[21−24,26,33] As shown in Figure 4, in both SUMO data sets Specialize was able to increase the number of identified SUMOylated peptides by 54−125% over what was identified by Mascot[34] at 5% FDR. A detailed comparison shows that Specialize was able to identify 41 out of 68 SUMOylated peptides found by Mascot. Further investigation into those SUMOylated peptides missed by Specialize showed that either the substrate peptides are very long (e.g., ≥ 27 aa, see Supplementary Figure 2a) or the fragment ions from the SUMO tag display relatively low intensity in the MS/MS spectra: an average of only 5% of the total intensity in the spectrum as compared to 10−20% of the intensity for cases that were identified by Specialize (see Supplementary Figure 2b). It is perhaps not surprising that Specialize did not identify this subgroup of SUMOylated peptides because these observed characteristics highlight the limitations of the peptide libraries used to train Specialize. For example, the peptides in the libraries have a fixed length of 12 residues, which reflects the average length of a tryptic peptide. However, this limited diversity in peptide length may lead to a scoring model that does not capture the fragmentation pattern for long peptides very well. Similarly, in the training data the SUMO tag always contributes a significant fraction (on average 20−25%) of the total intensity in the MS/MS spectra (see Supplementary Figure 3). As a result Specialize gives considerable weight to these fragment ions from the SUMO tag when evaluating whether a SUMOylated peptide is a good match to an MS/MS spectrum. When many fragment ions from the SUMO tag are missing or of relatively low abundance in the MS/MS spectrum, Specialize is likely to assign a low score. We argue that this may actually be a desirable feature for automatic methods to have, since the presence of these fragment ions from the SUMO tag help confirm that the PTM is a SUMOylation rather than some other combination of sequence variation and modifications that happen to result in the same peptide parent mass. Nevertheless, by capturing the specific fragmentation characteristics of SUMOylated peptides, Specialize was clearly shown to be able to substantially increase the identification of SUMOylated peptides in various data sets.

DISCUSSION

A key requirement for the development of efficient and accurate computational tools for the automatic identification of MS/MS spectra is the availability of a sufficiently large set of identified spectra from distinct peptides. However, creating such a data set for atypical classes of peptides is difficult without efficient informatics tools to identify these spectra in the first place, a recurring "chicken-and-egg" problem. Traditionally these reference data sets were only possible when mass spectrometrists manually curated hundreds to thousands of MS/MS spectra, but such an approach is very labor intensive and not scalable. We demonstrated that combinatorial peptide libraries are an efficient way to address this challenge by quickly generating large numbers of unique modified peptides at very low cost. There is also no need to enrich for modified peptides from a large background of unmodified peptides because modifications are directly attached to the peptides during synthesis. MS/MS spectra from the modified peptides are readily identified using a search strategy that takes advantage of the design in the peptide libraries. Using this approach, we observed two main characteristics that make fragmentation of SUMOylated peptides different from those of linear, unlinked peptides. First, the SUMO tag fragments contribute significantly to the total ion intensity in the spectrum and require search algorithms to consider both fragment ions from the peptide and the PTM when matching spectra against SUMOylated peptides. Second, the residual attachment of a SUMO tag to the substrate peptide generates highly charged fragment ions that are not commonly observed in linear, unlinked peptides. These differences in fragmentation statistics makes current database search methods, which have mainly been designed based on linear, unlinked peptides, inappropriate for identification of MS/MS spectra from SUMOylated peptides. In our benchmark analysis of synthetic peptides, we observed that the identification rate for a regular database search tool dropped by 2−10-fold when applied to MS/MS spectra from SUMOylated peptides. On the other hand, the incorporation of PTM-specific fragmentation statistics into Specialize increased the identification rate of SUMOylated peptides and made it comparable to that of linear, unlinked peptides. Further testing on several data sets unrelated to the training data demonstrated that Specialize is able to identify significantly more SUMOylated peptides from biological samples when compared to InsPecT or Mascot. These samples originated from multiple species, and different techniques were used to enrich for the SUMOylated peptides, leading to different SUMO C-terminus tags present on substrate peptides. Specialize's ability to identify SUMOylated peptides across these samples demonstrates its robustness in identifying SUMOylated peptides from various sources in an unbiased manner. One current limitation of our approach is that the training data may not generalize to all possible SUMOylated peptides; this is illustrated by the handful of SUMOylated peptides identified by Mascot that were not identified by Specialize. It is common for different search engines to perform better on different subclasses of peptides.[35−39] As with Specialize this is usually because they tend to perform best on data similar to those used to develop the algorithm. However, in universal tools such as PepNovo,[40] Percolator,[41] or MSGFDB,[42] it is possible to retrain the models for different types of data and classes of peptides.[43] Similarly, one could potentially train a specific Specialize model for each subtype of SUMOylated peptides in order to maximize the sensitivity of the search tool at detecting SUMOylated peptides (similar to what is already done for peptides with different charge states). In fact, one of the major advantages of Specialize is that this would be readily addressable in our framework as one could retrain it by synthesizing additional peptide libraries with more diversity in sequence composition and length or by using additional spectra of SUMOylated peptides from any other sources. Finally, the core concepts of the proposed approach for developing PTM-specific search methods are not specific to SUMOylation and can potentially be used to develop new tools to identify peptides with other complex PTMs by designing and synthesizing new peptide libraries for each complex PTM.

METHODS

Combinatorial Peptide Libraries of SUMOylated Peptides

We used combinatorial peptide synthesis to generate peptide libraries with the following sequence patterns:

(I) K(QQQTGG)A[X]D[X]ES[X]LRAK
(II) TALH[X]K(QQQTGG)[X]S[X]TFR
(III) A[h]K(QQQTGG)[X][DE]T[X]FRAK

Each peptide library was synthesized with a SUMO tag (QQQTGG) attached via a lysine residue at positions 1, 6, and 3 along the substrate peptides, respectively (see Figure 2). The letter in square brackets indicates that multiple residues are possible at that position. The possible residue choices are [X] = ARDEHLKMFPSTYV, and [h] = FILVYW. The sequence patterns were designed to generate sufficient sequence variability as well as to provide a realistic model for SUMOylated peptides seen in real samples. In particular the sequence pattern for library III contains the canonical sequence motif ([XK(D/E)]) for SUMOylated peptides.[19]

After synthesis, the peptides libraries were analyzed and identified using tandem mass spectrometry. Samples from each library were injected via an autosampler for separation by reverse phase chromatography on a NanoAcquity UPLC system (Waters, Dublin, CA). Peptides were loaded onto a Symmetry C18 column (1.7 m BEH-130, 0.1 mm × 100 mm, Waters, Dublin, CA) with a flow rate of 1 μL a minute and a gradient of 2% Solvent B to 25% Solvent B (where Solvent A is 0.1% formic acid/2% ACN/water and Solvent B is 0.1% formic acid/2% water/ACN) applied over 60 min with a total analysis time of 90 min. Peptides were eluted directly into an Advance CaptiveSpray ionization source (Michrom BioResources/Bruker, Auburn, CA) with a spray voltage of 1.4 kV and were analyzed using an LTQ Velos Orbitrap mass spectrometer (ThermoFisher, San Jose, CA). Precursor ions were analyzed in the FTMS at a resolution of 60,000. MS/MS was performed in the LTQ with the instrument operated in data dependent mode whereby the top 15 most abundant ions were subjected for fragmentation.

Synthetic MCL1 Data Set

All possible chymotryptic peptides with internal lysine residues in the human myeloid cell leukemia protein (MCL1_Human) were synthesized with a SUMO2 tag (QQQTGG) attached to the lysine residue. This corresponds to a total of 20 SUMOylated peptides, including variants of the same peptide with SUMO attached to different lysine positions. This set of synthesized peptides serves as a benchmark data set to test our algorithm and also as a reference spectral library for identifying SUMOylated peptides in a real sample. To test our algorithm's ability to identify SUMOylated peptide in a complex mixture, the synthetic SUMOylated peptides from MCL1 (125fmol/peptide) were also spiked into 1 μg whole cell lysate of the human Jurkat cell. The samples were then analyzed by LC—MS/MS as described for the combinatorial peptide libraries.

Identification of SUMOylated Peptides from Combinatorial Peptide Libraries

As illustrated in Figure 2, a two-stage search strategy was used to identify the MS/MS spectra from the three synthetic peptide libraries. For peptide library I, the SUMO tag is attached to the library peptide at the first residue, which is conceptually similar to library peptides having a prefix extension of QQQTGG. Thus MS/MS spectra from peptide library I can be identified by searching a custom database where a prefix QQQTGG is added at the N-terminus of every possible peptide sequence in library I. In addition to these target sequences, an *E. coli* protein sequence database (downloaded from NCBI with Taxonomy ID 511145, ver. 08/25/2009) was used as the decoy database. The *E. coli* database was appended to the target library peptide sequences in order to generate a sufficient large database for the target/decoy approach[44] to get a reasonable estimation of FDR, since a sequence database containing only library peptide sequences would be too small for TDA to accurately estimate FDR.[45] While a decoy database of larger size than the target database may result in overestimation of the actual FDR, we opted for this more conservative approach because the identified spectra were used as reference training data for our proposed scoring models. The database search was performed using InsPecT[29] with a 1% spectrum-level false discovery rate (FDR). This allows one to identify an initial set of MS/MS spectra from SUMOylated peptides that are then used to build a SUMO-specific database search method (see next section) to identify MS/MS spectra from peptide libraries II and III, which are a more realistic representation of endogenous SUMOylated peptides. Library II contains peptides with a SUMO tag attached near the middle of the peptide, and library III contains peptides whose sequence pattern conforms to the canonical sequence motif [XK(D/E)][19] for SUMOylated peptides. After spectra from SUMOylated peptides were identified from libraries II and III, they were incorporated in our training data and used to build a better scoring model for SUMOylated peptides. Finally, this improved method was used to search the spectra from all three libraries one more time to get a final list of MS/MS spectra from synthetic SUMOylated peptides. To avoid overfitting and FDR underestimation by training and testing Specialize on the same set of PSMs, we used 2-fold cross-validation.[30,41] The PSMs were randomly split into two subsets of equal size, and Specialize was trained on one subset and tested on the other subset (and vice versa, as in standard 2-fold cross-validation). Since InsPecT does not support small precursor mass tolerance, it was run with 3 Da parent mass tolerance and 0.5 Da fragment mass tolerance. Specialize search was run with a 50 ppm precursor mass tolerance and 0.5 Da fragment mass tolerance. To make the search space comparable, when we compared the search result between Specialize and InsPecT, Specialize was also run with a 3 Da parent mass tolerance. With 50 ppm precursor mass tolerance Specialize identified a total of 2357, 4967, and 2990 MS/MS spectra from libraries I, II, and III, respectively, while with 3 Da precursor mass tolerance it identified 2320, 4967, and 2929 spectra from SUMOylated peptides, respectively.

Building a PTM-Specific Database Search Method for SUMOylated Peptides

In general MS/MS spectra from SUMOylated peptides have two defining characteristics: (1) they tend to contain a mixture of SUMO tag fragment ions and substrate peptide fragment ions, and (2) the attachment of the SUMO tag to the substrate peptide makes higher-charged fragment ions much more prominent than on spectra of unlinked peptides. To model the first characteristic we assume that each SUMOylated peptide can only fragment once and conceptually think of a SUMOylated peptide as a mixture of two peptides: a substrate peptide carrying a modification of mass +599 Da (the mass of QQQTGG) at the lysine residue and a peptide with sequence QQQTGG carrying a modification with mass of the substrate peptide at the C-terminus (see Figure 1b). In common MS/MS database search, one tries to evaluate how well a *single* candidate peptide matches to an MS/MS spectrum; for SUMOylated peptides we evaluate how well a *pair* of peptides (substrate peptide and SUMO tag) matches to a MS/MS spectrum. In previous work (MixDB[46]) we introduced a probabilistic model that describes how well a pair of peptides matches to a mixture MS/MS spectrum from coeluting peptides. The statistical framework used here extends that used in MixDB by further capturing the specific fragmentation pattern of branch-linked peptides.

Briefly, an MS/MS spectrum is represented as a vector of n bins, each representing a mass interval of width δ Da (δ depends on instrument resolution). An experimental MS/MS spectrum is represented as a vector $S = s_1, s_2, ..., s_n$ where s_i represents the peak intensity rank (ranked from most to least intense) of the highest-intensity peak in each bin. Similarly, a theoretical spectrum of a peptide $P = p_1, p_2, ..., p_n$ is represented as a vector where p_i indicates the ion-type of the fragment ion (e.g., b-ion or y-ion) with mass in that bin. The model captures peptide fragmentation statistics by using a set of annotated MS/MS spectra to learn the probability that each type of ion generates an observed peak with a given rank: $Prob(s|p)$. Similarly, a noise model, $Prob(s|0)$, can be learned using unannotated peaks in the spectrum (where the symbol 0 represents noise). The scoring function for a Peptide Spectrum Match (PSM) is thus defined as the likelihood ratio of the probability that the observed spectrum S is generated from the candidate peptide P versus the probability that the observed spectrum is generated from noise: $Score(S,P) = \sum Score(s_i,p_i) = \sum \log[(Prob(s_i|p_i))/(Prob(s_i|0))]$. Since a spectrum from a SUMOylated peptide is a mixture spectrum from two peptides, we can represent a SUMOylated peptide as two vectors $SUMO(P) = (U,T)$. The vector $U = u_1, u_2, ..., u_n$ encodes all possible fragment ions from the substrate peptide (having the SUMO tag as a lysine modification), while the vector $T = t_1, t_2, ..., t_n$ contains all possible fragments from the SUMO tag (having the substrate peptide as a C-terminus modification). In order to account for their different fragmentation patterns, separate scoring models were learned to score U and T against S. For example, for b-ion Specialize will uses a different scoring model for substrate peptides ($Score(s,b_{\text{substrate}}$) and the SUMO tag ($Score(s,b_{\text{tag}})$). Thus, the likelihood score that a spectrum S is generated from a pair of peptides (U,T) is defined as: $Score(S,(U,T)) = \sum \max(Score(u_i,p_i),Score(t_i,p_i))$. The max operation is used to model the dependency between the substrate peptide and the SUMO tag; when theoretical fragment ions from both *U* and *T* match to the same peak in the spectrum, the model assign the peak only to the theoretical fragment ion with higher probability. This avoids using the same peak twice to support the identification of substrate or tag peptides, which if not explicitly prevented will incorrectly bias toward unusually high scores for pairs of peptides with shared masses for many of their theoretical fragment ions.

In order to further capture the fragmentation statistics of branch-linked peptides Specialize separates the fragment ions from a SUMOylated peptide into linked and unlinked fragments (see Figure 1). Linked fragments are defined as fragment ions that are covalently linked to a second peptide. Specialize introduces new ion types to account for linked fragments. The original MixDB scoring model considered the standard ion types: b, $b(iso)$, $b{-}H_2O$, $b{-}NH_3$, y, $y(iso)$, $y{-}H_2O$, $y{-}NH_3$, where (iso) indicates the isotopic peak of b or y ions. Specialize further adds the ion types b_X, $b(iso)_X$, $b{-}H_2O_X$, $b{-}NH_{3X}$, y_X, $y(iso)_X$, $y{-}H_2O_X$, $y{-}NH_{3X}$ to represent the corresponding linked-fragment ions that can be generated from SUMOylated peptides. For each ion type Specialize considers charge states from one to the precursor charge of the observed MS/MS spectrum. With these new ion types, the fragmentation properties of linked-fragment ions were learned during training, and different probability/weights were assigned to linked and nonlinked fragment ions when matching a SUMOylated peptide against an MS/MS spectrum. We noted that Specialize does not attempt to find new, nonstandard ion types from peptide libraries; however, this capability could be implemented using the offset frequency function.[42,47]

Since it is not known in advance whether each spectrum comes from a SUMOylated peptide, both SUMOylated peptide candidates and non-SUMO peptide candidates are considered during database search. SUMOylated peptide candidates are scored using models with both linked and unlinked fragment ions as described above, and unlinked peptides are scored using only models with unlinked fragment ions. The top scoring peptide candidate, whether SUMOylated or not, is taken as the final match for the particular query spectrum. Because Specialize used the assigned precursor charge state in the MS/MS spectrum to determine the list of candidate peptides and the appropriate scoring model, spectra with unassigned charge states were not considered. We note that it is important to consider both SUMOylated and unlinked peptide candidates when searching a spectrum against a database even though the main goal is to identify SUMOylated peptides. This is because an MS/MS spectrum generated from a long, unlinked peptide can be mistaken as a shorter peptide candidate carrying a SUMO modification at a lysine site near the N or C-terminus of the peptide. These incorrect SUMOylated candidates can sometime obtain good scores, especially when they share a prefix/suffix with the correct unmodified peptide. Thus considering both SUMOylated and unlinked candidates for every query spectrum can reduce the chances of such false positive IDs.

After determining the highest-scoring match for each spectrum, top scoring peptide spectrum matches (PSMs) from SUMOylated peptides are scored using a Support Vector Machine (SVM)[48,49] to distinguish true matches from false positive ones. Because the current implementation of Specialize focuses on the identification of peptides with complex PTMs, non-SUMOylated PSMs were not considered for further scoring. The features used in SVM were (1) likelihood score as described above; (2) normalized score: likelihood score from (1) divided by the number of amino acids in the candidate peptide; (3) explained MS/MS intensity: total intensity of annotated peaks divided by total intensity of the spectrum; (4, 5) fraction of b and y ions present: number of b and y ions present in the spectrum divided by the number of b/y ions possible from the peptide (2 features); (6, 7) longest consecutive series of b and y ions (2 features); and (8) average mass error between theoretical and observed masses.

For SUMOylated peptides, each of the above features can be computed for the substrate peptide and SUMO tag, thus resulting in a total of 16 features. Together with the combined likelihood score that considers fragments from both the substrate peptide and SUMO tag (as described above), this defines the final list of 17 features used in the SVM model for SUMOylated peptides. The SVM model was trained using the identified MS/MS spectra from the combinatorial libraries. For each training data set, the correct PSMs were used as positive training data while top-scoring PSMs from the decoy database were used as negative training data.

Estimation of False Discovery Rate for SUMOylated Peptides

All database searches were performed against a concatenated database consisting of the target database and a decoy database created by reversing each protein sequence in the target database. For each database search result, all top scoring PSMs with peptides having the SUMO modifications were considered as SUMOylated PSMs. All SUMOylated PSMs were extracted and a false discovery rate (FDR) specific for SUMOylated PSMs was determined using the standard Target/Decoy Approach.[44] Briefly, let $\text{SUMO}_{\text{target}}$ be the number of SUMOylated PSMs matched to the target database and $\text{SUMO}_{\text{decoy}}$ be the number of SUMOylated PSMs matched to the decoy database, the FDR for SUMOylated PSM was calculated as follows:

$$\text{FDR} = \frac{\text{SUMO}_{\text{decoy}}}{\text{SUMO}_{\text{target}}}$$

Identification of SUMOylated Peptides in Biological Data Sets

For the synthetic MCL1 SUMOylated peptides (*pure MCL1 data set*), Specialize was run with same parameters while allowing the following two SUMO tags: QQQTGG and Q(−17.0265)QQTGG where Q(−17.0265) indicates pyroglutamate formation. The data were searched against a database containing all synthetic MCL1 peptide sequences with an appended *E. coli* sequence database (downloaded from NCBI, ver . 08/25/2009) as decoy. All SUMOylated peptide PSMs were extracted, and then a 5% FDR was enforced using the standard target/decoy strategy (TDA). We used a FDR threshold slightly higher than the 1% that is usually used in typical proteomic experiment because the number of spectra from SUMOylated peptides in the sample is usually small (i.e., 30−200 in the MCL1 data set). As a result, it is difficult to get a robust estimation of FDR using the TDA approach when only a very small number of decoy SUMOylated PSMs are allowed to pass the FDR threshold (i.e., 0−2 PSMs). For the human Jurkat cell lysate data set with spiked-in MCL1 peptides (*Jurkat data set*), searches were done against a database containing all synthetic MCL1 peptide sequences and a Human protein sequences (downloaded from NCBI Refseq, ver. 10/29/2010).

To estimate the an upper bound for the number of SUMOylated peptides that could have been identified from the set of acquired spectra, identified MS/MS spectra from the pure MCL-1 data set described above were used to build a spectral library of MCL1 SUMOylated peptides. This spectral library was then used to search the Jurkat data to determine a list of possible SUMOylated peptides that were selected by the instrument for MS/MS analysis out of all acquired spectra.To estimate the an upper bound for the number of SUMOylated peptides that could have been identified from the set of acquired spectra, identified MS/MS spectra from the pure MCL-1 data set described above were used to build a spectral library of MCL1 SUMOylated peptides. This spectral library was then used to search the Jurkat data to determine a list of possible SUMOylated peptides that were selected by the instrument for MS/MS analysis out of all acquired spectra. The search was done using M-SPLIT[50] using default parameters.

To compare the performance of Specialize and Mascot, the Jurkat, the *Arabidopsis*[32] and the Human SUMO[20] data sets were analyzed by both Specialized and Mascot. For the *Arabidopsis* data set, because only a subset of the MS/MS data were provided to us by the authors, only results in the following data files are considered in this manuscript: MM_cold_091609o.mzXML; MM_Sumo_hot_091609qṁzXML; Vierstra_Sumo_062209d.mzXML; Vierstra_sumo_070109dṁzXML. For the Human SUMO data set,[20] we considered only MS/MS data that were generated in the Collision-induced dissociation (CID) mode since our training data was generated in CID only. All searches were done using 50 ppm precursor mass tolerance and 0.5 Da fragment mass tolerance with variable modification N-terminal acetylaton (Nterm+42.011) and methionine oxidation (M+15.995). Because different SUMO modifications are generated in different data sets the following SUMO tags or modifications were allowed on lysine residue during the searches. For the Jurkat data set, QQQTGG (+599.266) and Q(−17.0265)-QQTGG (+582.23) were considered; for the *Arabidopsis* SUMO data set, QTGG (+343.33) and Q(−17)TGG (+326.12) modifications were considered; and for the Human SUMO data set, QQTGG (+454.18) and Q(−17.0265)QTGG (471.21) were considered. All top-scoring PSMs with the SUMO modifications were extracted and filtered to have a precursor mass error less than 15 ppm and a 5% FDR was enforced using the TDA approach. The sequence databases used were the *Arabidopsis thaliana* protein sequence database (downloaded from UniProt, ver. 5/13/2012) and the Human protein sequences (downloaded from NCBI Refseq, ver. 10/29/2010). For the Human and *Arabidopsis* data sets decoy databases were generated as usual[44] by reversing each protein sequence in the target databases.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional figures as described in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 1-858-534-8666. Fax: 1-858-534-7029. E-mail: bandeira@ucsd.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partly supported by the National Institutes of Health grants GM078596 (J.W. and P.E.B.) and 3-P41-GM103484 (N.B. and J.W.) from the National Institute of General Medical Sciences.

■ ABBREVIATIONS

MS/MS, Tandem mass spectrometry; PTM, post-translational modification; SUMO, small ubiquitin-like modifier; PSM, peptide spectrum match; PPSM, peptide/peptide spectrum match; FDR, false discovery rate; SVM, support vector machine; MCL1, myeloid cell leukemia protein

■ REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(2) Krishna, R. G.; Wold, F. Post-translational modification of proteins. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1993**, 265−298.

(3) Yang, X. J. Multisite protein modification and intramolecular signaling. *Oncogene* **2004**, *24* (10), 1653−1662.

(4) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255−261.

(5) Witze, E. S.; Old, W. M.; Resing, K. A.; Ahn, N. G. Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* **2007**, *4* (10), 798−806.

(6) Jensen, O. N. Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* **2006**, *7* (6), 391−403.

(7) Rikova, K.; Guo, A.; Zeng, Q.; Possemato, A.; Yu, J.; Haack, H.; Nardone, J.; Lee, K.; Reeves, C.; Li, Y.; et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **2007**, *131* (6), 1190−1203.

(8) Moritz, A.; Li, Y.; Guo, A.; Villen, J.; Wang, Y.; MacNeill, J.; Kornhauser, J.; Sprott, K.; Zhou, J.; Possemato, A.; et al. Akt-rsk-s6 kinase signaling networks activated by oncogenic receptor tyrosine kinases. *Sci. Signaling* **2010**, *3* (136), ra64.

(9) Spiro, R. G. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* **2002**, *12* (4), 43R−56R.

(10) Geiss-Friedlander, R.; Melchior, F. Concepts in sumoylation: a decade on. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (12), 947−956.

(11) Pearce, M. J.; Mintseris, J.; Ferreyra, J.; Gygi, S. P.; Darwin, K. H. Ubiquitin-like protein involved in the proteasome pathway of mycobacterium tuberculosis. *Sci. Signaling* **2008**, *322* (5904), 1104.

(12) Ueda, K.; Hayaishi, O. Adp-ribosylation. *Annu. Rev. Biochem.* **1985**, *54* (1), 73−100.

(13) Savitski, M. M.; Lemeer, S.; Boesche, M.; Lang, M.; Mathieson, T.; Bantscheff, M.; Kuster, B. Confident phosphorylation site localization using the mascot delta score. *Mol. Cell. Proteomics* **2011**, *10* (2), No. M110.003830.

(14) Beausoleil, S. A; Villén, J.; Gerber, S. A; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285−1292.

(15) Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J. R.; Kuster, B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **2013**, *31* (6), 557−564.

(16) Meulmeester, E.; Melchior, F. Cell biology: Sumo. *Nature* **2008**, *452* (7188), 709−711.

(17) Zhang, Y. Q.; Sarge, K. D. Sumoylation of amyloid precursor protein negatively regulates a [beta] aggregate levels. *Biochem. Biophys. Res. Commun.* **2008**, *374* (4), 673−678.

(18) Steffan, J. S.; Agrawal, N.; Pallos, J.; Rockabrand, E.; Trotman, L. C.; Slepko, N.; Illes, K.; Lukacsovich, T.; Zhu, Y. Z.; Cattaneo, E.; et al. Sumo modification of huntingtin and huntington's disease pathology. *Science* **2004**, *304* (5667), 100.

(19) Rodriguez, M. S.; Dargemont, C.; Hay, R. T. Sumo-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J. Biol. Chem.* **2001**, *276* (16), 12654−12659.

(20) Matic, I.; Schimmel, J.; Hendriks, I. A.; van Santen, M. A.; van de Rijke, F.; van Dam, H.; Gnad, F.; Mann, M.; Vertegaal, A. C. O. Site-specific identification of sumo-2 targets in cells reveals an inverted sumoylation motif and a hydrophobic cluster sumoylation motif. *Mol. Cell* **2010**, *39* (4), 641−652.

(21) Knuesel, M.; Cheung, H. T.; Hamady, M.; Barthel, K. K. B.; Liu, X. A method of mapping protein sumoylation sites by mass spectrometry using a modified small ubiquitin-like modifier 1 (sumo-1) and a computational program. *Mol. Cell. Proteomics* **2005**, *4* (10), 1626−1636.

(22) Matic, I.; van Hagen, M.; Schimmel, J.; Macek, B.; Ogg, S. C.; Tatham, M. H.; Hay, R. T.; Lamond, A. I.; Mann, M.; Vertegaal, A. C. O. In vivo identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an in vitro to in vivo strategy. *Mol. Cell. Proteomics* **2008**, *7* (1), 132−144.

(23) Schimmel, J.; Larsen, K. M.; Matic, I.; van Hagen, M.; Cox, J.; Mann, M.; Andersen, J. S.; Vertegaal, A. C. O. The ubiquitin-proteasome system is a key component of the sumo-2/3 cycle. *Mol. Cell. Proteomics* **2008**, *7* (11), 2107−2122.

(24) Blomster, H. A.; Imanishi, S. Y.; Siimes, J.; Kastu, J.; Morrice, N. A.; Eriksson, J. E.; Sistonen, L. In vivo identification of sumoylation sites by a signature tag and cysteine-targeted affinity purification. *J. Biol. Chem.* **2010**, *285* (25), 19324−19329.

(25) Galisson, F.; Mahrouche, L.; Courcelles, M.; Bonneil, E.; Meloche, S.; Chelbi-Alix, M. K.; Thibault, P. A novel proteomics approach to identify sumoylated proteins and their modification sites in human cells. *Mol. Cell. Proteomics* **2011**, *10* (2), No. M110.004796.

(26) Wohlschlegel, J. A.; Johnson, E. S.; Reed, S. I.; Yates, J. R., III Improved identification of sumo attachment sites using c-terminal sumo mutants and tailored protease digestion strategies. *J. Proteome Res.* **2006**, *5* (4), 761−770.

(27) Pedrioli, P. G. A.; Raught, B.; Zhang, X. D.; Rogers, R.; Aitchison, J.; Matunis, M.; Aebersold, R. Automated identification of sumoylation sites using mass spectrometry and summon pattern recognition software. *Nat. Methods* **2006**, *3* (7), 533−539.

(28) Hsiao, H.H.; Meulmeester, E.; Frank, B.T.C.; Melchior, F.; Urlaub, H. "ChopNSpice", a mass-spectrometric approach that allows identification of endogenous sumo-conjugated peptides. *Mol. Cell. Proteomics* **2009**, 2664−2675.

(29) 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*, 4626−4639.

(30) Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence*; Morgan Kaufmann Publishers: San Francisco, CA, 1995; Vol. *14*, pp 1137−1143.

(31) Li, J.; Zimmerman, L.J.; Park, B.H.; Tabb, D.L.; Liebler, D.C.; Zhang, B. Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular Systems Biology* **2009**, DOI: 10.1038/msb.2009.54.

(32) Miller, M. J.; Barrett-Wilt, G. A.; Hua, Z.; Vierstra, R. D. Proteomic analyses identify a diverse array of nuclear processes affected by small ubiquitin-like modifier conjugation in arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (38), 16512−16517.

(33) Jeram, S. M.; Srikumar, T.; Pedrioli, P. G. A.; Raught, B. Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites. *Proteomics* **2009**, *9* (4), 922−934.

(34) Cottrell, J. S.; London, U. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.

(35) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics* **2013**, *12*, 2383−2393.

(36) Searle, B. C; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple ms/ms search methodologies. *J. Proteome Res.* **2008**, *7* (1), 245−253.

(37) Dagda, R. K; Sultana, T.; Lyons-Weiler, J. Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics. *J. Proteomics Bioinf.* **2010**, *3*, 39.

(38) Shteynberg, D.; Deutsch, E. W; Lam, H.; Eng, J. K; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), No. M111.007690.

(39) Kwon, T.; Choi, H.; Vogel, C.; Nesvizhskii, A. I; Marcotte, E. M. MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. *J. Proteome Res.* **2011**, *10* (7), 2949−2958.

(40) Frank, A.; Pevzner, P. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−973.

(41) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923−925.

(42) Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J. R.; Pevzner, P. A. The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: applications to database search. *Mol. Cell. Proteomics* **2010**, *9* (12), 2840−2852.

(43) Payne, S. H; Yau, M.; Smolka, M. B; Tanner, S.; Zhou, H.; Bafna, V. Phosphorylation-specific ms/ms scoring for rapid and accurate phosphoproteome analysis. *J. Proteome Res.* **2008**, *7* (8), 3373−3381.

(44) Elias, J. E; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.

(45) Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification. *BMC Bioinf.* **2012**, *13* (Suppl 16), S2.

(46) Wang, J.; Bourne, P. E.; Bandeira, N. Peptide identification by database search of mixture tandem mass spectra. *Mol. Cell. Proteomics* **2011**, *10* (12), No. M111.010017.

(47) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **1999**, *6*, 327−342.

(48) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20* (3), 273−297.

(49) Joachims, T. Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods - Support Vector Learning*; Eds. Schölkopf, B.; Burges, C. and Smola, A. MIT-Press, 1999.

(50) 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.